UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA
Departamento de Filología Inglesa I

EL C-TEST: ALTERNETIVA O COMPLEMENTO DE OTRAS PRUEBAS EN EL APRENDIZAJE DEL INGLÉS COMO LENGUA EXTRANJERA

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

María de los Milagros Esteban García

Bajo la dirección del doctor
Honesto Herrera Soler

Madrid, 2007

- **ISBN: 978-84-669-3041-3**

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA

Departamento de Filología Inglesa I

EL C-TEST: ALTERNATIVA O COMPLEMENTO DE OTRAS PRUEBAS EN EL APRENDIZAJE DEL INGLÉS COMO LENGUA EXTRANJERA

TESIS DOCTORAL

MARÍA DE LOS MILAGROS ESTEBAN GARCÍA

Director: Dr. Honesto Herrera Soler

Madrid, 2007

A mi familia

ÍNDICE

CAPITULO 2. PERSPECTIVA HISTÓRICA DE LA EVALUACIÓN DE LA LENGUA

CAPÍTULO 3. RASGOS DE LOS EXÁMENES O PRUEBAS

CAPÍTULO 4. LA EVALUACIÓN DEL VOCABULARIO

CAPÍTULO 5. LAS PRUEBAS DE CIERRE

CAPÍTULO 6. EL C-TEST

CAPÍTULO 9. ANÁLISIS EMPÍRICO DE LA VALIDEZ DEL C-TEST

CAPÍTULO 10. ANÁLISIS DE REGRESIÓN LINEAL DEL C-TEST

CAPÍTULO 11. ACTUACIÓN EN EL C-TEST EN FUNCIÓN DEL ESTATUS DEMOGRÁFICO

CAPÍTULO 12. ANÁLISIS DEL CUESTIONARIO RETROSPECTIVO DE OPINIÓN

CONCLUSIONES Y SÍNTESIS DE RESULTADOS

AGRADECIMIENTOS

Esta tesis es el resultado final de varios años de trabajo durante los cuales me he sentido acompañada por muchas personas cercanas, sin cuyo apoyo esta Memoria no habría sido una realidad.

En primer lugar, he de agradecer al Dr. D. Honesto Herrera Soler, director de la tesis, su esfuerzo y dedicación, su disponibilidad, paciencia y ayuda en la elaboración de la tesis. Agradezco el seguimiento que ha realizado, su orientación y presencia constantes, desde los primeros momentos hasta la culminación del trabajo. Gracias a él he disfrutado al recorrer este largo camino.

Doy las gracias también a su Departamento por la acogida y los ánimos.

En segundo lugar, quiero expresar mi agradecimiento sincero a las profesoras de Inglés de Enseñanza Secundaria que han colaborado al aplicar las pruebas de evaluación a sus alumnos, siguiendo siempre fielmente nuestras indicaciones, porque han desarrollado una labor silenciosa pero fundamental para este trabajo. Han puesto a nuestra disposición toda la información necesaria y nos han facilitado la tarea. En especial, a Guillermina Garrido y Pilar Bruguera (IES San Isidoro de Sevilla), María Manso de Zúñiga (IES Ágora), Mª Ángeles Reglero (IES Vicente Aleixandre), Marita Matesanz e Isabel Sanz (IES Humanejos). Gracias por vuestra disponibilidad y ayuda desinteresada.

Con vosotras, el agradecimiento a todos los alumnos y más aún a los que formaron parte del estudio empírico. Porque, en definitiva, son ellos los que dan sentido a esta tesis.

Gracias a tantos profesores, ejemplo de trabajo bien hecho y a menudo poco reconocido.

También he de mencionar mi agradecimiento a los amigos, y a todos los que, de una forma u otra, han colaborado en este trabajo.

Y, por supuesto, a mi familia. A mis padres, que me enseñaron a valorar desde siempre el trabajo y el afán de superación. A Gemma, por su ayuda en la fase de redacción final, con su experiencia en estas lides. Y a Javier.

La fase final de esta Tesis ha podido ser realizada gracias a la concesión de Licencia por Estudios durante el curso 2004-05 por parte de la Consejería de Educación de la Comunidad de Madrid.

PRINCIPALES ABREVIATURAS UTILIZADAS EN ESTA TESIS

AESLA	Asociación Española de Lingüística Aplicada
AILA	Association Internationale de Linguistique Appliquée
ALTE	Association of Language Testers in Europe
ANOVA	Analysis of Variance
CALL	Computer Assisted Language Learning
CLA	Communicative Language Ability
CLA	Classical Latent Additive Test Model
CLT	Communicative Language Testing
CRM	Criterion-referenced Measurement
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ESP	English for Specific Purposes
IFE	Inglés para Fines Específicos
ILE	Inglés como Lengua Extranjera
IRT	Item Response Theory
LDP	Letter Deletion Procedure
LSP	Language for Specific Purposes
L1	First Language
L2	Second Language
LT	Language Testing
NRM	Norm-referenced Measurement
MC	Multiple Choice
PAAU	Pruebas de Aptitud para el Acceso a la Universidad
RRP	Principio de redundancia reducida
SLA	Second Language Acquisition
TLU	Target Language Use
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
UCH	Unitary Competence Hypothesis
VLS	Vocabulary Learning Strategies

PUBLICACIONES PREVIAS RELACIONADAS CON EL TEMA DE TESIS

Los resultados obtenidos a partir de los dos estudios piloto previos a esta tesis fueron expuestos en congresos de AESLA y posteriormente publicados:

- Esteban, M., Herrera, H. y Amengual, M. (2001) Niveles de correlación entre el C-test y las pruebas de Inglés de Selectividad. Comunicación al XIX Congreso Nacional de AESLA, Universidad de León.

- Esteban, M., Herrera, H. y Amengual, M. (2001) ¿Puede el C-test ser una alternativa a otras pruebas en la enseñanza del inglés como segunda lengua? *La lingüística española a finales del siglo XX. Ensayos y propuestas*, Tomo I. AESLA 1999. Universidad de Alcalá.

- Esteban, M. y Herrera, H. (2003) El C-test: instrumento apropiado para la evaluación de la competencia en inglés como lengua extranjera. *Las lenguas en un mundo global*. Universidad de Jaén, 2003.

- Esteban, M. (2005) Niveles de correlación entre el C-test y la prueba de Inglés de Selectividad. En Herrera Soler, H. y García Laborda, J. (Coord.) *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*. Valencia: Editorial UPV.

INTRODUCCIÓN

1.1. Enunciado del tema de tesis, motivación y objetivos

Esta tesis desarrolla el análisis pormenorizado de un tipo de prueba objetiva de elementos discretos: el C-test o "prueba C"[1]. Es una prueba de cierre que fue desarrollada a partir de los *clozes* tradicionales. Sus creadores, Klein-Braley y Raatz (1981), lo consideraron un instrumento de evaluación muy adecuado para medir la competencia lingüística global en lengua extranjera. Posteriormente, diversos autores han continuado investigando la validez de la prueba en distintos contextos.

Nuestro trabajo supone una revisión de sus características, diseño y aplicación en alumnos de Inglés como Lengua Extranjera en Bachillerato. En él se analizan las ventajas y los puntos débiles que derivan de este diseño.

El presente estudio nació de nuestro interés por encontrar instrumentos nuevos, prácticos, válidos y fiables para la evaluación del Inglés como Lengua Extranjera. En el diseño del C-test reconocimos una prueba novedosa llena de posibilidades, que además facilita la labor del profesorado.

Actualmente, la investigación lingüística hace uso de los métodos aplicados en las Ciencias Sociales para poder generalizar después los resultados: "Language research is based on data" (Rieveld y van Hout 2005). Se parte de la recopilación de los datos para su posterior análisis e interpretación. Por tanto, para garantizar los valores atribuidos al C-test, no basta con aplicarlo a alumnos españoles y comprobar los resultados de forma aislada, es vital validar con rigor distintos aspectos de la prueba, estudiando, por ejemplo, su correlación con otras pruebas estandarizadas que midan el mismo constructo. En el proceso de validación contamos con la valiosa

[1] En este trabajo preferimos respetar el término inglés. El origen de esta denominación, que refleja el fuerte vínculo entre C-test y *clozes* (Klein-Braley 1997: 63), se explica en el capítulo 6.

ayuda de los medios informáticos y estadísticos que nos permitieron el análisis objetivo de los datos.

Se eligieron las Pruebas de Aptitud para el Acceso a la Universidad (PAAU) como medida independiente, por ser en este momento un referente objetivo y estandarizado, ya consolidado dentro del sistema educativo español[2]. Estas pruebas, que se realizan una vez superado el 2º curso de Bachillerato y durante años han supuesto la vía de acceso a las Universidades españolas, incluyen ejercicios de diversas materias comunes y de modalidad. La prueba de Inglés forma parte de las materias comunes. Hasta ahora, todos los alumnos que pretenden entrar en la Universidad han de enfrentarse a dicha prueba de Inglés.

Como punto de partida de la tesis se hace necesario el estudio detallado del estado de la cuestión en Evaluación de Lenguas Extranjeras. El marco teórico que justifica y guía la investigación realizada en este trabajo se desarrolla en los primeros capítulos del mismo y desgrana desde los aspectos generales de la Evaluación de la Lengua o *Language Testing* (LT) hasta los más concretos, relacionados con el C-test. Nos centramos en las pruebas de lengua pragmáticas de elementos discretos y, en particular, en los *clozes* o pruebas de cierre como expresión del principio de redundancia reducida (RRP) de Spolsky (1973) y del de gramática de expectativas de Oller (1979). Dedicamos nuestra labor fundamental al estudio riguroso del C-test como tipo de prueba de cierre aún no suficientemente explorado. De hecho, como veremos, en la literatura desarrollada en torno al C-test encontramos resultados contradictorios.

A continuación, la Perspectiva Empírica, en la segunda parte de la tesis, incluye una descripción de las distintas pruebas aplicadas, la metodología seguida, y por fin, el tratamiento estadístico de los datos obtenidos en la fase experimental y las inferencias correspondientes. Por último, el proceso culmina con las conclusiones e implicaciones pedagógicas derivadas de nuestra investigación.

[2] El examen de Selectividad apareció en 1974 como prueba de acceso a los estudios universitarios (Ley 30/1974 de 24 de julio) y, a pesar de los cambios producidos en el sistema educativo, se ha mantenido hasta nuestros días (Fernández y Sanz 2005). En la literatura, dependiendo de la Universidad de referencia, dichas pruebas de acceso aparecen indistintamente con las denominaciones abreviadas PPAU, PAU o PAAU. En esta tesis adoptamos la última versión, y nos referimos a ellas indistintamente como PAAU o Selectividad. La segunda denominación es la más extendida y popular, pues alude claramente al carácter selectivo de la prueba.

El primer capítulo de la tesis plasma el marco teórico sobre Evaluación de la Lengua. Comienza con una mirada retrospectiva, que parte de los orígenes de la disciplina, para luego analizar su actividad en el momento actual y proyectarse hacia el futuro de la misma. Pretende mostrar las investigaciones que se están realizando en el campo de la Lingüística Aplicada y más concretamente en Evaluación de la Lengua, cada vez más numerosas y ricas (Alderson y Banerjee 2001). Desde todos los ámbitos se augura un futuro prometedor para la disciplina.

Con este trabajo pretendemos averiguar qué puede aportar el C-test a la realidad educativa española y, en concreto, a los distintos sujetos implicados en la tarea de la enseñanza de lenguas: profesores, alumnos e instituciones educativas. Nos interesa especialmente concretar cómo responde el C-test a las demandas y necesidades del profesorado de idiomas en materia de evaluación, preocupado por encontrar instrumentos apropiados de medida de la competencia en la Lengua Extranjera.

Algunos estudios valoran la contribución del C-test como instrumento de evaluación que hace uso del principio de redundancia reducida (Eckes y Grotjahn 2006; Rashid 2002; Babaii y Ansary 2001; Connelly 1997; Klein-Braley 1997; Dörnyei y Katona 1992; Klein-Braley y Raatz 1981, 1984; Raatz 1983), otros cuestionan su validez aparente y de constructo (Bradshaw 1990; Cohen *et al.* 1984; Jafarpur 1995, 1999; Feldmann y Stemmer 1987; Kokkota 1988). Ante tal panorama, de cierta confusión, los autores insisten en la necesidad de nuevas investigaciones.

Nuestro trabajo toma el testigo y, siguiendo las líneas de investigación que sugiere la literatura, intenta determinar:

- Qué mide el C-test, como prueba discreta y a la vez holística, y con qué tipos de prueba correlaciona mejor.
- Cuáles son sus características en términos de fiabilidad y validez (de constructo, de contenido, concurrente, etc.), a la luz de los criterios que fijan los expertos en la materia.
- Qué reacción suscita en los alumnos (validez aparente).
- Qué ventajas ofrece este diseño a los profesores de Inglés como Lengua Extranjera (factibilidad).

Porque, en definitiva, es nuestro objetivo definir si el C-test:

- Demuestra ser una prueba válida y fiable para medir la competencia global de los alumnos españoles de 2º de Bachillerato en Inglés como Lengua Extranjera.
- Se podría utilizar como alternativa a otras pruebas en la enseñanza del Inglés como Lengua Extranjera, concretamente a la PAAU de Inglés vigente.
- O bien debería utilizarse únicamente como complemento a los tipos de examen tradicionales.

De la mano de estos objetivos fundamentales han ido apareciendo otros puntos de estudio colaterales que han contribuido a dar la forma definitiva a nuestra investigación:

- Aspectos intrínsecos al propio diseño de la prueba; como las diferencias de recuperación de las omisiones dependiendo de las características del texto de partida (tema, variación léxica, densidad), el tipo de palabra mutilada (términos léxicos y funcionales), y el formato utilizado (con omisiones guiadas o no).
- Influencia de variables externas en los resultados (género, formación previa e IES).
- Posibles ventajas y/o aplicaciones de este diseño en las clases de Lengua Extranjera.

Después de haber llevado a cabo dos estudios piloto, una vez fijados los objetivos, el diseño y las etapas del trabajo empírico, comenzó la fase práctica con la creación de un C-test de 100 omisiones, siguiendo los parámetros de Klein-Braley y Raatz (1981, 1984).

Para determinar la validez concurrente y de constructo del C-test debíamos tomar otras pruebas como referencia o criterio externo. Chapelle y Abraham (1990) estudiaron su correlación con diferentes tipos de *cloze*, con un ensayo y el Group Embedded Figures Test (GEFT); Dörnyei y Katona (1992) tomaron cuatro exámenes

(Department Proficiency Test, TOEIC, un *cloze* y una entrevista oral); Ikeguchi (1998) usó el STEP exam; tanto Babaii y Ansary (2001) como Babaii y Moghaddam (2006) administraron el TOEFL, y Eckes y Grotjahn (2006), el TestDaF alemán.

En nuestro caso, se estudian las correlaciones del C-test con la prueba de Inglés de las PAAU, ya que es actualmente el referente oficial y estandarizado en nuestro país, una prueba externa a la escuela que mantiene su vigencia y cuya validez se da oficialmente por supuesta, a pesar de las voces críticas que reclaman su renovación y la mejora de algunos aspectos (Herrera 1999, 2005; García Laborda 2005; Fernández y Sanz 2005; Watts y García Carbonell 2005).

Por ello, elegimos una muestra de alumnos de 2º curso de Bachillerato para desarrollar nuestro trabajo empírico. Pero encontramos algunas limitaciones, como el hecho de que no todos los alumnos que cursan 2º de Bachillerato realicen las PAAU. Por otra parte, la información que aportan las Universidades una vez corregida la prueba se limita a la calificación global de cada sujeto.

Teniendo en cuenta estas realidades, decidimos administrar directamente en las aulas otra prueba de Inglés previamente aparecida en Selectividad (*Cavemen?*), de este modo solucionamos de forma operativa los dos problemas que planteaban las PAAU oficiales al diseño de nuestra investigación. Contamos con la información suministrada por la prueba tipo PAAU realizada en clase por todos los sujetos de la muestra y, además, con la calificación de la prueba de Inglés de los alumnos presentados a las PAAU oficiales de junio de 2001.

Como veremos más adelante, de la cuantiosa información suministrada por la prueba tipo PAAU aplicada en el aula surgieron interesantes aportaciones.

Además, puesto que nuestra motivación es eminentemente pedagógica, optamos por incluir también como referente las calificaciones de la 2ª Evaluación en la asignatura de Inglés. Este dato aporta la valoración de los profesores respectivos acerca de la competencia, progreso y aprovechamiento del alumno en la lengua. Aglutina los resultados obtenidos en distintas pruebas formales (orales y escritas), datos de la observación sistemática del profesor, apreciación del progreso y esfuerzo personal, etc. según el currículo oficial y la programación de la asignatura.

En cuanto a la validez aparente del C-test, fuertemente cuestionada por algunos autores (Bradshaw 1990; Jafarpur 1995), entendimos que la utilización de protocolos *think-aloud* no respondía a nuestras posibilidades, dado el volumen de la

muestra. En su lugar, se optó por la elaboración y administración de un cuestionario retrospectivo de opinión.

De este modo, quedaron configurados los elementos centrales de nuestra investigación, que ahora detallamos:

- El análisis de los rasgos del C-test como instrumento de evaluación de la lengua, principalmente de su validez de constructo y concurrente, a través del estudio de las correlaciones con otras pruebas y del procedimiento de regresión lineal.
- La determinación de los factores que influyen en el diseño de C-tests.
- El análisis de la validez aparente del C-test para alumnos españoles a partir de los datos de un cuestionario.
- Las perspectivas de futuro de la prueba en el sistema educativo español.

Además, se analizó de forma somera la influencia de dos variables externas en los resultados obtenidos: género y centro educativo (IES) de los sujetos de la muestra.

1.2. Hipótesis

Movidos por la convicción de que la investigación de todo instrumento de Evaluación de la Lengua debe abordarse desde una perspectiva pedagógica y teniendo en cuenta a todos los sujetos implicados en el proceso de enseñanza-aprendizaje, pretendemos establecer la validez y fiabilidad del C-test y su funcionamiento en el contexto de la Enseñanza Secundaria en España, mediante el análisis de la propia prueba y su correlación con otra prueba estandarizada; el examen de Inglés de las PAAU.

Esta tesis intenta responder a las siguientes hipótesis de trabajo:

HIPÓTESIS 1. Partiendo de las características de la prueba podemos predecir que el C-test deberá correlacionar bien con otras pruebas estandarizadas que midan la competencia global en lengua extranjera, como las PAAU, y también con las calificaciones obtenidas por los alumnos en la asignatura de Inglés.

HIPÓTESIS 2. De ello se sigue que, por sus características, al ser una prueba objetiva de elementos discretos, para un mismo sujeto, el C-test correlacionará mejor con las pruebas de tipo objetivo que con las de tipo subjetivo y holístico.

HIPÓTESIS 3. En este tipo de prueba el alumno recuperará mejor los términos funcionales que los de contenido léxico.

HIPÓTESIS 4. Los cambios en el formato influyen directamente en los resultados obtenidos; si se incluye el número de letras que corresponde a cada omisión se facilita la tarea del alumno.

HIPÓTESIS 5. Por su novedad y su carácter fragmentario, algo confuso al principio, puede conducir al rechazo. El C-test carece de validez aparente.

HIPÓTESIS 6. No habrá diferencias significativas al aplicar la variable de género.

HIPÓTESIS 7. No se prevé que existan diferencias de funcionamiento del C-test al aplicar la variable IES (zona de ubicación del centro de enseñanza).

En el proceso de validación de las hipótesis planteadas intentaremos responder a las siguientes preguntas de investigación:

1. ¿Discrimina el C-test de forma adecuada y fiable entre los sujetos, atendiendo a su competencia lingüística?

2. ¿Existe correlación significativa entre las puntuaciones obtenidas por un sujeto en un C-test y en la prueba de Inglés de las PAAU? ¿Y con respecto a la valoración que hace el profesor acerca de su progreso en la asignatura?

3. Si subdividimos la prueba de Inglés de las PAAU en las distintas preguntas que la forman, ¿hay diferencias entre la correlación del C-test con las puntuaciones obtenidas en preguntas de tipo objetivo y subjetivo? Si las hay, ¿a qué se deben y cómo se explican?

4. ¿Incide el formato utilizado en el C-test en la recuperación de las omisiones? Es decir, ¿supone una ayuda eficaz guiar las omisiones indicando el número de letras omitidas?

5. ¿Incide el tipo de término, de carácter léxico o funcional, afectado por la mutilación?

6. ¿Sería pertinente plantear un nuevo diseño de C-test, eliminando la “regla del dos” y, en su lugar, incluyendo exclusivamente omisiones “a la medida”?

7. ¿Hasta qué punto depende el funcionamiento del C-test del tipo y características del texto sobre el que esté diseñado?

8. ¿Cómo valoran los sujetos al C-test? ¿Se podría considerar que la prueba carece de validez aparente? ¿Qué datos se extraen del cuestionario de opinión al respecto?

9. ¿Qué influencia ejercen las variables género, formación previa y centro de estudios de los sujetos en los resultados obtenidos?

10. ¿Qué puede aportar el C-test como instrumento de evaluación del Inglés como Lengua Extranjera a nuestra actividad docente?

1.3. Organización y contenido de la tesis

En cuanto a la organización de esta tesis, mencionaremos que se divide en dos partes. La primera estudia el estado de la cuestión en Evaluación de la Lengua o *Language Testing* (LT), revisa y establece el marco teórico que fundamenta este campo, en el que se inscribe la tesis. La segunda aporta la investigación empírica que se ha llevado a cabo: los datos, el desarrollo y resultados de ésta. Incluye además las inferencias y conclusiones alcanzadas a la vista de los resultados, y las implicaciones pedagógicas que derivan de ellas.

Además, la tesis se inicia con esta Introducción que pretende presentar la investigación realizada, y concluye con un Apéndice final que recopila algunos materiales de interés utilizados en ella (C-tests, cuestionarios, etc.).

La primera parte de la tesis se estructura en torno a seis capítulos, cuyo contenido desglosamos a continuación.

El capítulo 1 revisa el concepto de Evaluación de la Lengua, comenzando por acotar los términos relativos a la evaluación. Alude a la legislación y recomendaciones vigentes en el sistema educativo español en materia de evaluación, pero se centra en las peculiaridades de las pruebas de lengua, los modelos de dominio de la lengua y los conceptos de “redundancia de la lengua” (Spolsky 1973) y “gramática de expectativas” (Oller 1979).

El capítulo 2 hace un recorrido histórico por la Evaluación de la Lengua desde los orígenes de la Lingüística Aplicada, como disciplina en cuyo seno se inserta la subdisciplina de la Evaluación de Lenguas. Repasa los principales enfoques de LT a lo largo del siglo XX, con especial atención al movimiento comunicativo y a la última década de LT. El análisis de algunas publicaciones periódicas especializadas

(*Language Testing, Language Learning, Language Teaching Abstracts*, etc.) nos permite tomar el pulso a la actualidad en evaluación. Culminamos el capítulo esbozando ciertos rumbos de futuro que podría tomar la disciplina en el siglo XXI.

El capítulo 3 está dedicado al estudio de los rasgos de las pruebas. En primer lugar examina los conceptos de validez y fiabilidad. Desglosa los distintos tipos de validez y, con mayor detenimiento, la validez de constructo (Messick 1989). Continúa con el análisis de las demás cualidades de las pruebas: autenticidad, carácter interactivo y factibilidad. Finaliza con una revisión del impacto o efecto rebote de las pruebas.

El C-test ha sido considerado a efectos prácticos como prueba específica de vocabulario (Chapelle 1994; Read 2000; Schmitt 2000), a pesar de que no fue ésta la intención de sus creadores. En el capítulo 4 se analiza el rol del vocabulario en el aprendizaje de lenguas extranjeras. Revisa la naturaleza y características del vocabulario: el concepto de palabra y la clasificación de las palabras atendiendo a distintos criterios. Aborda también cuestiones relativas a su adquisición o aprendizaje y a la evaluación del vocabulario.

El capítulo 5 es fundamental para el desarrollo de la tesis. En él se identifica al C-test como prueba de cierre. Comienza con una aproximación al concepto de *cloze*, que se remonta a la Psicología de la Gestalt. Revisa sus características y los distintos tipos de prueba de cierre; de ratio fija, de ratio variable (Bachman 1982, 1985), de elección múltiple y el C-test (Klein-Braley y Raatz 1981).

Con el capítulo 6 concluye la primera parte. Profundiza en el C-test como prueba de redundancia reducida que pretende mejorar algunas deficiencias de los *clozes* tradicionales. Se aportan detalles de su diseño y características, siguiendo las indicaciones de sus creadores, Klein-Braley y Raatz (1981). Finalmente, se analiza la literatura sobre el C-test y las investigaciones recientes más significativas que informan nuestro trabajo experimental.

La segunda parte de la tesis, Perspectiva Empírica, incluye el trabajo experimental desarrollado: el diseño de las pruebas, su aplicación en distintos IES, los resultados obtenidos y las conclusiones.

El capítulo 7 describe los pasos seguidos en el diseño y aplicación de dos estudios piloto, que supusieron nuestra primera aproximación al C-test como

instrumento de evaluación del Inglés como Lengua Extranjera. Se revisan los resultados obtenidos, las conclusiones y su incidencia posterior en el diseño de la investigación empírica que justifica esta tesis.

El capítulo 8 introduce la metodología de la investigación desarrollada. Describe los principales elementos que han constituido nuestro trabajo empírico: los sujetos participantes en el estudio, los distintos materiales utilizados en el mismo y otras características del contexto de la investigación. Culmina con la explicación del procedimiento utilizado y algunos aspectos relativos al tratamiento estadístico de los datos.

El capítulo 9 presenta los datos del análisis empírico que se ha desarrollado con los instrumentos estadísticos pertinentes. Parte del análisis intrínseco del C-test aplicado; resultados, diseño, subtests y funcionamiento, para validar la prueba como instrumento de evaluación de la competencia en lengua inglesa. Se centra en el estudio de la validez criterial concurrente del C-test frente a otras pruebas tomadas como medida independiente (PAAU y calificaciones en la 2ª Evaluación) a través, principalmente, del estudio de las correlaciones. Estudia también la fiabilidad del C-test mediante el método de "análisis por mitades" y el Alfa de Cronbach.

En el capítulo 10 se utiliza el procedimiento estadístico de regresión lineal para explorar las relaciones entre los subtests del C-test como variable independiente (VI) y las otras pruebas aplicadas: *Cavemen?*, PAAU de junio de 2001 y calificaciones de la 2ª Evaluación, tomadas como variables dependientes (VDs).

El estudio de la incidencia de variables externas en las pruebas completa el análisis empírico del C-test.

En el capítulo 11 se valora cómo afectan los factores demográficos en la actuación de los sujetos. Nos centramos en la incidencia del género de los sujetos y el IES de procedencia en los resultados obtenidos en las pruebas.

El capítulo 12 aborda la validez aparente del C-test a través del análisis de los datos obtenidos mediante el cuestionario retrospectivo de opinión.

A lo largo del proceso de análisis y siguiendo el orden de presentación de los datos, se da respuesta a las preguntas de investigación que aparecen plasmadas en el apartado 1.2 de esta Introducción, para llegar a confirmar o rechazar las hipótesis de trabajo planteadas.

La tesis culmina con la síntesis de resultados y las conclusiones alcanzadas.

Finalmente, se aportan algunas propuestas de carácter pedagógico y se sugieren ideas para futuras investigaciones.

CAPÍTULO 1. APROXIMACIÓN TEÓRICA AL CONCEPTO DE EVALUACIÓN: LA EVALUACIÓN DE LA LENGUA

1.1. Introducción

La evaluación forma parte de nuestra vida. De forma más o menos consciente constantemente evaluamos distintos aspectos de la vida, tanto nuestra actuación como la de los que nos rodean, con respecto a unos puntos de referencia.

En el ámbito de la enseñanza la evaluación adquiere especial relevancia, es una tarea ineludible para el profesor: "teaching involves assessment" sentencia Rea-Dickins (2004: 249).

A pesar de todo no es fácil hacer una delimitación conceptual de la evaluación. Evaluar con intención formativa no equivale a medir o clasificar, ni a aplicar pruebas. Tiene que ver con estas actividades, con las que injustamente se la identifica, pero las trasciende. Según Álvarez Méndez (2001: 11), la evaluación educativa debería entenderse como "actividad crítica de aprendizaje", también para el profesorado.

Bachman y Palmer (1996: 8) parten de la misma idea que Rea-Dickins: "virtually all language teaching programs involve some testing". Por razones deontológicas los profesores de lenguas no pueden ignorar la evaluación, es su responsabilidad elegir o crear los instrumentos adecuados y formarse en este campo. En los últimos años han tomado especial relevancia los aspectos éticos de la evaluación[3]. Apelan directamente al rol del profesor, que no termina con la creación y aplicación de pruebas. Canale (1988: 15) expone: "Once one has been involved in

[3] Buena muestra de la creciente preocupación por la ética es el volumen 14 (3) de la revista *Language Testing* (1997), número especial que recoge los artículos presentados en el simposio sobre ética en la evaluación de la lengua dentro del congreso de AILA (1996). En él se plantean algunas cuestiones fundamentales. Incluye las colaboraciones de destacados especialistas (Spolsky y Hawthorne, Elder, Norton y Starfield, Hamp-Lyons, Shohamy, Lynch y Davies) y una clarificadora introducción de Davies acerca de los límites de la ética en este campo.

gathering information, one becomes responsible in some way to see that it is used ethically".

Desde el planteamiento de la evaluación como actividad básica y no simplemente "de relleno", en España ya en la década de los 80, Alcaraz y Ramón (1980: 5) introducen algunos aspectos de la evaluación que retomaremos en capítulos posteriores. Expresan el enorme potencial de la evaluación y la responsabilidad del profesor al respecto en los siguientes términos:

> La evaluación es una fase importante en el proceso didáctico; no es una actividad de relleno. A ella debe prestar el profesor la atención debida no sólo porque mide y da resultados, sino también porque es un refuerzo del aprendizaje; y si es realista, válida y justa, de acuerdo con las posibilidades del alumno, produce formidables efectos motivadores en el aprendizaje.

En el contexto educativo los intentos de evaluar han de estar necesariamente bien estructurados, pero también es importante que exista una actitud de apertura hacia cualquier cambio que pueda mejorar la evaluación. Así lo expresa Hughes (1989: 4): "If it is accepted that tests are necessary, [...] we should do everything that we can to improve the practice of testing", siempre tomando como punto de partida la premisa de que toda mejora en la evaluación supondrá una mejora de la calidad de la enseñanza (Wragg 2003; Álvarez Méndez 2001).

Esta responsabilidad como docentes nos ha llevado al campo de la Evaluación de la Lengua, con la convicción de que toda aportación puede ser útil: "Then I believe that practitioners, rather than being consumers of other´s people research, should adopt a research orientation to their own classrooms" (Nunan 1992: xii).

Procede, por tanto:

- Revisar y precisar lo que se entiende por Evaluación de la Lengua.
- Acotar los términos que hacen referencia a la evaluación en lengua inglesa.
- Revisar la legislación y recomendaciones vigentes en materia de evaluación en el sistema educativo español.
- Analizar las peculiaridades de las pruebas de lengua y los modelos de dominio de la lengua.

- Hacer una primera aproximación a los conceptos de "redundancia de la lengua" (Spolsky 1973) y "gramática de expectativas" (Oller 1979).
- Conocer la clasificación de las pruebas de lengua según distintos criterios, con objeto de identificar qué tipo de prueba es el C-test.

1.2. El concepto de Evaluación de la Lengua

Resulta paradójico que, como advierte Amengual Pizarro (2003: 45), la definición precisa de este concepto no sea frecuente en la abundante literatura especializada. Es más, en la mayor parte de las obras se omite la definición para enfatizar los objetivos de la evaluación. Incluso la mítica obra de Lado (1961) *Language Testing* evita una definición explícita, dándola por supuesta.

A pesar de todo, la Evaluación de la Lengua (LT) se ha consolidado como disciplina independiente dentro de la Lingüística Aplicada[4]. Compartimos la idea de Amengual Pizarro (2003) de que la falta de una definición precisa de LT no hace sino aportar amplitud y flexibilidad al concepto.

El punto de partida de la inmensa mayoría de las obras sobre evaluación es considerar que la evaluación no es un hecho aislado, sino que se da siempre con un fin y en un contexto concretos (Hughes 1989; Bachman 1990; Bachman y Palmer 1996). Por eso, en evaluación no existe una "receta" que solucione todos los problemas y se adapte a todas las situaciones.

Como hemos apuntado en la introducción, los profesores evalúan el progreso de sus alumnos de muchas maneras, desde las informales, que forman parte de la rutina cotidiana del aula, hasta las pruebas más formales propuestas en los distintos sistemas educativos. Según Wragg (2003: 14) el objetivo más frecuente de la evaluación es "ofrecer retroalimentación a los enseñantes y a los alumnos para que sepan qué se ha aprendido y qué no se comprende todavía", por tanto, la evaluación se relaciona directamente con el aprendizaje de los alumnos.

Bachman y Palmer (1996) aconsejan a los profesores que sean prudentes y realistas en sus expectativas acerca de la evaluación.

[4] Véase el capítulo 2, apartados 2.1.2 y 2.1.3, sobre la perspectiva histórica de la Evaluación de la Lengua dentro de la Lingüística Aplicada.

La naturaleza de la evaluación es muy diversa. No obstante, en los trabajos relevantes sobre el tema se pueden determinar los rasgos más comunes asociados a la evaluación. El primero es su función cuantitativa. Así, Bachman y Palmer (1996: 19) respaldan la idea de las pruebas como instrumento de medida: "the primary purpose of tests is to measure". Este rasgo prioritario las distingue de otros elementos de los programas educativos.

Por otra parte, a partir de los resultados obtenidos en las pruebas, el profesor se verá abocado a emitir una serie de juicios y valoraciones sobre el proceso de enseñanza-aprendizaje (Alcaraz y Ramón 1980; Alderson 1990). Alderson (1990) destaca el aspecto valorativo de la evaluación, que requiere emitir juicios constantemente acerca de la propia actividad y de cada uno de sus elementos.

Aunque, como se acaba de comentar, no se prodigan mucho las definiciones de "evaluación del Inglés", Alcaraz y Ramón (1980: 7-8) se arriesgan y la definen en los siguientes términos:

> Entendemos por evaluación del inglés la medida del progreso del discente en su aprendizaje de la lengua inglesa con el fin de emitir un juicio de valor. Para llevar a cabo esta medición el profesor se sirve de dos procedimientos:
> a) La valoración contínua.
> b) Las pruebas formalizadas.

El aspecto valorativo, que puede interpretarse como evaluación cualitativa, nos conduce a la segunda característica de la evaluación reconocida en la literatura: su propósito pedagógico.

Rea-Dickins (2004: 249) resalta la importancia de la evaluación en la práctica docente para la toma de decisiones "about lesson content and sequencing, about materials, learning tasks and so forth". También Bachman y Palmer (1996: 8) hacen una enumeración de objetivos concretos de carácter pedagógico para los que pueden ser utilizadas las pruebas de lengua:

> They can provide evidence of the results of learning and instruction, and hence feedback on the effectiveness of the teaching program itself. They can also provide information that is relevant to making decisions about individuals [...] Finally, testing can also be used as a tool for clarifying instructional objectives.

Cuando se habla de evaluación, el término suele evocar en los profesores automáticamente la idea formal de los exámenes o pruebas. Sin embargo, en la literatura se refiere a cualquier forma de recopilación de información, desde la observación directa del progreso de los alumnos hasta la aplicación de pruebas específicas (Rea-Dickins 2004; Álvarez Méndez 2003; Recomendaciones de la Dirección General de Renovación Pedagógica sobre la LOGSE 1992) y al uso posterior que de todos los datos hace el profesor.

Aunque no se ha investigado demasiado sobre el profesor de lengua extranjera como agente de la evaluación del alumno, Rea-Dickins (2004: 253) señala su delicada posición: "sometimes torn between their role as facilitator and monitor of language development and that of assessor and judge of language performance as achievement".

Si consideramos la evaluación en sentido amplio, veremos que no sólo se evalúa al alumno, sino a todos los elementos personales y materiales que intervienen en el proceso, incluida también la actuación del profesor, los programas, los materiales e instrumentos utilizados, el propio proceso de enseñanza-aprendizaje e incluso el sistema educativo en que se inscribe.

Pero este tipo más general de evaluación no es objeto de nuestro trabajo. Nos centraremos, pues, en la evaluación de la competencia lingüística del alumno, y el instrumento de evaluación clave para nuestro trabajo será el C-test.

1.2.1. Aproximación terminológica

A pesar de la falta de definición explícita comentada en el epígrafe anterior, en la enseñanza del Inglés como Lengua Extranjera es importante clarificar el concepto de Evaluación de la Lengua (LT) y fijar las diferencias entre los distintos términos que aluden a ella en lengua inglesa: *testing, measurement, assessment* y *evaluation.* Cada uno de ellos se refiere a un aspecto de la evaluación. Pero a veces se confunden en la literatura y se utilizan como sinónimos, e incluso en la práctica, a menudo, se refieren al mismo tipo de actividad. No resulta fácil fijar sus límites

conceptuales, y, sin embargo, para centrar el tema que nos ocupa es necesario precisar desde un punto de vista operativo qué se entiende por cada una de ellas.

Bachman (1990: 51) afronta el reto de definir *test*, *measurement* y *evaluation*, pero opta por no hacerlo con los términos *assessment* y *appraisal*, que considera ambiguos. Con respecto a *assessment*, el propio autor (2004: 6) explica: "there seems to be no consensus on what precisely it means. Furthermore, a number of other terms are frequently used more or less synonymously to refer to assessment".

1.2.1.1. *Testing*

Comenzaremos acotando el concepto de *language test* por ser el más concreto[5]. Para referirnos a él en esta tesis utilizaremos indistintamente los términos *test*, *examen* o *prueba*. El diccionario de la Real Academia de la Lengua Española incluye la palabra inglesa *test* como sinónimo de las anteriores. No obstante, puesto que se suele asociar con las pruebas de tipo objetivo, preferimos optar por el término inglés para referirnos a la prueba en que se basa la tesis: el C-test. Esta denominación nos parece más apropiada que su posible traducción al español como "prueba C".

Como hemos visto en el epígrafe anterior, la literatura coincide en señalar dos aspectos fundamentales que identifican a las pruebas de lengua (*language tests*): ser instrumento de medida y fuente de información para la posterior emisión de juicios de carácter pedagógico del profesor. En realidad, ambos rasgos son comunes a cualquier examen, sea o no de lengua.

Cronbach (1971: 26, citado en MacNamara 1997:10) pone de relieve el carácter cuantitativo al definir el concepto de prueba como "procedimiento

[5] Bachman (1990) hace notar que también existe en inglés el término "*examination*". A veces se distingue entre *test* y *examination*, pero no hay consenso en cuanto a los criterios que se deben seguir a la hora de identificar sus rasgos característicos. Pilliner (1968) apunta que la diferencia puede estar en el grado de objetividad o subjetividad de la prueba. Sin embargo, el Diccionario de la RAE no limita el concepto de test a prueba objetiva. Tampoco lo hace el Diccionario de uso del español de María Moliner (2000), aún así hemos de reconocer que los tests a menudo se identifican con las pruebas objetivas y en concreto con las de elección múltiple. Por eso, el María Moliner recoge la siguiente acepción: "Examen de respuestas breves en que cada pregunta tiene varias opciones como posibles soluciones: "Un examen tipo test"".

sistemático para observar el comportamiento de un sujeto y describirlo con la ayuda de una escala numérica o un sistema de clasificación".

Carroll (1968: 46 citado en Bachman 1990: 20) expone que un examen es un instrumento del que se puede inferir un determinado comportamiento: "A psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characterisitics of an individual".

Según Bachman (1990: 21) lo que diferencia a la prueba de otros tipos de medida es precisamente esto, que ya desde su diseño pretende obtener "a specific sample of behaviour".

Bachman (1990) y Bachman y Palmer (1996), en la línea de Carroll, subrayan el carácter de instrumento de medida de los exámenes de lengua, pero dan un paso más al introducir los aspectos pedagógicos:

> Language tests can be valuable sources of information about the effectiveness of learning and teaching. Language teachers regularly use tests to help diagnose student strengths and weaknesses, to assess student progress and to assist in evaluating student achievement... (Bachman 1990: 2-3)

> Language tests can be a valuable tool for providing information that is relevant to several concerns in language teaching. (Bachman y Palmer 1996: 8)

Oller (1979: 2) aporta la definición más clara para el profano. Se fija concretamente en los exámenes de lengua extranjera tal como son percibidos por cualquiera que los haya experimentado: "For them, a language test is a device that tries to assess how much has been learned in a foreign language course, or some part of a course".

Por tanto, un examen o prueba es un acontecimiento puntual en el proceso de evaluación y en el aún más amplio proceso de aprendizaje de una lengua. Teniendo en cuenta las aportaciones de Carroll (1968), Cronbach (1971), Hughes (1989), Bachman (1990), Bachman y Palmer (1996) y Rea-Dickins (2004), entre otros, podemos concluir que los exámenes constituyen un instrumento útil de medida diseñado para suscitar un determinado comportamiento del que se infiere la adquisición de determinadas habilidades, en nuestro caso, lingüísticas. Aportan una información importante acerca del proceso de enseñanza-aprendizaje que el

profesor utilizará después según se trate de objetivos clasificatorios y/o educativos o pedagógicos.

Otro aspecto destacado de las pruebas es su dimensión social (Canale 1987, 1988)[6]. En este trabajo no profundizaremos en él, pero sería injusto no mencionar su importancia, especialmente para entender las pruebas estandarizadas externas a gran escala, o las encaminadas a la obtención de un determinado título. Chapelle y Douglas (1993: 15) apuntan:

> Conceptions of what language tests can and should do derive from a socioacademic consensus on the nature of language and appropriate methods of measurement as well as the perceived necessity for particular types of information.

Volvemos a la revisión terminológica relacionada con la evaluación. Hemos acotado el concepto de test, examen o prueba, que es concreto y limitado. *Language Testing* generalmente se refiere al aspecto cuantitativo de la aplicación de pruebas (tests) más o menos formales para obtener información acerca del aprendizaje lingüístico del alumno.

1.2.1.2. *Evaluation* y *assessment*

Los conceptos *assessment* y *evaluation* son más amplios que el de *testing*. Abarcan todo un conjunto de procesos y procedimientos utilizados en la toma de decisiones con un propósito educativo. Pueden incluir, por tanto, la administración de pruebas, pero van más lejos.

La literatura no siempre evidencia la distinción entre ambos conceptos; Bachman (1990: 51) considera que tanto *assessment*, como *appraisal*, son simplemente "stylistic variants of "evaluation" and "test"".

[6] Chapelle y Douglas (1993: 14) explican la concepción de Canale (1987, 1988), para quién una prueba es: "[...] an event, conceived of socioacademic beliefs, implemented in academic society where it conveys to test takers and instructors messages about language and learners' roles (Canale 1987) and where it is used to gain information with social consequences".

A pesar de todo, generalmente, el proceso de valoración o emisión de juicios del profesor a partir de los resultados de las pruebas se denomina *assessment*. Bachman (2004: 9) comenta:

> **Evaluation**, which involves making value judgements and decisions, can best be understood as one possible *use* of assessment, although judgements and decisions are often made in the absence of information from assessment.

Nunan (1992) puntualiza las diferencias entre *assessment* y *evaluation*. Determina que el término *assessment* se refiere al proceso que nos permite decir en qué medida un alumno ha conseguido los objetivos que pretendía alcanzar en su aprendizaje (por medio de pruebas, observación directa, cuestionarios, etc.). La amplitud del concepto de *evaluation* es aún mayor, supone la recogida de datos pero también implica el análisis y la toma de decisiones con respecto a los propios programas educativos (Nunan 1992: 185).

Bachman (1990: 22) define *evaluation* como "the systematic gathering of information for the purpose of making decisions", siguiendo a Weiss (1972). Apunta que si la información recopilada (cuantitativa o cualitativa)[7] es relevante, aumentan las probabilidades de que se tomen las decisiones correctas. De ahí la importancia de un buen diseño de pruebas.

La diferencia entre *testing* y *evaluation* es clara: las pruebas en sí mismas sólo miden, no evalúan:

> Evaluation, therefore, does not necessarily entail testing. By the same token, tests in and of themselves are not evaluative. It is only when the results of tests are used as a basis for making a decision that evaluation is involved. (Bachman 1990: 22-23)

Por tanto, concluye Bachman, se puede evaluar sin utilizar pruebas: "evaluation need not involve measurement or testing" (op. cit.: 49). Debemos puntualizar que, aunque las pruebas por sí mismas no evalúan, sobre todo en determinadas situaciones, con frecuencia sí constituyen un referente fundamental de que dispone el profesor para desarrollar la evaluación.

[7] La información cuantitativa se refiere a los resultados de las pruebas, la cualitativa a la obtenida a partir de otros instrumentos de evaluación. Véase Bachman 1990: 22.

En los contextos educativos, Allan (1999: 20) recomienda una combinación adecuada de *testing* y *assessment* para llegar a la evaluación (*evaluation*) completa y eficaz del alumno. Pone de relieve el carácter instrumental de las pruebas, frente al procedimental de la evaluación (*assessment, assessing procedures*).

> A further benefit of assessing learners, rather than just testing them is that the variety of possible approaches allows a number of wider educational objectives to be reflected. [...] Good language teachers need an understanding of, and an ability to use, a wide repertoire of test instruments and assessment procedures. The effective evaluation of learner performance in language programmes does not require teachers to make a choice between testing and assessment, but rather to use the appropriate combination of both.

A los problemas conceptuales hay que añadir los que plantea la traducción de estos términos al español. Nuestra lengua no tiene tal diversidad para expresar el concepto de evaluación. En español la palabra "evaluación" puede abarcar y denominar los procesos de *testing*, *assessment* y *evaluation*, que son sin embargo bien distintos, aunque complementarios. De ahí que a veces surjan problemas para traducirlos con precisión. En esta tesis, cuando el contexto no resulta suficiente para la adecuada comprensión, hemos decidido mantener la palabra inglesa con objeto de lograr mayor rigor; en otras ocasiones se ha optado por aportar una explicación aclaratoria.

1.2.1.3. *Measurement*

Por último, un breve comentario relativo al término *measurement*.

Es éste un término general. Hace referencia al carácter cuantitativo de la evaluación y se utiliza sobre todo en Estados Unidos como sinónimo de *Testing*. Se suele asociar el término *measurement* al de *quantification*. Bachman (1990: 18) lo define así: "Measurement in the social sciences is the process of quantifying the characteristics of persons according to explicit procedures and rules".

También puntualiza la diferencia entre *measurement* y *evaluation*: "I believe it is important to distinguish the information-providing function of measurement from the decision-making function of evaluation" (op. cit.: 23).

Prueba de la equivalencia *testing-measurement* como término general es la publicación periódica de los volúmenes *Educational Measurement* del *American Council of Education,* que agrupan las investigaciones recientes en el campo de la evaluación. Su paralelismo en cuanto a contenido con la publicación europea *Language Testing* es evidente.

1.2.2. Límites de la evaluación

A pesar de que todo profesor desearía tener la seguridad de que su tarea evaluadora es impecable y de que las pruebas que aplica son el mejor instrumento de medida, sabemos que no es así: “our tests are not perfect indicators of the abilities that we want to measure” (Bachman 1990: 30), y que debemos interpretar los resultados obtenidos con prudencia.

Según Bachman (1990), las limitaciones de la evaluación vienen dadas por múltiples factores que afectan a la especificación, observación y cuantificación.

La situación de un examen de lengua tiene sus propias características, desde las personales y cognitivas hasta las derivadas del contexto (hora del día, lugar, temperatura, tiempo asignado, tipo de tarea requerida, etc.). De todas ellas, la que más afecta es la habilidad lingüística del sujeto, que es precisamente la que se pretende medir.

Para ello, hemos de especificar a nivel teórico la destreza que queremos medir y, a nivel operativo, los aspectos de la actuación lingüística que nos servirán como indicadores de que se posee esa habilidad o competencia. Bachman (1990: 31) dice que esta especificación “defines the relationship between the ability and the test score”. Pero nunca podemos tener en cuenta todas las habilidades que refleja una prueba, por razones prácticas hemos de simplificar y, por tanto, nuestra interpretación de los resultados será necesariamente limitada.

Por otra parte, los procesos de observación y cuantificación también son limitados, porque toda medida de la habilidad mental es indirecta, incompleta, imprecisa, subjetiva y relativa (Bachman 1990: 32).

Cuando definamos la autenticidad como rasgo de las pruebas[8], veremos que las pruebas de lengua son indicadores indirectos de las destrezas que queremos medir. Además, tienen una gran carga de subjetividad, que comienza en el mismo diseño. Y la competencia lingüística es por su naturaleza siempre relativa; en una lengua extranjera nunca es nula, ni tampoco perfecta.

Estos aspectos limitan la interpretación y el uso de los resultados de las pruebas, por eso Bachman (1990: 50) recomienda seguir tres pasos en la creación de pruebas de lengua:

1. definición teórica clara de las destrezas que queremos medir,
2. especificación de las condiciones en que se va a desarrollar la prueba,
3. utilización de las escalas de medida adecuadas.

> ...(1) provide clear and unambiguous theoretical definitions of the abilities we want to measure; (2) specify precisely the conditions, or operations that we will follow in eliciting and observing performance, and (3) quantify our observations so as to assure that our measurement scales have the properties we require.

1.3. La evaluación en el sistema educativo español

Según Álvarez Méndez (2003: 40), en educación “Cualquier reforma que no parta del análisis de la situación en la que se encuentra la escuela y todo lo que la envuelve está abocada al fracaso”.

Hagamos una breve incursión en la historia reciente de nuestro sistema educativo[9]. Comenzaremos aludiendo a la Ley General de Educación de 1970 (LGE), que supuso la introducción de nuevas perspectivas en el sistema educativo español. En ella se concibe la evaluación como “la valoración del rendimiento educativo” (art. 11.1) teniendo en cuenta “los progresos del alumno en relación con su propia capacidad” (art. 19.1). La evaluación se entiende como actividad sistemática integrada en el proceso de formación del alumno, continua y personalizada. Se enfatiza su función formativa, que la hace parte fundamental de la

[8] Véase el apartado 3.5 del capítulo 3.
[9] Las distintas leyes de educación españolas pueden consultarse en: www.boe.es y/o www.mec.es.

actividad educativa y no la limita a una actividad pospuesta, como son los exámenes. Con la LGE se implantó el Curso de Orientación Universitaria, y en 1974 se aplicó el examen de Selectividad como prueba de acceso a los estudios universitarios.

Le siguieron distintas reformas educativas, como la Ley Orgánica 1/1990 de Ordenación General del Sistema Educativo de 3 de octubre (LOGSE). Esta ley incide en el carácter continuo de la evaluación, que no puede verse "reducida a actuaciones aisladas en situaciones de examen o prueba, ni identificarse con las calificaciones o con la promoción" (En Álvarez Méndez 2003: 68).

En los Reales Decretos que desarrollan la LOGSE (R. D. 1344 y 1345/1991 de 6 de septiembre) se fijan los criterios de evaluación para la Educación Primaria y Secundaria en cada una de las áreas o materias del currículo. Se descarta la evaluación normativa y se subraya de nuevo el carácter continuo e individualizado que ya proponía la Ley General de Educación (LGE). Además se valora la participación del alumno en el proceso mediante la autoevaluación.

También las orientaciones y recomendaciones de la Dirección General de Renovación Pedagógica sobre la LOGSE (1992) reconocen el *papel decisivo* de la evaluación en la formación integral de los alumnos.

En cuanto al proceso concreto de enseñanza-aprendizaje de la Lengua Extranjera, la LOGSE lo planteaba dentro de un contexto de escuela comprensiva. Para ello proporcionaba orientaciones didácticas en las que basaba el diseño de este proceso y que debían servir de guía tanto a la programación como a la actuación del profesor en el aula.

A pesar de la afirmación de Álvarez Méndez sobre las reformas educativas con la que comenzábamos este apartado, con frecuencia los cambios en el sistema educativo de un país dependen más de las vicisitudes políticas que del estudio de las necesidades reales detectadas en dicho sistema[10].

La legislación educativa española afrontó en los últimos años otra reforma educativa con la Ley Orgánica de Calidad de la Educación (LOCE) de 2002,

[10] Incluso cuando los cambios afrontan nuevas necesidades o realidades, a menudo se implantan sin que se sepa a ciencia cierta cuál será su funcionamiento. Sólo el tiempo y la evaluación del propio sistema mostrarán los resultados reales obtenidos. Airasian 1998b (en Gipps 1994) lo expresa en los siguientes términos: "Many educational innovations are adopted even though they have high levels of uncertainty; because of the nature of education the wisdom of adopting these innovations and the range of their effects are rarely known in advance".

propuesta para mejorar algunos aspectos de la LOGSE. Una de sus novedades era la desaparición del examen de Selectividad y su sustitución por la Prueba General de Bachillerato (PGB), que sería requisito para la obtención del título, y por tanto, estaría incluida en la Enseñanza Secundaria. Pero el cambio en la política española, en marzo de 2004, abrió un nuevo panorama en Educación. Se paralizó el calendario de aplicación de la Ley de Calidad para incorporar modificaciones, esbozadas en el Proyecto de la nueva Ley Orgánica de Educación (LOE) de 22 de julio de 2005. La actual Ley Orgánica 2/2006 comenzó su vigencia el 3 de mayo y el Real Decreto 806/2006 de 30 de junio, estableció el calendario de aplicación.

En este punto de inflexión nos encontramos en el momento actual[11].

1.3.1. Pautas de actuación LOGSE para el profesorado de Lenguas Extranjeras

Puesto que la LOE está apenas iniciando su vigencia, con muchos aspectos todavía pendientes de desarrollo, la LOGSE sigue siendo un importante punto de referencia para los profesores en materia de evaluación.

La Dirección General de Renovación Pedagógica (1992), siguiendo el espíritu de la LOGSE, propuso al profesorado de Educación Secundaria Obligatoria unas líneas de actuación en evaluación de la Lengua Extranjera basadas en criterios pedagógicos y en las últimas teorías vigentes en Lingüística Aplicada. A continuación exponemos algunas ideas clave relativas a la enseñanza de idiomas, que se apuntaron en dicha legislación educativa y no nos parecen obsoletas.

1. El proceso de aprendizaje de una lengua es un proceso de construcción creativa, por tanto, el profesor debe favorecer la actividad mental constructiva del alumno mediante la exposición a la lengua y la propuesta de actividades de diversos tipos. Para que el aprendizaje sea funcional es necesario que se enseñe la lengua extranjera en situaciones reales y variadas de comunicación,

[11] Porte (2002: 110) alude a la situación de inestabilidad y confusión en materia educativa, derivada de los recientes avatares políticos: “The last few years in Spain have seen continuous comings-and goings as regards educational proposals and policy at secondary and university levels. Many of these recommendations and ordinances from ministerial level must have left more than one language teaching practitioner at times uncertain or confused”.

de la vida cotidiana, y que se proporcione la posibilidad de practicar la lengua. La aplicación práctica constituye un factor motivador.

2. Se pretende conseguir unos contenidos de tipo conceptual, procedimental y actitudinal que se interrelacionan en el acto de comunicación. En la enseñanza de idiomas los contenidos de tipo actitudinal son especialmente importantes; tanto la actitud previa del alumno ante la lengua, como la apertura y respeto a otras formas de expresión.

3. El profesorado debe cuidar la relación de la lengua extranjera con otras áreas del currículo, especialmente con la lengua materna, y plantear temas interdisciplinares.

4. El proceso de aprendizaje ha de estar centrado en el alumno. La escuela atiende a grupos cada vez más heterogéneos. Esta diversidad de los alumnos requiere una metodología variada en el área de idiomas, adaptando las actividades comunicativas a los diversos niveles y ritmos con que podemos encontrarnos en el aula, en un clima de cooperación que fomente el carácter formativo de la comunicación interpersonal.

5. El tratamiento de los temas transversales será más eficaz si hay un clima positivo en el aula, partirá de la realidad de los alumnos para llegar a una aceptación, valoración y respeto por lo extranjero.

1.3.2. Propuestas LOGSE para la evaluación de Lenguas Extranjeras

Como hemos visto, la LOGSE planteaba la evaluación educativa como instrumento al servicio del proceso de enseñanza-aprendizaje.

Proponía dos tipos de evaluación: sumativa y formativa. La sumativa informa al alumno de su situación con respecto al currículo oficial en cada momento del aprendizaje. El examen es uno de los instrumentos de que dispone el profesor para llevar a cabo la evaluación sumativa, pero no el único. La evaluación formativa le

muestra en qué punto del proceso de su propio aprendizaje se encuentra en cada momento.

La Dirección General de Renovación Pedagógica, en sus orientaciones didácticas y para la evaluación de Lenguas Extranjeras (1992: 154), fijó como objetivo fundamental de la evaluación: "verificar en qué medida el alumno es capaz de utilizar la lengua aprendida en situaciones de comunicación", aportó recomendaciones para el profesorado y propuso dos instrumentos básicos:

1. la observación sistemática (a través del diario del alumno, cuestionarios, comentarios escritos de los alumnos, discusiones sobre la marcha de la clase, grabación de actividades, etc.)

2. las pruebas (principalmente de tipo comunicativo, basadas en la interacción) contextualizadas y diversificadas.

En el apartado 1.2 hemos comentado que, según Bachman y Palmer (1996), la primera función de las pruebas es la cuantitativa, y la pedagógica quedaría relegada al segundo lugar. La Dirección General de Renovación Pedagógica (1992: 159), por el contrario, expresaba: "La importancia fundamental de una prueba, o de una serie de ellas, reside en que de su resultado el profesor sacará conclusiones que repercutirán en la programación y la metodología". En el contexto concreto de la Educación Secundaria española los aspectos valorativos o cualitativos toman el papel prioritario.

Entre los cambios más discutidos y contestados que pretendía introducir la Ley de Calidad (LOCE) de 2002 estaba el modo de acceso a la Universidad. Desde 1974 se accede mediante las Pruebas Unificadas de Acceso a la Universidad, conocidas como Selectividad o PAAU. La calificación obtenida en dichas pruebas, junto a la media del expediente académico de Bachillerato, define la puntuación de cada alumno y sus posibilidades de acceder a una u otra Facultad. En su lugar, la Ley de Calidad proponía la Prueba General de Bachillerato (PGB), una "reválida" oficial.

Las indicaciones sobre las características de la nueva PGB aparecieron en el BOE, Reales Decretos 1741/2003 y 1742/2003. Una de las novedades que más iban a afectar a la enseñanza de lenguas extranjeras era la incorporación de los aspectos

orales a la prueba. Pero con la paralización de la Ley de Calidad (LCE) se decidió prolongar la vigencia del sistema de las PAAU, cuyo modelo se mantiene con la nueva Ley de Educación (LOE) (Fernández y Sanz 2005), aunque el RD 806/2006 prevé una nueva prueba de acceso, cuyas características están aún por determinar, para el curso académico 2009-2010.

Como aparece reflejado en la Introducción y se verá en capítulos posteriores, este trabajo toma como referencia a las PAAU para analizar, entre otros aspectos, la validez criterial concurrente del C-test, mediante el estudio de las correlaciones entre ambas pruebas. Fue elegida para este propósito como medida independiente por ser una prueba externa, estandarizada y de carácter nacional, de referencia durante años en el sistema educativo español.

1.3.3. Panorama actual: La escuela ante las nuevas realidades sociales

Finalmente, debemos hacer notar la situación de cierta inestabilidad que se vive hoy en el sistema educativo español. Crece el fracaso escolar. Es un periodo de ajustes, cambios, y adaptación a nuevas realidades, como la inmigración o el aumento progresivo de la violencia en las aulas (indisciplina, acoso, etc.). Evidentemente los cambios de la sociedad afectan directamente a la escuela[12]. Ésta responde a las nuevas necesidades con la creación de modelos nuevos, como las denominadas “aulas de enlace” y con la aparición de nuevas figuras dentro del ámbito escolar, como la del trabajador social. Resulta fundamental el trabajo de los Departamentos de Orientación en los centros.

El aula de lenguas extranjeras no permanece ajena a este proceso de cambio. La actual globalización propicia la necesidad de comunicarse en varias lenguas para desenvolverse de forma adecuada en el entorno internacional[13]. El Inglés se consolida como lengua de las Nuevas Tecnologías y de las relaciones

[12] Ver el Informe Pisa (2003) de la OECD sobre fracaso escolar en www.pisa.oecd.org y el Informe Cisneros VII (2005) del Instituto de Innovación Educativa y Desarrollo Directivo en www.acosoescolar.com.

[13] Esta realidad llevó al Council of Europe a la confección de un marco común de referencia para la enseñanza de lenguas: Common European Framework of Reference for Languages: Learning, Teaching, Assessement (2001).

internacionales, lo que supone una "demanda social" que actúa como impulso motivador para su aprendizaje como lengua extranjera.

Otra de las circunstancias determinantes es la incorporación a las aulas de alumnos inmigrantes procedentes de países cuya primera lengua no es el español. En nuestro sistema educativo, para muchos la tarea es doble; han de aprender simultáneamente al menos dos lenguas extranjeras: español e inglés. El profesor de Inglés se enfrenta diariamente a nuevos retos, como la diversidad de niveles en un mismo grupo de alumnos. Contar con buenos instrumentos de evaluación supone una gran ayuda en el proceso de enseñanza-aprendizaje.

1.4. Las pruebas de evaluación de la lengua

Al comenzar este capítulo hemos apuntado que la evaluación es una fase vital en el proceso de enseñanza aprendizaje, de la que las pruebas constituyen una mínima parte. El proceso de evaluación cada vez tiene más en cuenta otros factores que intervienen en el aprendizaje, no importa sólo el grado de consecución de los objetivos académicos (*Alternative Assessment*)[14]. La evaluación se realiza en distintos momentos y con finalidades diferentes.

Aunque no todos los contextos escolares requieren obligatoriamente la realización de pruebas escritas específicas, pues depende de múltiples factores, como la ratio de alumnos por clase, la motivación del alumnado, nivel, situación, etc., hemos de reconocer que sí se utilizan con regularidad en el aula. A menudo son los propios alumnos y/o sus padres los que demandan su utilización porque quieren conocer su situación (nivel de conocimientos) con respecto a la asignatura y al resto del grupo.

Cuando pensamos en un examen, con frecuencia imaginamos el examen escrito tradicional, y en este trabajo, ciertamente, centraremos la parte experimental en una prueba escrita, el C-test. Pero en realidad hay muchas otras formas de medir la actuación lingüística y todas se complementan. En la práctica docente es el

[14] En la actualidad el movimiento conocido como Evaluación Alternativa (*Alternative Assessment*) incluye todas aquellas formas de evaluación distintas de la tradicional: abarcan un periodo mayor de tiempo, son de tipo formativo más que sumativo, y producen un efecto rebote beneficioso. Uno de estos procedimientos es la autoevaluación (Véase el capítulo 2, apartado 2.5.4.2.4.).

profesor quien debe buscar el modo más equilibrado de valerse de varios instrumentos de medida para lograr una evaluación válida y fiable.

Así, aunque en el contexto del aula de Lenguas Extranjeras no deben nunca ser la única fuente de información sobre la actuación del alumno, parece claro que los exámenes proporcionan a profesores y alumnos un instrumento útil de información y análisis acerca de la situación individual de cada alumno y su evolución con respecto al grupo, sobre el resultado del proceso de instrucción, los métodos y materiales utilizados, e incluso del propio programa. Por eso nos parece importante definir los rasgos propios de las pruebas de lengua, revisar sus tipos e intentar buscar nuevas fórmulas que los mejoren.

En el apartado siguiente estudiaremos las peculiaridades específicas de las pruebas de lengua, sus objetivos y los pasos que implica su diseño.

1.4.1. Peculiaridades de las pruebas de evaluación de la lengua

Para comenzar veremos un aspecto importante que constituye una peculiaridad de las pruebas de lengua y, por tanto, no podemos ignorar: en ellas se utiliza la lengua simultáneamente como objeto de la evaluación y como medio[15] (Oller 1979; Bachman 1990).

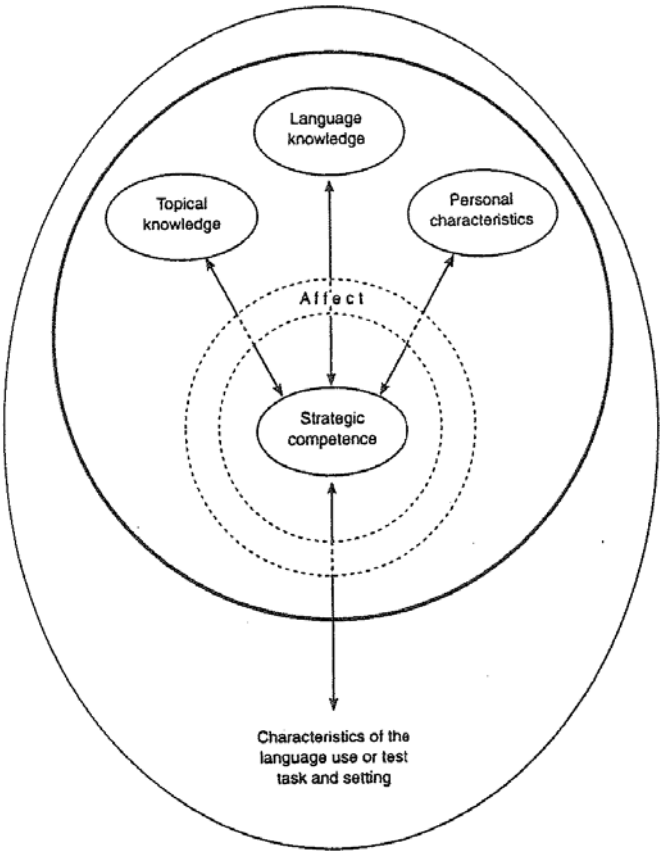
Cuando nos referimos a la evaluación del dominio de la lengua, debemos ser conscientes de que, queramos o no, cualquier tarea propuesta al alumno mide de un modo u otro su comprensión lingüística. Por otra parte, toda prueba de cualquier materia depende del manejo de la lengua.

> In one way or another, practically every kind of significant testing of human beings depends on a surreptitious test of ability to use a particular language. Consider the fact that the psychological construct of “intelligence” or IQ, at least insofar as it can be measured, may be no more than language proficiency. In any case, substantial research (see Oller and Perkins, 1978) indicates that language ability probably accounts for the lion’s share of variability in IQ tests. (Oller 1979: 2)

[15] Oller (1979: 34) explica: “Language is both an object and a tool of learning”, siguiendo la idea de Cherry (1957: 28): “this seems to imply that language is not just just a means of expressing the ideas that we already have, but rather that it is a means of discovering ideas that we have not yet fully discovered”.

En las pruebas de lengua, según Bachman y Palmer (1996), el objetivo fundamental es medir la competencia lingüística, pero además hay que tener en cuenta las características personales (edad, género, nacionalidad, estatus, lengua nativa, nivel de educación, experiencias previas, etc.), el conocimiento del mundo, la cultura y los esquemas afectivos, ya que influyen en la actuación del alumno (Fig. 1.1). Y toda prueba debe facilitar la actuación del alumno, siempre que sea posible. El uso de la lengua implica la interacción de las variables individuales con las propias de la situación en que se utiliza, en este caso, el examen de lengua.

Figura 1.1. *Some components of language use and language use performance* (Bachman y Palmer 1996: 63)



Bachman (1990: 54) describe los objetivos de las pruebas de idiomas:

> The two major uses of language tests are: (1) as sources of information for making decisions within the context of educational programs; and (2) as indicators of abilities or attributes that are of interest in research on language, language acquisition, and language teaching. In educational settings the major uses of test scores are related to evaluation, or making decisions about people or programs.

Bachman y Palmer (1996) aseguran que una prueba de lengua bien diseñada debe enriquecer al alumno y proporcionarle la oportunidad de reflejar todas sus habilidades lingüísticas, y para el profesor ha de suponer un instrumento justo y apropiado de medida. No obstante, en el proceso evaluador, proceso humano por excelencia, los docentes se enfrentarán con problemas que un examen sólo no puede resolver. La ética, profesionalidad y el buen criterio del profesor se imponen en esos momentos.

A continuación, citamos textualmente sus propuestas para que los docentes logren una mayor competencia en el uso de las pruebas:

Our philosophy of language testing

1. Relate language testing to language teaching and language use.
2. Design your tests so as to encourage and enable test takers to who use your test, accountable for the way your test is used, perform at their highest level.
3. Build considerations of fairness into test design.
4. Humanize the testing process: [...]
5. Demand accountability for test use; hold yourself, as well as any others
6. Recognize that decisions based on test scores are fraught with dilemmas, and that there are no universal answers to these.

(Bachman y Palmer 1996: 13)

1.4.2. Diseño y creación de pruebas

Bachman y Palmer (1996: 85) definen el concepto de *test development* como el proceso completo de creación y utilización de una prueba. Es decir, desde “its initial conceptualization and design” hasta que se consigue ”one or more archived tests and the results of their use”. Según la situación y el tipo de prueba, se requerirá una mayor o menor inversión en términos de tiempo y esfuerzo.

En cualquier caso, para que la experiencia del desarrollo de una prueba sea satisfactoria, conviene que todo el proceso esté previamente planificado cuidadosamente.

Según Bachman y Palmer (1996: 87) este proceso se organiza en tres estadios generalmente secuenciales: “test design, operationalization, test administration” (Véase Fig. 1.2).

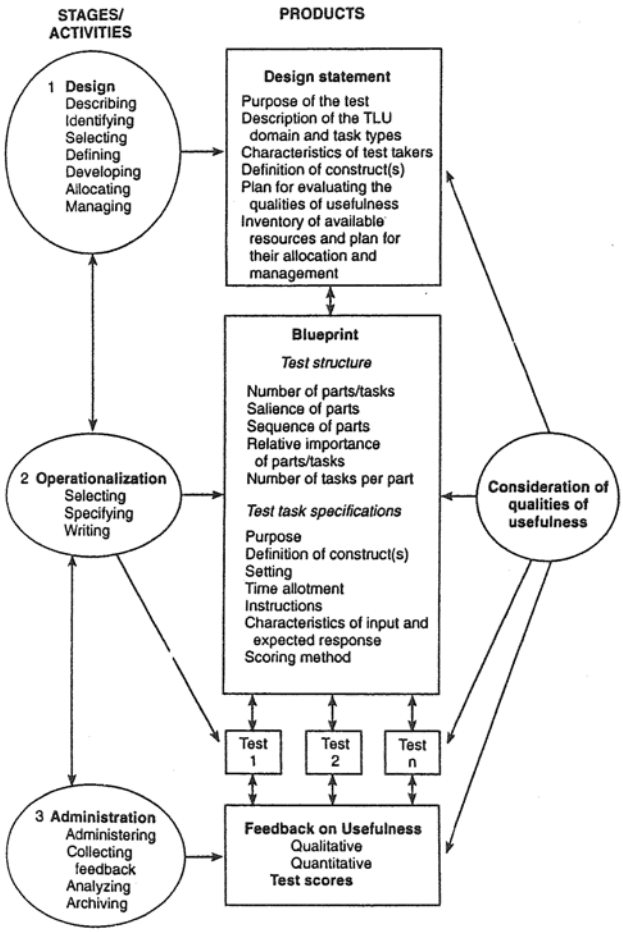
La fase de diseño de un examen comienza con la descripción del dominio lingüístico (TLU), el constructo y los tipos de tarea que queremos medir según los propósitos de la prueba. Incluye además la descripción de los examinandos, la valoración de la utilidad de la prueba y la identificación de los recursos disponibles (humanos, materiales y de tiempo) para su administración y corrección.

La segunda fase supone especificar las tareas que se van a incluir en la prueba, prever la estructura general, detallar las instrucciones que se van a dar al examinando y concretar los procedimientos de corrección.

Para medir la habilidad lingüística de un individuo hay que determinar la escala que servirá como medida. Normalmente se utilizan números que indican el grado de adquisición de las distintas destrezas y que después el profesor ha de interpretar.

Finalmente, la fase de administración de una prueba supone su aplicación a un grupo de individuos, la recogida de información y el posterior análisis (*feedback*).

Figura 1.2. *Test development* (Bachman y Palmer 1996: 87)



1.4.3. Qué evaluar en las pruebas de Lengua Extranjera

El objetivo final del profesor de Lenguas Extranjeras es que los alumnos logren aprender la lengua objeto de estudio, y que sean capaces de utilizarla con éxito en todos los contextos que sea necesario. Por tanto, ha de evaluar el grado de competencia lingüística de los alumnos que previamente han seguido un determinado proceso de aprendizaje de una lengua extranjera (Oller 1979).

En este sentido, como ya hemos mencionado, la Dirección General de Renovación Pedagógica (1992: 154) señala en su desarrollo de la LOGSE que "el objetivo principal de la evaluación será verificar en qué medida el alumno es capaz de utilizar la lengua aprendida en situaciones de comunicación reales o simuladas, pero en todo caso auténticas".

Pero surge una pregunta fundamental que retomamos de nuevo en el apartado 1.6 del presente capítulo: ¿qué es exactamente aprender una lengua? Y otra que se sigue de la anterior: ¿cómo sabemos que los resultados obtenidos por un alumno en una prueba reflejan realmente su capacidad de manejar la lengua en otras situaciones de comunicación? Bachman y Palmer (1996: 78) insisten en que para hacer inferencias a partir del nivel de un sujeto en una prueba de lengua "we must be able to demonstrate how test performance corresponds to non-test language use".

A este respecto, Bachman (1990: 21) subrayaba que no toda muestra de lengua es válida para que el profesor infiera la competencia lingüística del alumno: "However, it is precisely because any given sample of language will not necessarily enable the test user to make inferences about a given ability that we need language tests".

Por esta razón, el valor de las pruebas radica en un buen diseño: "the value of tests lies in their capability for eliciting the specific kinds of behaviour that the test user can interpret as evidence of the attributes or abilities which are of interest" (op. cit.: 22). Según Bachman y Palmer (1996: 43), para ello es necesario "to describe the characteristics of language tasks and test tasks".

Estas cuestiones nos llevan directamente al concepto de dominio de la lengua, que revisamos en el apartado siguiente.

1.5. Modelos de dominio de la lengua

A lo largo del tiempo, y teniendo como base los distintos enfoques sobre enseñanza y evaluación de la lengua, han surgido diversos modelos de dominio de la lengua (*language proficiency*) que han determinado la aparición de pruebas de evaluación también diferentes. Bachman (1990: 81) reconoce la necesidad de basar las pruebas en una teoría de dominio de la lengua:

> [...] if we are to develop and use language tests appropriately, for the purposes for which they are intended, we must base them on clear definitions of both the abilities we wish to measure and the means by which we observe and measure these abilities.

Chalhoub-Deville (1997: 3) revisa la literatura en busca de los modelos de dominio de la lengua que más han influido en la evaluación durante las dos últimas décadas y constata “a lack of consensus among models in their representation of proficiency” o, dicho de otro modo, ”no single representation of proficiency exists”.

Según Alderson (1991) la existencia de distintos modelos de dominio de la lengua plantea un dilema al profesor para decidir cuál de ellos aplicar en el diseño de pruebas.

Generalmente el propósito de las pruebas es valorar el grado de competencia lingüística a partir de los resultados que el alumno obtiene en ellas. Spolsky (1973), basándose en la dicotomía *competence vs. performance* de Chomsky (1965), explica que la competencia subyace a la actuación lingüística.

Chalhoub-Deville (1997) clasifica los modelos de dominio de la lengua en dos grupos: modelos de componentes (*componential models*) y modelos de niveles de competencia (*levels of proficiency*). Los pertenecientes al primer grupo describen los elementos del domino lingüístico, mientras que los segundos entienden el dominio de la lengua como habilidad progresiva y describen las sucesivas etapas.

Chalhoub-Deville se inclina hacia los modelos de dominio de la lengua de tipo componencial, aunque reconoce que falta investigación empírica al respecto. Nos centramos en ellos a continuación.

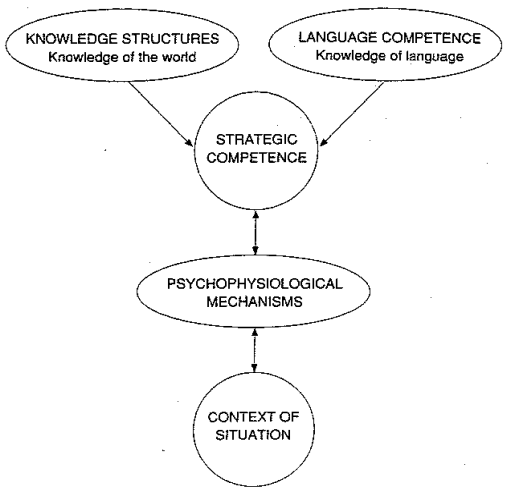
1.5.1 Modelos de componentes

Entre los modelos de componentes destacamos el de Oller (1976) denominado *Unitary Competence Hypothesis* (UCH) y el *Communicative Language Ability* (CLA) de Bachman (1990), que entronca con el enfoque comunicativo para la enseñanza de lenguas y desarrolla la idea de Canale y Swain (1983). El marco del European Council (2001) para la evaluación de la lengua se basa en el modelo de Bachman.

El modelo de Oller (1976) se basa en el análisis de los resultados de distintas pruebas y determina el dominio de la lengua en términos de un único factor general. Fue criticado por no tener en cuenta los aspectos funcionales y sociolingüísticos de la lengua (Cummins 1979). El propio autor reconoció posteriormente (Oller 1983) que el factor general se puede descomponer en otros componentes, tal como reclamaban otros autores (Bachman y Palmer 1982).

Por otra parte, Bachman (1990: 107) define su modelo como ”a framework for describing communicative language ability as both knowledge of language and the capacity for implementing that knowledge in communicative language use”. El modelo de competencia comunicativa (CLA) de Bachman (1990) incluye tres componentes interrelacionados: competencia lingüística, competencia estratégica y mecanismos psicofisiológicos (Fig. 1.3). Cada uno de ellos se subdivide en otros rasgos hasta llegar a un modelo completo y detallado cuya intención es ser “a guide, a pointer... to chart directions for research and development in language testing” (op. cit.: 82).

Figura 1.3. Componentes de la habilidad lingüística comunicativa (Bachman 1990: 85)



Posteriormente Bachman y Palmer (1996: 67) retoman el modelo de Bachman (1990). Explican que la combinación de conocimiento lingüístico y estrategias metacognitivas aporta al sujeto que utiliza la lengua: ”the ability, or capacity, to create and interpret discourse, either in responding to tasks on language tests or in non-test language use”.

La actual prueba de Selectividad española se basa en el marco teórico de Bachman (Herrera 1999: 91; Amengual Pizarro 2003: 53). Desarrolla uno de los tres componentes del dominio de la lengua: la competencia lingüística. A su vez, ésta se compone de competencia organizativa y competencia pragmática.

La organizativa “comprises those abilities involved in controlling the formal structure of language for producing or recognizing grammatically correct sentences” (Bachman 1990: 87) e incluye competencia gramatical y textual.

Bachman también nos dice que la competencia gramatical “includes those competencies involved in language usage [...] such as the knowledge of vocabulary, morphology, syntax and phonology/graphology” (ibíd.), mientras que la textual “includes the knowledge of the conventions for joining utterances together to form a text” (op. cit.: 88).

La competencia pragmática se refiere a la relación entre el hablante y el contexto de la comunicación: “includes illocutionary competence, or the knowledge of the pragmatic conventions for performing acceptable language functions, and sociolinguistic competence” (op. cit.: 90).

En la parte experimental de la tesis se explica con mayor detalle la estructura de la actual prueba de Selectividad.

El marco teórico de CLA de Bachman ha sido ampliamente reconocido y aceptado, aunque McNamara (1990) considera que, precisamente por ser tan completo, su aplicación en el diseño de pruebas puede resultar difícil.

A pesar de todo, en su reciente revisión *Looking back, looking forward: rethinking Bachman,* el autor expresa:

> Bachman (1990) remains for me the most significant book ever published on language testing, and a great intellectual contribution to the field of applied linguistics; the Bachman model showed that language testing, far from being marginal or peripheral in applied linguistics, could be a central site for the articulation of notions (the nature of language ability) which are fundamental to the whole field. (McNamara 2003: 471)

Skehan (1991) y Chalhoub-Deville (1997) auguran que, a pesar de su valía en el momento actual, el modelo de Bachman será superado por otros en el futuro:

> "God's truth" models come and go, and while the Bachman model is the best that we have at present, it is inevitable that it will be superseded and weaknesses will be revealed. (Skehan 1991: 15)

Finalmente, Chalhoub-Deville (1997: 16) insiste en que los modelos de dominio de la lengua han de ser validados en el contexto concreto en que se utilicen: "operational models, in general, and assessment frameworks, specifically -even if based on sound theoretical models- need to be empirically examined in the contexts in which they are used".

Esta validación periódica es fundamental, sobre todo en exámenes cuya superación suponga importantes consecuencias para el futuro del examinando (i. e. *high-stake examinations* como las pruebas de Selectividad en España).

1.6. El concepto de redundancia de la lengua

También Spolsky (1973) se planteó en qué consiste saber una lengua. Su enfoque nos interesa especialmente porque le llevó a profundizar en una idea fundamental para entender las pruebas de cierre y, más concretamente, el C-test: la redundancia de la lengua. El artículo *What does it mean to know a language, or how do you get someone to perform his competence?* (1973) refleja sus conclusiones.

Spolsky comparte la idea de Fries (1945): saber una lengua no consiste tanto en saber su vocabulario, cuyo conocimiento es limitado incluso para los hablantes nativos[16], como en manejar el sistema de sonidos y las estructuras de la lengua de forma automática.

Fries reconoce que es necesario conocer un cierto número de palabras para saber un idioma, pero sobre todo es necesario saber utilizarlas en distintos contextos (*functional ability*). Además hay una serie de elementos discretos que subyacen a la habilidad funcional.

[16] Véase el capítulo 4, apartado 4.2.1.1, sobre la evaluación del vocabulario.

No obstante, según Spolsky (1973: 167), la teoría de Fries olvida dos importantes aspectos de la lengua: redundancia y creatividad, “the fact that language is redundant and the fact that it is creative”.

El aspecto redundante de la lengua se empezó a estudiar con la teoría estadística de la comunicación (Shannon y Weaver 1949), según la cual un mensaje transmite una información que reduce nuestra incertidumbre: “a message carries information to the extent that it effects a reduction in uncertainty by eliminating certain probabilities” (Spolsky 1973: 167).

Somos capaces de adivinar o deducir gran parte de la información que se nos transmite, más cuanto mayor conocimiento del funcionamiento de la lengua tengamos. Spolsky (1973: 167) plantea un ejemplo ilustrativo: cuando vemos a alguien escribir su nombre, cada letra que se añade reduce las posibilidades.

> When I see him write the letter P, the uncertainty has been reduced by a large amount, for he has excluded all the names that begin with any other letter [...] As more letters are added, the amount of information conveyed by each letter becomes less, until certainty is reached.

Los mensajes contienen elementos de los que se puede prescindir sin que se rompa la comunicación. La redundancia “reduces the possibility of error and permits communication where there is some interference in the communicating channel” (op. cit.: 168). La persona que no maneja la redundancia en una lengua encuentra problemas y comete más errores, por eso normalmente los hablantes no nativos al comenzar su aprendizaje necesitan todas las redundancias de la lengua, e incluso más: gestos, que se les hable más despacio, etc.

La lengua permite la comunicación aunque una parte de los signos del mensaje no aparezcan. Según Spolsky (1973: 170), conocer bien una lengua implica comprender mensajes con redundancia reducida. Y el principio de redundancia de la lengua justifica la utilización de las pruebas de cierre en la evaluación de la lengua extranjera:

> [...] knowing a language involves the ability to understand a distorted message, to accept a message with reduced redundancy.
> From this follows the usefulness of such language testing techniques as the noise test referred to and the cloze procedure.

El autor (op. cit.: 170) se inclina hacia las pruebas integradoras frente a las de elementos discretos: “the assessment of proficiency in a language must rather be based on functioning in a much more linguistically complex situation that is provided by the one-element test”.

En cuanto al aspecto creativo de la lengua, Spolsky (1973: 173) explica que también es expresión de la competencia lingüística del individuo: “the ability to handle new sentences is evidence of knowing the rules that are needed to generate them”.

Por tanto, según Spolsky, saber una lengua es tener competencia lingüística en ella, conocer sus normas, y se expresa por la capacidad de crear en ella y de utilizar el principio de redundancia reducida si es necesario.

Basándose en la idea chomskiana de que la competencia en una lengua subyace a toda actuación lingüística, Spolsky (1973) sugiere dos posibilidades o tipos de prueba para evaluar el dominio de la lengua:

1. Las pruebas orales.
2. Las de redundancia reducida.

La primera opción consiste en hacer pruebas orales o entrevistas que muestren el funcionamiento de la lengua habitual en situaciones normales de comunicación; método inviable cuando se trata de exámenes a gran escala o pruebas de tipo institucional (oposiciones, PAAU, etc.), debido a que requiere muchos medios materiales y personales para su correcta administración (número de profesores, lugar, tiempo, etc.). Supone también fijar unos criterios muy claros de evaluación que aseguren fiabilidad en la puntuación para no sesgar la objetividad de la prueba. Otro problema es que la entrevista en sí misma no es una situación habitual y entran en juego una serie de variables propias del examen oral (desconocimiento del interlocutor, ansiedad, etc.). La segunda opción, y en su opinión la más adecuada, es usar el principio de redundancia reducida, es decir, la habilidad lingüística del alumno para comprender y expresarse en una lengua extranjera cuando introducimos ruidos en el mensaje. Este método tiene varias vías de aplicación práctica: ejercicios de dictado, tests de ruido, tests de cierre, de elección múltiple, etc. Como veremos más adelante, el C-test es un tipo de prueba de cierre y, por tanto, se basa en el principio de redundancia reducida.

1.7. El concepto de "gramática de expectativas"

El concepto de "gramática de expectativas" (*expectancy grammar*), clave para entender desde el punto de vista psicológico los procesos que subyacen al uso de la lengua, fue introducido por Oller (1979).

Como veremos, según Oller, al adquirir una lengua el ser humano crea un sistema de expectativas. Y un examen de lengua nos pide que "echemos mano" de ese sistema lingüístico: "it is claimed that for a proposed measure to qualify as a language test, it must invoke the expectancy system or grammar of the examinee" (Oller 1979: 16).

El análisis lingüístico tradicional se ha ocupado del aspecto cognitivo de la lengua; cómo el ser humano codifica una información en una secuencia de sonidos que forman sílabas, palabras y frases. Pero la lengua codifica también información afectiva o emocional (mediante la expresión gestual, el tono de voz, *body language*, etc. en la comunicación oral, y mediante otros recursos en la escrita).

En el acto de comunicación, los interlocutores estamos continuamente aportando información; anticipando lo que vamos a escuchar o las reacciones e inferencias del otro ante la información que se aporta. Oller denomina "creative inference" a este proceso. Gracias a la gramática que vamos asumiendo (*internalized grammar*) sabemos lo que podemos esperar desde el punto de vista lingüístico, qué combinaciones de fonemas o palabras son imposibles en un idioma concreto, cuáles aportan una novedad, etc.

Para Oller (1979: 24) la "gramática pragmática de expectativas" es un sistema que ordena los elementos lingüísticos de forma secuencial en el tiempo y con relación al contexto extralingüístico:

> The term pragmatic expectancy grammar further calls attention to the fact that the sequences of classes of elements, and hierarchies of them which constitute a language are available to the language user in real life situations because they are somehow indexed with reference to their appropriateness to extralinguistic contexts.

La gramática de expectativas rige el uso de la lengua para cualquier persona que la conozca y maneje: "In the normal use of language, no matter what level of

language or mode of processing we think of, it is always possible to predict partially what will come next in any given sequence of elements" (op. cit.: 25).

La explicación de Oller completa la teoría de Spolsky sobre la redundancia de la lengua expuesta en el punto anterior, y ambas justifican la utilización de pruebas de cierre.

Para Oller (1979: 32) la evaluación de la lengua "is primarily a task of assessing the efficiency of the pragmatic expectancy grammar the learner is in the process of constructing". Desde esta perspectiva, la validez de una prueba viene dada por su capacidad para medir la gramática de expectativas que va desarrollando el alumno: "Valid language tests are defined as those tests which meet the pragmatic naturalness criteria -namely those which invoke and challenge the efficiency of the learner's expectancy grammar" (op. cit.: 34).

Las pruebas de cierre, y entre ellas el C-test, demandan al alumno la aplicación de la gramática de expectativas. Son pruebas de tipo pragmático: "*cloze tests meet both of the naturalness criteria for pragmatic languages tests*" (op. cit.: 42). Por tanto, Oller las considera un instrumento válido de medida de la competencia lingüística.

Como hemos visto, las pruebas de cierre responden tanto al concepto de redundancia reducida de Spolsky como a la gramática pragmática de expectativas que propone Oller.

1.8. Tipos de pruebas de evaluación de la lengua

En el apartado 1.2 hemos comentado que no existe un test ideal, apropiado para todos los alumnos y válido para todas las situaciones, ni una receta para su creación. Elegir o diseñar un buen examen depende del propósito, los destinatarios y la situación en la que se aplique (Hughes 1989; Bachman 1990; Bachman y Palmer 1996).

Con objeto de encuadrar al C-test como prueba de lengua, veremos las clasificaciones de los exámenes en el contexto de los cursos de lengua. Las aportan numerosos autores, tales como Oller (1979), Hughes (1989), Bachman (1990), Bachman y Palmer (1996), atendiendo a criterios diversos.

Una primera clasificación tiene en cuenta el objetivo de la prueba y las decisiones del profesor que implica. La segunda divide las pruebas según su naturaleza en directas e indirectas. Atendiendo al número de elementos lingüísticos que midan, serán de elementos discretos o integradoras. Dentro de las integradoras situamos todo el elenco de pruebas pragmáticas. Según el método de corrección pueden ser objetivas o subjetivas. Y si tenemos en cuenta el grupo en que se aplican, las dividimos en normativas y criteriales (*norm-referenced* y *criterion-referenced*).

1.8.1. Según su propósito

Seguimos la taxonomía de Hughes (1989) que clasifica las pruebas según el objetivo con que se aplican. Coincide en gran parte con la que hacen Bachman y Palmer (1996), en este caso, teniendo en cuenta el tipo de decisión que implican por parte del profesor con respecto al alumno.

1.8.1.1. Pruebas de competencia lingüística (*Proficiency tests*)

Son las diseñadas para medir la competencia del alumno en la lengua objeto de estudio, independientemente del proceso de aprendizaje que hayan seguido en esa lengua. Bachman (1990: 71) las considera “theory-based tests” frente a las pruebas de logro, que son “syllabus-based”.

Normalmente estas pruebas implican que el profesor tome decisiones para seleccionar a los alumnos: “Selection decisions involve determining which individuals should be admitted to a particular educational program or offered a particular job” (Bachman y Palmer 1996: 97).

El contenido de este tipo de examen dependerá de los objetivos que se planteen para considerar que el candidato ha alcanzado el nivel deseado para un propósito concreto: Se basa en “a specification of what candidates have to be able to do in the language in order to be considered proficient” y su función es “to show

whether candidates have reached a certain standard with respect to certain specified abilities" (Hughes 1989: 9-10).

Ejemplos de esta clase de prueba son las estandarizadas, como las de la Universidad de Cambridge (PET, First Certificate, etc.) u otros exámenes cualesquiera encaminados a la consecución de un determinado título, pero no directamente relacionados con la realización de un curso previo.

1.8.1.2. Pruebas de adquisición de objetivos programados (*Achievement tests*)

Son las pruebas con las que los profesores estamos en contacto más directo en el aula, puesto que se relacionan directamente con los cursos de lengua y pretenden establecer el grado de consecución de los objetivos programados.

Se basan en la programación y el currículo (contenidos) que se haya seguido en un curso determinado. El diseño de estos exámenes depende, por lo tanto, de la programación del curso, de los materiales utilizados, etc. (*syllabus-content approach*).

Hughes distingue dos clases de pruebas de adquisición o consecución de objetivos programados:

- *final tests*
- *progress tests*

Es decir, las que se realizan al final de un curso, y las que se aplican largo del proceso de aprendizaje para medir el progreso de los alumnos con relación a los objetivos a corto plazo. Estas últimas no tienen porqué ser tan formales y rigurosas como las finales. Cada profesor puede adaptarlas según su estilo, son su responsabilidad, de manera que reflejan "the particular "route" that an individual teacher is taking towards the achievement of objectives" (Hughes 1989: 13).

La información que se obtiene a partir de las pruebas de adquisición de objetivos puede ser útil tanto para la evaluación formativa del alumno "to help students guide their own subsequent learning, or for helping teachers modify their

teaching methods and materials" como para la sumativa "of students' achievement or progress at the end of a course of study" (Bachman y Palmer 1996: 98).

Hughes (1989: 12) hace notar que las pruebas finales de este tipo pueden coincidir en sus características con las de competencia lingüística en algunas ocasiones: "If a test is based on the objectives of a course, and these are equivalent to the language needs on which a proficiency test is based".

Éste sería el caso de las PAAU, por una parte la prueba de Inglés podría considerarse de competencia lingüística, pues superación supone alcanzar un determinado nivel de competencia que permite el acceso a la Universidad. Pero, por otra, se basa en unos objetivos directamente relacionados con el currículo y la programación didáctica de la asignatura para el segundo curso de Bachillerato.

El trabajo experimental de nuestra tesis debe inscribirse en este epígrafe. No obstante, más adelante veremos que el C-test es una prueba versátil que podría utilizarse casi en cualquiera de los tipos de prueba que estamos mencionando.

1.8.1.3. Pruebas de diagnóstico (*Diagnostic Tests*)

Se aplican antes de comenzar un determinado curso de lengua. Sirven como indicador de los problemas y necesidades de los alumnos. Permiten al profesor detectarlos para planificar la enseñanza posterior.

Hughes (1989: 13) define la finalidad u objetivo de las pruebas de diagnóstico: "They are intended primarily to ascertain what further teaching is necessary".

El citado autor lamenta la falta de buenos tests de diagnóstico, que serían muy útiles para la enseñanza personalizada y el autoaprendizaje.

1.8.1.4. Pruebas de nivel (*Placement Tests*)

Las pruebas de nivel sirven para situar al alumno en el nivel que le corresponda dentro de un programa de cursos, o para organizar al alumnado en grupos homogéneos en cuanto al grado de conocimiento o manejo de las destrezas lingüísticas. Bachman y Palmer (1996: 97) explican: "Placement decisions involve

determining in which of several different levels of instruction it would be most appropriate to place the test taker".

Aunque existen algunos estandarizados, lo ideal es que se preparen los adecuados para cada situación concreta.

1.8.2. Según la naturaleza de las tareas propuestas

A continuación, abordamos la división de las pruebas atendiendo a la naturaleza de las tareas que proponen al alumno. Distinguimos entre pruebas directas e indirectas.

1.8.2.1 Pruebas directas

Comenzaremos con la definición de Rea (1985: 26):

> A test may be defined as "direct" to the extent that it requires the integration of linguistic, situational, cultural, and affective constraints which interact in the process of communicating. [...] "Directness" is therefore crucially concerned with situational and communicative realism.

Así pues, son directas aquellas pruebas que proponen al alumno la realización de la tarea concreta que se pretende medir. Hughes (1989: 15) explica: "If we want to know how well candidates can write compositions, we get them to write compositions". Es obvio que resulta más sencillo cuando se quieren medir las destrezas de tipo productivo.

Hughes recomienda que tanto las tareas como los textos utilizados sean auténticos en la medida de lo posible, aunque la propia situación de las pruebas impida la total autenticidad. El autor (1989: 15) subraya el atractivo de las pruebas directas frente a las indirectas porque:

> ...it is relatively straightforward to create the conditions which will elicit the behaviour on which to base our judgements [...] the assessment and interpretation of students' performance is also quite straightforward.

A lo que hay que añadir, además, la mayor probabilidad de que el impacto producido sea positivo: “there is likely to be a helpful backwash effect” (Hughes 1989: 15).

Rea (1985: 26) alude a otro rasgo que caracteriza a las pruebas directas, su validez aparente: ““direct” measures have popular appeal to face validity. In other words, the more a test looks as if it is testing what it is intended to measure, the better it is”.

1.8.2.2. Pruebas indirectas

Las pruebas indirectas proponen al alumno tareas en las que subyacen las habilidades que se quieren medir. La ventaja de las pruebas indirectas, en palabras de Hughes (1989: 16), es que “they offer the possibility of testing a representative sample of a finite number of abilities which underlie a potentially indefinitely large number of manifestations of them” y sus resultados son, por tanto, más generalizables.

Sin embargo, las pruebas directas son más fáciles de diseñar. Hughes las recomienda para los exámenes de competencia lingüística y los de consecución de objetivos programados.

El C-test, entre otros, constituye un buen ejemplo de prueba indirecta.

1.8.3. Según el número de elementos lingüísticos que se mida en cada prueba

Carroll (1961) fue el primero en diferenciar entre tests de lengua discretos e integradores. Oller (1979: 70) los considera “two extremes on a continuum”. Como veremos, dentro de los integradores, los tests pragmáticos constituyen una clase especial.

1.8.3.1. Pruebas de elementos discretos *(Discrete point tests)*

Las pruebas de elementos discretos miden un solo elemento de la lengua, una estructura gramatical concreta. Según Oller (1973b: 190) su principal limitación es que no reflejan el uso real de la lengua. Además encuentra otras desventajas que citamos textualmente: “they often fail to provide the student with practice in useful language skills [...] They require substantial skill on the part of the person who prepares them”.

Algunas de las ventajas de las pruebas de elementos discretos son evidentes, como su fiabilidad y carácter práctico. Destaca también su facilidad de administración y corrección, vital cuando se necesita examinar a un gran número de sujetos en breve espacio de tiempo. Actualmente su uso ha disminuido, Arnaud (1984: 14) expresa la situación en los siguientes términos: “Discrete-item tests of separate components of language, although still in use for practical reasons, have tended to fall out of fashion as language testing theory has begun to place more emphasis on validity”.

1.8.3.2. Pruebas integradoras

Son las que demandan tareas en las que es necesario utilizar varios elementos lingüísticos. Se definen por contraposición con las de elementos discretos:

> ... integrative tests attempt to assess a learner’s capacity to use many bits all at the same time, and possibly while exercising several presumed components of a grammatical system, and perhaps more than one of the traditionally recognized skills or aspects of skills. (Oller 1979: 37)

Dentro de las pruebas integradoras se incluye el dictado, las de comprensión lectora, la redacción y las pruebas de cierre como “one of the most promising types”.

A pesar de que se les achaca falta de fiabilidad en la corrección, según Oller (1973b), los resultados obtenidos en las pruebas integradoras correlacionan muy bien con la valoración del profesor y con otras pruebas. Esto se debe a que reflejan las situaciones reales de comunicación mejor que las pruebas de elementos discretos.

> ...tend to correlate better with teacher judgements, better among themselves, and better with other measures of language skills than do any of the discrete-point types because they more nearly reflect what people actually do when they use language. (Oller 1973b: 198)

1.8.3.2.1. Pruebas pragmáticas

Oller (1979: 38) define las pruebas pragmáticas, dentro de las integradoras, como

> ...any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extralinguistic context.

Son las pruebas pragmáticas las que mejor reflejan la competencia lingüística del alumno (Oller 1979: 64).

A menudo se considera que el concepto de prueba integradora es sinónimo de pragmática. Según el autor, la confusión viene dada porque las pruebas integradoras pueden ser pragmáticas, cuando la tarea lingüística que pide la prueba se relaciona de forma significativa con el contexto extralingüístico, y las pruebas pragmáticas son siempre integradoras.

De nuevo el dictado, las redacciones, narraciones, entrevistas orales, etc. y las pruebas de cierre son ejemplos de prueba pragmática que cita Oller (1979). Así pues, clasificamos al C-test como prueba integradora y pragmática.

Aunque dedicamos un capítulo completo a revisar en profundidad las características de las pruebas de cierre y sus tipos, adelantamos algunos rasgos que las hacen pruebas pragmáticas.

Las pruebas de cierre nacieron con Taylor (1953). Inventó el término "*cloze*" para designar a los exámenes que demandan completar un texto en el que se han omitido previamente determinados elementos. La tarea que estas pruebas plantean al alumno es semejante a un problema gestaltiano de "*closure*". Para resolverlo, el alumno se ve obligado a hacer inferencias de todo tipo (lingüísticas y extralingüísticas). Debe poner en funcionamiento su "gramática pragmática de expectativas" y su actuación en la prueba muestra "the efficiency of the learner's

developing grammatical system" (Oller 1979: 44). Como veremos, el C-test comparte los rasgos descritos.

1.8.4. Según el método de corrección

Atendiendo al método seguido en la corrección de las pruebas diferenciamos entre las de tipo objetivo y las subjetivas. Veremos que cierta subjetividad es inherente a las pruebas, aunque sea sólo en su diseño, y que la fiabilidad no es rasgo exclusivo de las pruebas objetivas.

1.8.4.1. Pruebas objetivas

Las pruebas objetivas son las que evitan o limitan al máximo la opinión del corrector. Ya Pilliner (1968) diferenciaba las pruebas subjetivas de las objetivas en términos del procedimiento de corrección. Un test objetivo es, en principio, fiable. Sin embargo, aún en las pruebas cuyo método de corrección es objetivo hay un margen para la subjetividad del profesor en el resto de las tareas, ya que el diseño de la prueba requiere siempre la toma de decisiones ineludibles: contenido, formato, texto, tema, etc.

Al igual que el resto de las pruebas de cierre, el C-test es reconocido como prueba objetiva. Su corrección no implica el juicio del corrector y su diseño reduce la subjetividad a la elección del texto y punto de comienzo de las omisiones.

1.8.4.2. Pruebas subjetivas

Cuando se requiere la emisión de juicios por parte del corrector, las pruebas se consideran subjetivas. El ensayo o redacción es una de ellas.

En las redacciones es casi imposible prescindir hasta cierto punto del juicio del corrector y sin embargo, aun así, se pueden lograr correcciones fiables, tanto si se adoptan enfoques analíticos como holísticos (Bacha 2001; Amengual 2003).

1.8.5. Según el marco de referencia utilizado

Teniendo en cuenta este criterio, las pruebas pueden clasificarse en: "*norm-referenced*" cuando medimos la actuación del alumno con respecto al grupo y "*criterion-referenced*" si consideramos el logro de los objetivos propuestos independientemente de la situación del grupo. Bachman (1990) insiste en que las pruebas normativas y las criteriales no se excluyen entre sí.

1.8.5.1. Pruebas normativas

Tomamos la definición que hace Bachman (1990: 72): "Norm-referenced tests (NR) are designed to enable the test user to make "normative" interpretations of results". Si una prueba normativa está bien diseñada los resultados estadísticos mostrarán unas características constantes. La distribución de las puntuaciones seguirá una curva típica. Otros datos de referencia interpretables son la media, la moda y la desviación típica[17].

Las pruebas normativas van orientadas a la discriminación del nivel del alumno en la lengua. Las pruebas estandarizadas (como las PAAU) son el prototipo de prueba normativa. Se basan en un contenido fijo e invariable de uno a otro examen. Se administran y corrigen siguiendo unos criterios previamente fijados. Además sus características se conocen bien, pues se han probado previamente en investigaciones y estudios piloto. Su validez y fiabilidad están aseguradas y demostradas empíricamente con grupos semejantes.

1.8.5.2. Pruebas criteriales

Como con las normativas, comenzamos con la definición que ofrece Bachman (1990: 74). Según sus palabras: "Criterion-referenced (CR) tests are designed to

[17] En la Perspectiva Empírica trabajamos aplicando estos conceptos estadísticos al análisis de las pruebas.

enable the test user to interpret a test score with reference to a criterion level of ability or domain of content".

Para una prueba criterial es prioritario especificar el nivel de habilidad o dominio del contenido. Debe ser "sensitive to levels of ability or degrees of mastery of the different components of that domain". Los resultados obtenidos en ella se interpretan como indicadores del nivel de habilidad alcanzado.

El propósito de las pruebas criteriales es clasificatorio: "to classify people according to whether or not they are able to perform some task or set of tasks satisfactorily" (Hughes 1989: 18).

1.8.6. Según su ámbito de aplicación y consecuencias

Podemos también diferenciar las pruebas teniendo en cuenta su ámbito de aplicación y las consecuencias que se derivan de la actuación en ellas. Nos fijaremos en dos extremos, por una parte las pruebas que se aplican regularmente en el aula y, por otra, las que se hacen a gran escala o *high-stakes tests*.

1.8.6.1. Pruebas de aula

Las pruebas de evaluación o control que se realizan en el aula no suelen necesitar un gran despliegue de recursos físicos ni materiales. El propio profesor es el que cubre todas las fases del proceso, desde el diseño hasta la aplicación y corrección. Y el propósito de las mismas generalmente es verificar el logro de los objetivos programados para un determinado período de tiempo.

1.8.6.2. Pruebas a gran escala

En el caso de los *high stakes tests* o pruebas estandarizadas a gran escala se requiere una planificación detallada y más recursos de todo tipo. Las diferencias radican en el ámbito de aplicación de la prueba, que suele administrarse a gran

número de examinandos, y en su propósito, que a veces permite o impide la consecución de un determinado título, diploma o posición.

Ambos, ámbito y finalidad, determinan que las circunstancias de la administración y corrección de estas pruebas sean totalmente diferentes.

Ya desde su diseño, las pruebas estandarizadas implican la participación de equipos de profesores y/o expertos, a menudo con fuertes inversiones de tiempo y dinero. Hay que mencionar, además, su repercusión social. En España, las pruebas de Selectividad o PAAU constituyen el mejor ejemplo de este tipo de prueba.

El C-test se inscribe en las pruebas integradoras y pragmáticas. Pretende medir el constructo de la competencia lingüística general. Por su naturaleza es una prueba indirecta. En cuanto al modo de corrección, es objetiva. Y, por sus características, puede utilizarse en cualquiera de las situaciones o momentos del proceso de aprendizaje, tanto como prueba de logro, de competencia lingüística o de nivel. Sólo resulta poco adecuada como prueba de diagnóstico.

Como veremos en la parte experimental de este trabajo, la hemos aplicado en el contexto de las clases regulares de Inglés de Bachillerato, pero pensamos que también podría formar parte de pruebas estandarizadas, como las PAAU, para las que supondría una valiosa aportación.

CAPÍTULO 2. PERSPECTIVA HISTÓRICA DE LA EVALUACIÓN DE LA LENGUA

2.1. Introducción

En este capítulo haremos un recorrido histórico por la Evaluación de la Lengua. Para empezar revisaremos brevemente los orígenes de la Lingüística Aplicada, como disciplina que nace a partir de la necesidad de aplicar soluciones que partan de la Lingüística a los problemas que plantea el auge de la enseñanza de idiomas, y en cuyo seno se inserta la subdisciplina de la Evaluación de Lenguas.

Continuaremos con un repaso somero de los principales enfoques de LT a lo largo del siglo XX y en los albores del siglo XXI. Dedicaremos especial atención al movimiento comunicativo, por su repercusión en la enseñanza de lenguas. Después revisaremos los pasos de la disciplina a partir de los años 80, para centrarnos en la última década de LT.

Dada la influencia de ciertas publicaciones periódicas especializadas (*Language Testing, Language Learning, Language Teaching Abstracts*, *System*, etc.) en el desarrollo de la disciplina, sobre todo durante los últimos años, dedicamos la parte más extensa del capítulo al estudio de las últimas tendencias de LT a partir de los artículos de investigación más recientes. Finalmente, tomando como referencia el cambio de milenio, terminaremos aventurándonos a apuntar el rumbo que podría tomar la disciplina en el presente siglo.

2.2. Orígenes de la Lingüística Aplicada

Esta disciplina comenzó con el objetivo de mejorar la enseñanza de idiomas extranjeros a partir de problemas concretos para los que se buscaba una solución

práctica. El estudio de las distintas propuestas dio paso a la formulación de teorías y al desarrollo de conceptos que, poco a poco, permitieron conocer mejor la lengua, sus mecanismos, la relación L1-L2, etc.

Brumfit (2001 citado en Bygate 2004) define la Lingüística Aplicada como "the theoretical and empirical study of real world problems in which language plays a central role".

Es interesante el recorrido histórico de Catford (1998) en busca de los orígenes de la Lingüística Aplicada. Aunque en sentido amplio podríamos decir que la Lingüística Aplicada a la enseñanza de lenguas se remonta a la Antigüedad, es en el siglo XIX cuando encontramos referencias explícitas a esta ciencia. En ese momento se siente la necesidad de unificar criterios en la enseñanza de idiomas. Catford alude a un grupo de lingüistas, entre los que destacaba Jespersen, que se reunieron en 1886 para diseñar un programa que guiara la enseñanza de lenguas. En 1899 Sweet publicó *The Practical Study of Languages*, que le valió la consideración de "padre" de la Lingüística Aplicada. Pero algo antes, Jan Baudouin de Courtenay (1870) había hecho ya por primera vez la distinción entre lingüística pura y aplicada. Baudouin (1904) expresaba: "Ever greater importance must be attached to the application of linguistics to didactics, both in the teaching and learning of foreign languages".

Después de estos inicios, en la década de los 40, habría que destacar algunos acontecimientos que impulsaron el desarrollo de la Lingüística Aplicada, como la creación del *English Language Institute* (ELI) en la Universidad de Michigan (1941), impulsado por Fries, y la aparición de la revista *Language Learning: A Quarterly Journal of Applied Linguistics* (1948). En el ELI enseñaron autores muy significativos, como Fries, Lado, Wallace, etc. La revista *Language Learning*, por su parte, se convirtió en vehículo de experiencias y foro de la investigación en la materia.

A partir de entonces aumentó el número de publicaciones, instituciones, asociaciones y congresos relacionados con la disciplina. Citamos sólo dos ejemplos: la fundación de la *School of Applied Linguistics* de la Universidad de Edinburgo (1957) y la creación de la *Association Internationale de Linguistique Appliquée* (AILA) en 1964 (véase el apartado 2.3.3.2). En épocas más recientes debemos mencionar, entre otros acontecimientos, la creación del *English Language Testing*

Service (ELTS) por el British Council y la aparición de la revista *Applied Linguistics* (1980).

Dentro de la Lingüística Aplicada centramos nuestro estudio en el campo de la Evaluación de la Lengua. Desde sus comienzos, la Lingüística Aplicada se ha ocupado de la evaluación, prueba de ello es la publicación, ya en 1968, de un número especial monográfico dedicado a *Problems in Foreign Language Testing* en la revista *Language Learning.*

Sin embargo, hasta 1984 no aparece una publicación específica sobre Evaluación de la Lengua. *Language Testing* es la publicación que llenó ese vacío. Un seguimiento exhaustivo de lo publicado en ella nos permite tomar el pulso a la disciplina en los últimos veinte años.

La Lingüística Aplicada está alcanzando en nuestros días su plena madurez: "Language Testing has *"come of Age"*", en palabras de Alderson (1995). También Bialystok (1998) manifiesta que la disciplina vive ahora su momento más rico e integrador. En Europa su desarrollo no permanece ajeno a los cambios sociales; como el final de la guerra fría, la creación de la Unión Europea, los cambios en las fronteras, el crecimiento económico de la Europa Occidental, las migraciones del sur al norte de Europa, el regionalismo emergente, etc. (Kees de Bot 2004: 57).

A lo largo de la historia podemos comprobar que las soluciones que han hecho avanzar a la ciencia se lograron gracias a la colaboración entre disciplinas. Es ésa la vía que proponen especialistas como Bialystok (1998), y que respaldamos desde este trabajo, para que el futuro de la Lingüística Aplicada sea de verdad clarificador y productivo.

2.3. La Evaluación de la Lengua: trayectoria histórica

Tanto los profesionales de la docencia como los especialistas en Lingüística Aplicada muestran un reconocimiento unánime de la importancia de la evaluación como parte integrante del proceso de enseñanza-aprendizaje. Todo proceso de enseñanza-aprendizaje necesita de la evaluación, como apuntamos en el capítulo 1. La evaluación forma parte de él e influye en él. Por ello, es conveniente revisar las

distintas técnicas e instrumentos de evaluación para asegurarnos de que afectan de forma positiva a la enseñanza.

Teniendo como base diferentes conceptos sobre la naturaleza de la evaluación dentro del proceso de enseñanza-aprendizaje, con el tiempo, la práctica docente ha ido adoptando distintas formas de evaluación. El objetivo final es siempre el aprendizaje de la lengua, pero para lograrlo, cada profesor debe "adaptar" en cierta medida los métodos propuestos por los especialistas a su contexto, peculiaridades, necesidades, etc.

El modelo psicométrico tradicional asumía que los tests habían de ser universales o iguales para todos los individuos, estandarizados y unidimensionales, pues cada ítem debía medir una sola destreza. Los exámenes servían para ordenar a los alumnos según los resultados obtenidos. Sin embargo, debido a sus limitaciones, este modelo se empezó a cuestionar a partir de la publicación de *Taxonomy of Educational Objectives* de Bloom (1950).

2.3.1 El movimiento estructuralista

Los años 50 y 60 vieron el florecimiento de la visión conductista estructuralista de la lengua, cuyo exponente más notable en evaluación es Lado, con su obra *Language Testing* (1961).

El movimiento estructuralista entiende que el lenguaje se estructura a partir de distintos componentes (fonológico, lexicosemántico y morfosintáctico) y unidades. Para el análisis de la lengua propone un enfoque científico que propicia los estudios contrastivos y el uso de instrumentos estadísticos.

Los estructuralistas consideran la competencia lingüística como suma de cuatro habilidades o destrezas (*listening, speaking, reading* y *writing*) y sus componentes (gramática, vocabulario, pronunciación). Esta aproximación a la lengua afecta al enfoque de su evaluación. El diseño de pruebas se basa en aspectos concretos y aislados del idioma. Así, Lado (1961) recomienda las pruebas de tipo objetivo, de elementos discretos.

2.3.2. El movimiento comunicativo

Destacamos el enfoque comunicativo por su contribución a la enseñanza de lenguas extranjeras y su gran repercusión en LT.

A finales de los 70 se empieza a entender el lenguaje como competencia comunicativa (Widdowson 1978; Canale y Swain 1980) y a tener más en cuenta los aspectos sociales del lenguaje. Se incrementa el interés por la enseñanza y evaluación de la lengua "real".

El marco teórico de competencia comunicativa ideado por Canale y Swain (1980) distinguía tres tipos de competencia: gramatical, sociolingüística y estratégica. Este marco ha planteado retos a los lingüistas y en él se ha basado gran parte de la investigación posterior sobre Evaluación de Lenguas.

La base del movimiento comunicativo es considerar que la competencia en una lengua no viene definida exclusivamente por los conocimientos gramaticales, sino por la capacidad de comunicarse en ella (Pica 2000). El aprendizaje de idiomas adquiere sentido precisamente porque posibilita la comunicación.

El método comunicativo supuso un gran cambio. Su principal novedad radica en que se centra en el alumno y sus necesidades. Cambia así el rol del profesor, que pasa de ser el centro a ser simplemente el que facilita el proceso de aprendizaje.

Este nuevo enfoque metodológico implicó el nacimiento de la evaluación de la competencia comunicativa o *Communicative Language Testing* (CLT) como reacción a las pruebas de tipo objetivo propuestas por Lado (1961).

Es preciso evaluar tanto la competencia lingüística como la actuación o producción en situaciones concretas. Con estas premisas como base se han desarrollado diversos modelos de evaluación, como el propuesto por Bachman (1990) y Bachman y Palmer (1996): *Communicative Language Ability* (CLA). Aludimos al modelo de Bachman en el capítulo 1 de la tesis, apartado 1.5.

La evaluación comunicativa aporta nuevas perspectivas; surge, por ejemplo, la preocupación por la autenticidad de las pruebas, aspecto que aún hoy suscita múltiples interrogantes.

Distintos autores fijaron las características de las pruebas comunicativas (Swain 1984; Davies 1990; Rea 1991), y algunos otros cuestionaron que se pueda

hacer una CLT adecuada, dada la dificultad para extrapolar los resultados de las pruebas comunicativas (Skehan 1988; Weir 1983; Morrow 1977).

Davies (1982: 149) expresaba su propuesta de incluir aspectos gramaticales en las pruebas comunicativas para lograr neutralizar la tensión entre validez y fiabilidad.

> The most useful tests are probably those that make a compromise, i.e. tests that make up on Reliability by testing linguistic competence through discrete point items, and make up on Validity by testing communicative competence through integrative items.

En esta línea se manifiesta Pica (2000: 15) cuando aboga por la integración de métodos tradicionales y comunicativos para suplir las deficiencias de ambos y mejorar la enseñanza de idiomas.

> Recent findings on the cognitive, social, and linguistic processes of L2 learning have suggested a principled approach to L2 instruction. Such principles are characterized by classroom strategies, participant structures, and activities which incorporate traditional approaches, and reconcile them with communicative practices.

2.3.3. La evaluación en las últimas décadas: estado de la cuestión

A partir de los años 80 LT se fue consolidando como disciplina, se amplió la preocupación por la evaluación de lenguas y surgieron nuevas técnicas. En primer lugar hay que mencionar la extensión que cobró este campo y su actividad constante a nivel internacional. Desde entonces hasta nuestros días se ha convertido en un área muy productiva, lo que hace difícil recoger todos los intentos de avanzar en evaluación de la lengua.

Continuamente aparecen estudios que plantean aspectos aún no suficientemente claros para la evaluación, como el *washback* o efecto rebote. Otros responden a preocupaciones nuevas, como la ética y la política educativa desde la perspectiva de la evaluación, o la aplicación de las nuevas tecnologías.

La mayor conciencia de los aspectos éticos ha llevado incluso a plantear la necesidad de una profesionalización en el campo de la evaluación.

Se camina hacia una colaboración entre disciplinas, como sugería Bialystok (1998). Así, por ejemplo, los avances en el campo de *Second Language Acquisition* (SLA) han permitido identificar distintos niveles de dominio de la lengua y esto ha abierto nuevas vías de investigación. Los métodos de investigación utilizados en la actualidad abarcan e integran técnicas cualitativas y cuantitativas. Destaca la actividad de varios centros; como las Universidades de Michigan, Edimburgo, etc.

En nuestro país, hay que mencionar el interés creciente que despiertan los estudios sobre Lingüística Aplicada[18]. La preocupación por la educación y la política educativa se constata en el aumento del número de tesis, proyectos y/o estudios promovidos por iniciativa pública o privada, por distintos programas de la Unión Europea, etc. A pesar de todo, todavía los estudios sobre Evaluación de la Lengua en España despliegan menor actividad que en otros países de nuestro entorno, sobre todo con respecto a otras áreas de la enseñanza de lenguas. Aunque no abundan los congresos dedicados exclusivamente a la evaluación de la lengua, sí se realizan aportaciones notables en los de ámbito internacional, como los de ALTE (*Association of Language Testers in Europe*).

Es significativo constatar la ausencia de un panel específico sobre evaluación en los congresos anuales de la Asociación Española de Lingüística Aplicada (AESLA), a la que nos referiremos en el apartado 2.3.3.2. No obstante, los trabajos relacionados con la evaluación se incluyen en otros paneles, generalmente en el dedicado a Enseñanza de Lenguas y Diseño Curricular.

2.3.3.1. Publicaciones especializadas en Evaluación de la Lengua

Ya desde la introducción de este capítulo hemos expresado que merece mención expresa la labor de distintas publicaciones que periódicamente recogen y divulgan las últimas aportaciones de los especialistas en Evaluación en la enseñanza de lenguas[19].

[18] Véase el informe de Graeme Porte (2003) para *Language Teaching Abstracts* sobre las investigaciones más recientes realizadas en España de 1999 a 2002 en el campo de la Lingüística Aplicada. El autor destaca el nuevo entusiasmo en la materia y la profusión de publicaciones. Hace, además, referencia explícita a los trabajos sobre validez y fiabilidad de las pruebas de Selectividad (Herrera, Esteban y Amengual 2001; Amengual y Herrera 2001).

[19] Actualmente la mayoría de estas publicaciones tiene una versión electrónica disponible en Internet.

Entre las revistas de mayor prestigio en el ámbito lingüístico educativo se encuentran *Language Testing*, *Language Learning*, *Language Teaching Abstracts*, *Applied Linguistics*, *ESP English for Specific Purposes*, *System*, *AILA Review*, *IRAL International Review of Applied Linguistics*, *TESOL Quarterly*, etc. Todas ellas desarrollan una labor importante en Lingüística Aplicada.

Language Testing es la publicación que se dedica de forma más directa al mundo de la Evaluación de Lenguas, pero todas las mencionadas son de gran interés puesto que, a pesar de centrarse en otros aspectos de la enseñanza de la lengua, publican también con mayor o menor frecuencia artículos relacionados con la evaluación. El análisis pormenorizado de su trayectoria nos dará el pulso del panorama actual, del estado de la cuestión en Evaluación de la Lengua.

2.3.3.2. Asociaciones

En el campo de la Evaluación de la Lengua es destacable la labor de diversas asociaciones. Algunas no se centran en la evaluación, sino que abarcan un ámbito mayor: el de la enseñanza de lenguas.

En los años 60 (Nancy, 1964) se creó la *Association Internationale de Linguistique Appliquée* (AILA), integrada por lingüistas europeos y profesores de idiomas. Coincidió con un momento de florecimiento de la Lingüística Aplicada y su creación supuso un gran estímulo para la actividad investigadora.

A partir de 1969 se celebran congresos de la Asociación cada tres años. Hoy acoge a más de 5000 socios de 43 asociaciones nacionales y se ha convertido en una vigorosa organización mundial. El volumen 17 de la revista *AILA Review* (2004) celebró los 40 años de la Asociación con una visión del panorama que ofrece la disciplina hoy: *World Applied Linguistics*. Colaboraciones de Bygate, Cavalcanti, Grabe, Kees de Bot, Kleinsasser, Pakir y Valdman, entre otros, muestran que la Lingüística Aplicada sigue su camino adaptándose a las circunstancias del mundo actual y afrontando nuevos retos.

Además hemos de mencionar la aparición de la *International Language Testing Association* (ILTA). Su manifiesta preocupación por la ética en la evaluación se refleja en el “Code of Ethics for ILTA”, adoptado en Vancouver (2000). En junio de

2007 se ha celebrado el 29th *Annual Language Testing Research Colloquium* (LTRC) de ILTA en Barcelona.

En 2004 se creó una nueva asociación de ámbito europeo, la *European Association for Language Testing and Assessment* (EALTA), cuyo objetivo citamos textualmente: "to promote the understanding of theoretical principles of language testing and assessment, and the improvement and sharing of testing and assessment practices through Europe". Celebró su segunda Conferencia Anual en junio de 2005 (Voss, Noruega).

En España, destaca la labor que desarrolla la Asociación Española de Lingüística Aplicada (AESLA), creada en 1982 y afiliada a AILA desde 1984. Esta asociación dio respuesta a la necesidad de contar con una organización española estable para difundir e impulsar las inquietudes que surgieran en el campo de la Lingüística Aplicada.

Desde su creación, AESLA celebra un congreso anual en el que ofrece a profesores, investigadores y demás personas e instituciones interesadas en la enseñanza de la lengua, la posibilidad de poner en común sus ideas y conocer las últimas tendencias en evaluación, de la mano de especialistas mundialmente reconocidos. Basta citar a algunos de los prestigiosos lingüistas invitados a los últimos Congresos de AESLA (XXIII, XXIV y XXV): Barndern, Chapelle, Cook, Downing, Edmondson, Ellis, Selinker, Turell, Lantoff, Kövecses, Steen, Llisterri, Wachs, Muñoz Liceras, Faber, Escandell, Pascual, Hyland, Meyer, Teubert, etc.

También dispone de un servicio de publicaciones que recoge las aportaciones y las divulga entre la comunidad científica, entre otras formas, mediante la revista RESLA y, desde 2003, con la revista electrónica RæL[20].

2.4. La Evaluación de la Lengua de 1984 a 1994: *State of the Art*

Como ya hemos mencionado en el apartado anterior, *Language Testing* es la publicación que más directamente se ocupa del campo de la Evaluación. De aparición trimestral, esta revista, cuyos editores son D. Douglas (Iowa State University) y J. Read (Victoria University of Wellington), cuenta con el apoyo y la

[20] Para más información, recomendamos visitar la página web de AESLA en http://www.aesla.uji.es.

colaboración de los especialistas en evaluación más importantes en el panorama mundial, como son J. C. Alderson, L. F. Bachman, M. Chalhoub-Deville, A. Cummming, A. Davies, G. Fulcher, P. Meara y un largo etcétera.

El primer número se publicó en 1984. Hoy *Language Testing* cuenta ya con más de dos décadas de vida. Su política editorial es clara: publica artículos teóricos o prácticos relacionados con la evaluación de segunda lengua y lengua extranjera, lengua materna, problemas y disfunciones lingüísticas, y proyectos o programas lingüísticos con implicaciones teóricas en la evaluación de lenguas. Atentos a todas las novedades en el campo de la evaluación de lenguas, cada volumen incluye además reseñas de libros de interés realizadas por colaboradores de la revista.

Dado el ámbito que abarca esta publicación, podemos decir sin miedo a equivocarnos, que el análisis de lo publicado en *Language Testing* durante su primera década de vida (1984-1994) nos permite conocer en profundidad el desarrollo de la investigación mundial en el campo de la Evaluación de la Lengua a partir de los años 80.

Siguiendo esta idea, Herrera Soler (1997: 116) revisó los primeros diez años de historia de *Language Testing*. En su estudio clasifica los artículos publicados en la revista durante ese periodo y destaca las aportaciones más interesantes que de ellos se derivan para la práctica docente: “A substantial number of the problems we have to cope with in our classroom every day have been dealt with in LT pages, and not a few answers to our difficulties can be found”.

El año 1984, fecha de lanzamiento de *Language Testing*, supuso un momento crucial en la historia de la evaluación de lenguas. Entre los artículos publicados ese primer año aparecían ya los temas más importantes que han seguido siendo objeto de investigación en años posteriores; como las características de toda prueba de evaluación (validez, fiabilidad, autenticidad), los métodos de evaluación (*criterion-referenced, norm-referenced*), exámenes oficiales estandarizados como el TOEFL, o nuevas teorías (IRT como alternativa a los métodos tradicionales) y técnicas de evaluación (*cloze, C-test, multiple choice*). Así lo expresa Herrera Soler:

> The most frequently dealt with topics were acquisition of a second language, methods, testing strategies and certain issues in linguistic fields; not an issue went by without an article on the Item Response Theory (IRT) and on the Testing of English as a Foreign Language (TOEFL) examination, the two main lines of research in LT.

El autor destaca la coherencia de la revista *Language Testing* y la riqueza de su contribución al campo de la Lingüística Aplicada:

> In our opinion then, LT has made a notable contribution over the last 10 years to the debate about how far language testing has gone toward understanding the abilities that teachers and institutions intend to measure. It is high time we, as teachers/testers of Second Language Acquisition or of Language for Specific Purposes, took advantage of these ten years of language testing research and that lamentations like the following one by Alderson (1988: 87) were progressively outdated: "It is rather sobering and perhaps depressing to note the minimal attention paid to testing..." (op. cit.: 132-133)

A continuación hacemos un breve recorrido histórico por la década. Para agrupar las distintas líneas de investigación en evaluación de la lengua desde 1984 a 1994, reflejada en los artículos de este periodo de *Language Testing*, seguimos la clasificación de Herrera (1997):

1. Teoría de respuesta al ítem (IRT).
2. Análisis de pruebas estandarizadas.
3. El problema de la autenticidad de las pruebas.
4. La autoevaluación.
5. La influencia de otros factores en la evaluación: el contexto y las características del alumno.
6. Las técnicas de examen.

2.4.1. Teoría de respuesta al ítem (IRT)

Uno de los modelos de evaluación más importantes fue la *Item Response Theory* o IRT de Rash, que para minimizar el error, calcula y relaciona el grado de dificultad de los distintos ítems y la capacidad o habilidad del individuo para resolverlos.

Abundan los artículos tanto de corte teórico como práctico sobre la aportación de esta teoría a la evaluación de lenguas. Se consideró una alternativa a los métodos tradicionales, pero los artículos publicados no sólo analizan sus ventajas sino también sus inconvenientes. En la década que nos ocupa destacan las

contribuciones teóricas de Henning *et al*. (1985 y 1989), Woods y Baker (1985), Carrol (1986), Adams *et al*. (1987), Tomlinson *et al.* (1988), Hudson (1993). Y las aportaciones prácticas en artículos que analizan el modelo IRT de Theunissen (1987), Jon y Glas (1987), Boldt's (1989), McNamara (1990, 1991), Choi y Bachman (1992) principalmente en tests de comprensión oral y escrita.

2.4.2. Análisis de pruebas estandarizadas

En segundo lugar, Herrera (1997) menciona toda una serie de artículos que se centran en investigaciones acerca de exámenes estandarizados, como el TOEFL. Son fundamentalmente prácticos, algunos suponen un intento de mejorar estas baterías, otros comparan el TOEFL con otras pruebas.

Parece claro que una buena batería de tests necesita una revisión y renovación permanente. Autores como Stansfield y Ross (1988) y Boldt (1989, 1992) pretenden con sus estudios mejorar las pruebas y analizar sus características de validez y fiabilidad. En 1990 Spolsky estudia la historia de este examen, sus orígenes y desarrollo. A menudo, las investigaciones van dirigidas hacia la comparación del TOEFL con otras baterías similares como el EFL *proficiency test*, el FCE, TSE, TWE, etc.

2.4.3. El problema de la autenticidad de las pruebas

Otro de los temas recurrentes que aparecen durante los primeros diez años de *Language Testing* es el de la autenticidad como característica fundamental de las pruebas, sobre todo de las de tipo comunicativo.

La autenticidad preocupa a los investigadores, puesto que a mayor grado de autenticidad de un test mayor probabilidad habrá de que éste sea representativo de la actuación lingüística del individuo y, a la vez, bien acogido por profesores y alumnos.

El movimiento comunicativo insiste en la utilización de materiales auténticos y la propuesta de tareas realistas. Pero los autores señalan que existe ya de partida

una contradicción entre un examen y el principio de autenticidad. Por el mero hecho de ser una prueba que mide unas determinadas actuaciones o destrezas, el examen no puede ser del todo auténtico.

En torno a este rasgo de las pruebas destacamos el artículo de Spolsky (1985) acerca de los límites de la autenticidad en la evaluación de la lengua.

No obstante, como veremos, el debate sobre el concepto y los límites de la autenticidad llega hasta nuestros días.

2.4.4. La autoevaluación

También la autoevaluación ha sido una preocupación constante de los estudiosos de la evaluación de la lengua. En 1989 se dedicaron varios artículos a este tema (Oscarson, Janseen, Bachman y Palmer). Hoy en día se dispone de múltiples instrumentos, con frecuencia mediante el uso de las nuevas tecnologías, para medir la propia competencia lingüística.

2.4.5. La influencia de otros factores en la evaluación: el contexto y las características del alumno

Por otra parte, durante esta década aparecieron investigaciones relacionadas con los distintos contextos culturales y con el componente afectivo: desde el estudio de Zeidner (1987) sobre la influencia de la raza, sexo y edad en la validez de una prueba, hasta el de Chihara *et al*. (1989), que se ocupaba del sesgo, es decir, del efecto que pueden producir los factores culturales en los *clozes*. También Zeidner y Bensoussan (1988) analizaron la actitud de los alumnos frente a pruebas orales y escritas, y Bradshaw (1989), en un test de nivel. Concluyeron que estos factores se deben tener en cuenta al preparar un test. Por otra parte, Scott (1996) evaluó otros factores como el formato, la limitación de tiempo, la familiarización con el tipo de examen, la ansiedad, etc. Algunos autores se fijaron incluso en el efecto de que los alumnos conozcan una estrategia de examen antes de enfrentarse a él (Allan 1992; Amer 1993).

2.4.6. Las técnicas de examen

Un grupo importante de los artículos publicados en *Language Testing* desde 1984 hasta 1994 se ocupa de las técnicas de examen. Nos centraremos en la controversia entre los distintos tipos de prueba de cierre: *cloze* tradicional, C-Test, LDP y *multiple choice*. Estos artículos dedicados a valorar las diferentes técnicas de examen suponen un punto de referencia fundamental para la investigación que se lleva a cabo en este trabajo.

Precisamente en el año 1984, fecha de aparición de la revista en que basamos nuestro análisis, Klein-Braley y Raatz publicaron en *Language Testing* algunas de sus primeras investigaciones acerca del C-test, en el artículo *A survey of research on the C-Test* (*LT* 1, 134).

También en 1984 se publicó en *Language Testing* una nota de Cohen, Segal y Weis sobre el C-test en hebreo, *The C-Test in Hebrew* (*LT* 1, 221). Poco antes, en 1981, Klein-Braley y Raatz habían introducido la técnica del C-test como posible remedio para evitar algunos de los inconvenientes de las pruebas de cierre.

Y estas primeras publicaciones de comienzos de los años 80 sobre el C-test y sus posibilidades no son el final, sino el comienzo de toda una serie de artículos relacionados con la técnica objeto de nuestro estudio. En 1985 Klein-Braley y Raatz continúan en la misma línea de investigación con la publicación de dos artículos: *A cloze-up on the C-Test: a study in the construct validation of authentic tests* (*LT* 2, 76) de Christine Klein-Braley y *Better theory for better tests?* (*LT* 2, 60) de Ulrich Raatz.

Klein-Braley (1985) compara el C-test con los *clozes* tradicionales, partiendo de la base de que ambos son pruebas pragmáticas (según la clasificación de Oller 1979) creadas a partir de materiales auténticos, y de redundancia reducida. Ambos tipos de test suponen la puesta en práctica de la misma teoría: la competencia general en la lengua se muestra en la actuación lingüística, en este caso cuando se es capaz de recuperar un texto cualquiera en el que previamente se han eliminado determinados elementos. A partir de un estudio de los "defectos" de los *clozes* Klein-Braley demuestra que el C-test es una prueba válida y técnicamente superior a ellos.

Raazt (1985), por su parte, plantea el problema de la validez de las pruebas y aplica el Classical Latent Additive Test Model, CLA Model, propuesto por

Moosbrugger y Müller en 1982, para demostrar de forma objetiva que las distintas partes o ítems del C-test son homogeneas.

Posteriormente en *Language Testing* continúan apareciendo artículos que investigan ésta u otras técnicas de examen. Kokotta (1988: 115-119) muestra un informe de su investigación con una técnica de examen cercana al C-test, *the letter-deletion procedure* (LDP). Esta nueva técnica muestra su flexibilidad con respecto a las pruebas de cierre tradicionales y al C-Test.

En el artículo *Cloze method: what difference does it make?*, Chapelle y Abraham (1990: 121-126) comparan distintas técnicas de evaluación de lenguas, como son los clozes de ratio fija y variable, el test de elección múltiple y el C-test

Poco después, Dörnyei y Katona (1992: 187-206) evalúan la validez del C-test con respecto al *cloze* tradicional entre estudiantes húngaros de Inglés como Lengua Extranjera. Sus conclusiones apuntan a que el C-test puede ser considerado un instrumento válido y fiable en la evaluación de lenguas.

Y en 1993, J. D. Brown estudia en profundidad las características de los *clozes* naturales y muestra, apoyándose en datos estadísticos, que los *clozes* tradicionales no siempre son tan fiables y válidos como se suponía. Propone la selección de los ítems para crear *clozes* a la medida, que realmente funcionen. Finalmente sugiere posibles líneas de investigación para posteriores estudios sobre la técnica.

Esta década fue, sin duda, la más productiva en investigaciones sobre el C-test. Por su relevancia, en capítulos posteriores volveremos a hacer alusión a los estudios mencionados en este apartado.

2.5. La evaluación de la lengua desde 1994 hasta nuestros días

2.5.1. Introducción y fuentes

Durante esta última década hemos sido testigos de la vitalidad que sigue caracterizando a la disciplina y además hemos vivido una fecha muy significativa: el cambio de milenio como punto de inflexión. Por tanto, a pesar de la proximidad, estamos en condiciones de hacer ya una revisión del estado de la cuestión en evaluación de la lengua desde 1994 hasta nuestros días.

El año 2000 animó a los expertos en evaluación a recapitular y mirar hacia el siglo que dejábamos atrás para buscar claves. Este año se inauguró en *LT* con un artículo especial de Lyle F. Bachman que revisaba el estado de la evaluación de lenguas ante el cambio de milenio: *Modern language testing at the end of the century: assuring that what we count counts.*

Bachman (2000) menciona en él los avances prácticos y teóricos, la variedad de enfoques y herramientas que se han usado en evaluación desde los años 80, las mejoras en las técnicas y formatos de examen, los aspectos éticos, etc. Pero lo más importante son las perspectivas de futuro: las claves, en su opinión, están en la profesionalización del campo de la evaluación y la investigación sobre validación.

Sin perder de vista el análisis de Bachman, para tener una visión general del desarrollo del campo de la evaluación durante la pasada década, tomamos como referencia principal dos publicaciones: *Language Testing* y *Language Teaching Abstracts.*

La primera por ser la que se ocupa casi exclusivamente de la evaluación de la lengua. La segunda por su característica de reseñar todo lo publicado en la materia. De ésta última nos interesa especialmente la revisión realizada por Alderson y Banerjee (2002-03), puesto que se centra en LT. El estudio de estas dos revistas nos parece un buen punto de partida para conocer los trabajos más recientes realizados en este campo y tomar el pulso a la actualidad en torno a la evaluación.

El año 1994 ponía punto final a la primera década de existencia de la revista *LanguageTesting* y por ello, al completar el décimo volumen, los editores revisaron la publicación y anunciaron sus intenciones de cara al futuro. Para iniciar la andadura de la segunda década, manifestaron su intención de seguir la misma filosofía y objetivos, incluso de ampliarlos:

> These aims continue to represent the present editors' views of the role of the journal. We reiterate them here while at the same time extending them. [...] No doubt assessment is often understood more widely than testing but the journal has always been interested in this wider view. (Editorial LT 1994)

Esta idea de ampliar el campo de las contribuciones de la revista pretende abarcar cualquier aspecto de la evaluación: los exámenes, su creación y funcionamiento, el análisis de datos de tests, la política educativa, SLA, etc. Pero sin

olvidar el aspecto más cercano y práctico de la evaluación, la investigación directa en el aula:

> We would like to publish more articles tackling issues in school-based education of children. *Tests in schools* are, and will continue to be, *the arena where most testing activity occurs* and the journal, in acknowledging that importance, seeks submissions reporting on that arena. (Editorial LT 1994) (el énfasis es mío)

Language Teaching Abstracts, por su parte*,* es una publicación especializada en enseñanza de idiomas. Recoge los *abstracts* de todos los artículos que van apareciendo en el ámbito de la enseñanza-aprendizaje de lenguas y los clasifica en distintos epígrafes, uno de los cuales es *Language Testing*. La revisión monográfica de Alderson y Banerjee (2001-2002) sobre *Language testing and assessment* no es más que una de las muchas que ofrece la revista en momentos puntuales, si bien supone para nuestro estudio un punto de referencia obligado.

2.5.2. Principales temas que plantea la Evaluación de la Lengua en los últimos años

Es innegable que el campo de la evaluación en la enseñanza de idiomas continúa en pleno desarrollo. Las investigaciones actuales abordan múltiples aspectos de la evaluación, como el impacto de las pruebas, y temas novedosos como la ética y las nuevas tecnologías. Aparecen también otros más tradicionales, como la teoría de la validez de las pruebas y la investigación de los distintos constructos o destrezas que subyacen en ellas.

Alderson y Banejee (2002) consideran que en buen número de los trabajos actuales están presentes de forma directa o indirecta algunas cuestiones que todavía preocupan a los especialistas en evaluación: la autenticidad, el diseño de tests de lenguas y la tensión entre validez y fiabilidad. También en esta tesis se abordan estos temas más adelante.

La mayor parte de los artículos que se publican hoy son de carácter experimental y pretenden corroborar de forma empírica los supuestos teóricos planteados. Para ello se basan en casos concretos y analizan los resultados con la ayuda de medios estadísticos. En general, los estudiosos intentan apoyarse en

datos de carácter empírico para justificar aspectos teóricos y aportar después conclusiones e implicaciones pedagógicas aplicables en el aula.

Algunos temas que en la década anterior fueron verdaderos hilos conductores de la actualidad en evaluación han perdido parte de su protagonismo en los años más recientes, como es el caso de la *Item Response Theory* (IRT) y del examen de TOEFL. No obstante, se sigue investigando sobre la IRT en otros ámbitos, como evidencian los últimos congresos de AESLA. Tampoco la autoevaluación ni los trabajos sobre autenticidad han tenido tanto peso en los últimos años como en la década inicial de *LT*.

Sin embargo, otros aspectos de la evaluación han aparecido con mayor frecuencia en estos años, por ejemplo los estudios de lingüística clínica, o los basados en la inmersión y en los problemas lingüísticos que se generan en la integración de colectivos de inmigrantes. Los estudios sobre lenguas e inmigración responden a la nueva realidad que se vive tanto en España como en los países de nuestro entorno. La Lingüística Aplicada tampoco puede ser ajena a las nuevas demandas sociales.

También observamos un aumento del número de artículos dedicados al estudio de la evaluación de las destrezas orales. Sin olvidar los de corte teórico, hay que destacar el gran volumen de artículos de carácter empírico sobre la evaluación de la comprensión y expresión oral, muchos de los cuales pretenden validar exámenes estandarizados. No en vano la tendencia manifestada por la legislación educativa española es la de potenciar las destrezas orales en la enseñanza de lenguas extranjeras y su evaluación iba a formar parte de la nueva PGB, propuesta en la LOCE. Como se comentó en el capítulo 1, apartado 1.3.2, estos planes quedaron paralizados y sustituidos por la actual LOE (2006), aún pendiente de completar su desarrollo.

Se echa de menos la existencia de artículos de tipo histórico. De los pocos que aparecen destacamos la revisión histórica y de política educativa que hizo Spolsky (1995), el estudio de Hawthorne (1997) acerca de la situación de la evaluación del Inglés en Australia y los condicionantes de tipo político, y el recorrido histórico por el campo concreto de la evaluación en ESP de Davies (2001).

Parecería lógico que en la era de la tecnología de la información abundaran propuestas de aplicación de programas informáticos a la Lingüística. Sin embargo, el

número de artículos no ha sido en este periodo tan numeroso como cabría esperar. Hemos de destacar el artículo de Alderson *et al*. (2000), que alude a un programa para evaluar pruebas sobre coherencia textual, y la revisión *Issues in computer adaptive Testing of reading proficiency: Selected papers*, de Chalhoub-Deville y Fulcher (2000).

Este aspecto nuevo de la Lingüística Aplicada en relación con la informática está en pleno proceso de desarrollo y suponemos que avanzará notablemente en un futuro próximo. La publicación específica *Computer Assisted Language Learning* (CALL) recoge esos avances, de la mano de autores como Decoo (2003), Taylor y Gitsaki (2003), Pennington (2004), Gruba (2004), Zapata (2004), etc. Sin embargo, no se centra en la evaluación, sino en cómo los medios técnicos pueden apoyar en el proceso de aprendizaje de lenguas, siempre al servicio del método elegido por el profesor.

Para concretar y centrar esta revisión, podemos decir que la investigación en LT se dirige actualmente hacia tres grandes direcciones que estudiamos con mayor detenimiento a continuación:

- La profundización en los rasgos de las pruebas.
- El estudio de los distintos tipos de tests.
- La búsqueda de soluciones para los problemas y retos que el mundo actual plantea a la disciplina.

2.5.3. Rasgos de las pruebas

Una gran parte de la investigación que se realiza actualmente en todo el mundo relacionada con la Evaluación de la Lengua tiene por objeto definir con claridad los rasgos de los exámenes. Destacamos los estudios sobre el impacto de las pruebas, los relacionados con los conceptos de validez y fiabilidad, y los que tratan sobre la autenticidad. En ellos se centra la preocupación de los investigadores, si tenemos en cuenta el volumen de los trabajos realizados.

2.5.3.1 *Washback*, impacto o efecto rebote

El impacto que producen los exámenes en la enseñanza y el aprendizaje se conoce con el nombre de efecto rebote o *washback*. A pesar de que se reconocía la existencia del *washback*, hasta fechas recientes no abundaban los estudios empíricos sobre este fenómeno. Además, a menudo se consideraban sólo los efectos negativos, pero la influencia de las pruebas puede ser tanto positiva como negativa. Alderson y Wall (1993) describían el término *washback* como neutral. El reto para los estudiosos del tema es el diseño de pruebas que produzcan el deseado impacto positivo (Bailey 1996).

En los últimos años está aumentando considerablemente el interés por este tema. Ya en 1993 Alderson y Wall sugerían varias vías de investigación del *washback* para determinar, por ejemplo, en qué aspectos influye: si afecta a lo que se enseña (contenidos, currículo) o también a cómo se enseña (metodología). Los estudios coinciden en señalar que las pruebas, sobre todo las más significativas o estandarizadas, realmente influyen en los contenidos que se enseñan y en los materiales utilizados. Es más complicado demostrar cómo influyen en la metodología del profesor e incluso en la motivación del alumno. Cheng (1997) demuestra que estos cambios metodológicos se producen de forma más lenta.

El tercer volumen de 1996 de *Language Testing* es un monográfico sobre el impacto de los exámenes en la enseñanza. Recoge artículos teóricos centrados en esta característica de los tests (Shohamy *et al*. 1996); desde la revisión histórica de este concepto en la evaluación (Bailey 1996) hasta el estudio de Messick (1996) acerca de los conceptos de validez y *washback*, pasando por los análisis prácticos de Wall (1996) y Watanabe (1996), junto al de Alderson y Hamp-Lyons (1996) sobre el *washback* en los cursos de preparación del TOEFL. Posteriormente, Hamp-Lyons (1997) retoma el tema y lo aborda desde el punto de vista ético.

El impacto es un tema amplio, rico y complejo abierto a nuevas investigaciones. Dedicamos el apartado 8 del capítulo 3 al estudio de este rasgo de las pruebas.

2.5.3.2. Fiabilidad y validez

Generalmente estos dos conceptos se consideran complementarios, una prueba ha de ser fiable para ser válida. A menudo se mezclan y los límites entre ellos se desdibujan. Alderson y Banerjee (2002: 102) desdramatizan este hecho: "In effect, this means that we need not agonise [...] over whether what we call reliability is "actually" validity".

La visión del concepto de validez como característica de las pruebas está cambiando y la distinción entre validez y fiabilidad es ya, según Alderson, "irrelevante". Sin embargo, sigue vigente la preocupación por cómo validar las pruebas de evaluación de la lengua.

Muchos autores son pesimistas en cuanto a la posibilidad de hacer una buena validación, no obstante, gran parte de las investigaciones actuales tiene por objeto la validación de distintas pruebas. A lo largo de la última década podemos destacar el estudio teórico de Messick (1996) sobre validez e impacto, el de Alastair (1994) entendiendo la validez como concepto unitario pero con múltiples facetas (Messick 1989), y los trabajos de Davidson (1994) acerca de la validez de una prueba estándar de carácter normativo (NRM, *norm-referenced measurement*).

Shohamy (1994), a partir de su estudio de la validez de dos pruebas orales, una directa y otra semi-directa, concluyó que los tests deben validarse desde varias perspectivas.

Para Scott *et al*. (1996), que estudiaron la validez en el examen de resumen de audiciones LSTE- versión española, siguiendo las pautas de Bachman, la validez viene dada por la autenticidad de situaciones e interacción en las tareas propuestas.

Distintos intentos (Kunnan 2001; Luoma 2001) siguen buscando proporcionar pistas adecuadas para la validación. A pesar de lo que ya se ha investigado sobre ello, es éste todavía un terreno por explorar.

2.5.3.2.1. Validez de los constructos y teorías sobre el uso de la lengua

La visión tradicional que clasificaba la validez en distintos tipos (de contenido, de constructo, aparente, predictiva, concurrente, etc.) da paso a otras

interpretaciones que plantean la validez de constructos como concepto unificado que engloba múltiples facetas y reconocen la validación de las pruebas como un proceso continuo (Chapelle 1999).

Con la preocupación creciente por fijar qué miden exactamente los exámenes y sobre todo cuáles son su utilidad y sus consecuencias educativas, aparece un nuevo concepto de validez, la consecuencial (*consequential validity*), que se relaciona directamente con el impacto o efectos de las pruebas.

El elemento central de las nuevas ideas sobre validez pasa a ser el constructo de los exámenes, es decir, lo que queremos medir; el conocimiento de la lengua y la habilidad para usarla. Según Alderson y Banerjee (2002: 80), no se puede evaluar si no se conoce bien lo que implica aprender una lengua*:* "Central to testing is an understanding of what language is, and what it takes to learn and use language, which then becomes the basis for establishing ways of assessing people's abilities".

Bachman (1991) destaca el avance que supone para la Evaluación de Lenguas el desarrollo de una teoría que considera la habilidad lingüística como multi-componencial. El modelo de Bachman reconoce que en la actuación en los exámenes se refleja tanto la habilidad lingüística (conocimientos y estrategias) como el método de examen (entorno, *input*, *expected response*). En este modelo se basa el Marco de Referencia para la Enseñanza de Lenguas que propone el Consejo de Europa (*Council of Europe's Common European Framework* 2001), al que *Language Testing* dedicó un número especial monográfico en 2005: *LT* 22 (3).

Si la habilidad o competencia en una lengua tiene múltiples componentes, éstos no tienen porqué desarrollarse a la vez ni en la misma medida (Perkins y Gass 1996). En el desarrollo de la adquisición de una lengua hay que tener en cuenta el progreso y los logros alcanzados en cada momento (Danon-Boileau 1997).

McNamara (1995), sin embargo, apunta que el modelo de Bachman olvida la dimensión social de la competencia lingüística, y recuerda que el aspecto de interacción es vital en la evaluación de lenguas.

2.5.3.2.2. Investigaciones sobre validación

Múltiples estudios sobre validación han basado su trabajo en la comparación de la actuación de los sujetos en distintas pruebas de Inglés estandarizadas y reconocidas, como el Cambridge y el TOEFL, o de ámbito más modesto. Una vez establecidas las correlaciones estadísticas, los estudios constataron las semejanzas y permitieron mejorar las pruebas.

Esta tesis también basa su estudio empírico en comparar la actuación de los alumnos de Bachillerato en distintas pruebas de Inglés (PAAU) para validar el C-test.

No obstante, además de estos métodos cuantitativos de validación, existen otros de tipo cualitativo.

2.5.3.3. La autenticidad

A partir de los años 70, tal y como mencionamos en el apartado 2.3.2 de este capítulo, de la mano del movimiento comunicativo, la autenticidad pasó a ser una preocupación para los especialistas en evaluación.

El tema es complejo, porque la propia naturaleza de los exámenes impide que las tareas propuestas en ellos sean semejantes a las situaciones de comunicación de la vida real. Aún así, se considera una característica deseable de todo examen.

El propio concepto de autenticidad está cambiando. Bachman y Palmer (1996: 23) definen la autenticidad como “The degree of correspondence of the characteristics of a given language test task to the features of a TLU (Target Language Use) task”.

Lo cierto es que, a pesar de todo, el tema de la autenticidad en la evaluación no ha estado tan presente en esta década como en la anterior.

Además, los estudios presentan claras contradicciones. Desde Wu y Stansfield (2001), que recomiendan la autenticidad de las tareas en un test de *Language for Specific Purpose* (*LSP*) para que sea válido y fiable, hasta autores como Lewkowicz (1997, 2000), que ponen en duda que la autenticidad de las tareas planteadas en las pruebas tenga consecuencias para la evaluación.

Un reciente trabajo de Lewkowicz (2000) hace un recorrido histórico de la evolución del concepto de autenticidad en la evaluación desde su aparición en la literatura durante los años 70. En él se cuestiona su importancia para los exámenes y se plantea la necesidad de seguir investigando en este campo.

Parece evidente que se necesita mayor volumen de investigación empírica que arroje luz sobre este tema.

2.5.4. Tipos de pruebas

2.5.4.1. Según el constructo

2.5.4.1.1. Evaluación de la comprensión escrita

La amplia literatura al respecto muestra la preocupación de los investigadores. Leer supone la existencia de una interacción de un sujeto con un texto escrito. Hay unidad al considerar que para afrontar el aprendizaje de la lectura en una lengua extranjera, es necesario dominar antes la lectura en la primera lengua (de la que se hacen transferencias). Pero las teorías difieren al definir qué se valora en los tests de lectura y al intentar explicar el origen de las dificultades que plantea la lectura en otra lengua.

Otro aspecto que plantea problemas es cómo influyen el *background knowledge* o "conocimiento del mundo" de los lectores y el tema del texto.

También algunas líneas de investigación estudian cómo afecta el método de examen utilizado (respuesta múltiple, pruebas de cierre, preguntas de respuesta corta, *C-tests*, repetición inmediata- *immediate recall*, traducción etc.).

Finalmente, mencionaremos que hay acuerdo al considerar que para medir la comprensión lectora es necesario utilizar varios métodos de examen. Sobre todo si tenemos en cuenta el hecho de que la habilidad para solucionar un determinado test o prueba no siempre garantiza la comprensión del texto por parte del lector.

2.5.4.1.2. Evaluación de la comprensión oral

Actualmente se está dando mayor relevancia que en épocas anteriores a la evaluación de esta destreza lingüística, cuyo estudio plantea no pocas dificultades. Los especialistas en lingüística aplicada reconocen su importancia: “The assessment of listening abilities is one of the least understood, least developed and yet one of the most important areas of language testing and assessment” (Buck: x-series editors’ preface, commented by Alderson and Bachman en Buck 2001).

Buck (2001) reivindica la comprensión auditiva como proceso de inferencias en el que intervienen el conocimiento lingüístico (fonología, vocabulario, sintaxis) y no lingüístico (interpretación). Además está condicionada por múltiples variables; unas personales, como el conocimiento previo del mundo, y otras físicas, como las condiciones de la audición, la velocidad, el acento o la entonación.

Algunas investigaciones se dirigen a los métodos de examen, tales como el dictado, la traducción de resúmenes, las pruebas de elección múltiple. Con la incorporación de los medios técnicos en la evaluación surgen también campos nuevos en la investigación de la lectura. Nos lleva por ejemplo a plantearnos cómo afecta a la comprensión auditiva la existencia de información visual, puesto que según algunos trabajos puede suponer una ayuda, pero también una distracción (Gruba 1997).

2.5.4.1.3. Evaluación de la gramática y el vocabulario

En los últimos años ha disminuido el número de estudios sobre la evaluación de la gramática, probablemente porque el interés de los expertos se dirige más a los aspectos comunicativos de la lengua (*communicative language teaching*) que a los gramaticales (Rea-Dickins 1997, 2001), tanto en la enseñanza como en la evaluación de la lengua. O quizás porque se considera que a menudo la gramática se evalúa de forma implícita al valorar el dominio de otras destrezas lingüísticas (*reading, writing*, *speaking,* etc.).

Sin embargo esta pérdida de importancia de la gramática en la enseñanza ha redundado en un aumento significativo de los estudios sobre la evaluación del vocabulario.

Especial actividad se ha desarrollado en la creación de pruebas para medir la amplitud del vocabulario adquirido (*vocabulary size*), partiendo de la base de que es necesario manejar una cierta cantidad de palabras en una lengua para expresarse en ella. Son dignos de mención los esfuerzos de Nation (1990) que creó una prueba de vocabulario en formato "elección múltiple" para unir una palabra con su sinónimo, y Meara (1996) con la prueba de reconocimiento de vocabulario conocida como *Yes/no vocabulary test,* en la que se introducen pseudo-palabras para comprobar si el sujeto sobrevalora sus conocimientos de vocabulario. Tras sus pasos, diversos autores han continuado con estudios de validación de distintas versiones de estos tests (Schmitt *et al*. 2001; Beglar y Hunt 1999; Beeckmans *et al.* 2001; Huibregtse 2002).

Schmitt (1999) cuestionó la validez de los ítems de vocabulario de los actuales tests TOEFL. Schmitt *et al*. (2001) revisaron dos nuevas versiones del *Vocabulary Levels Test*. Takala y Kaftandjieva (2000) buscaron el posible sesgo del género en una prueba de vocabulario. Y Read y Chapelle (2001) desarrollaron un marco para la evaluación del vocabulario en la enseñanza de una segunda lengua.

Algunos estudios pretenden abordar la evaluación del vocabulario de forma aislada y otros buscan hacerlo desde una perspectiva más global. Entre estos intentos debemos dirigir nuestra atención hacia Laufer y Nation (1999) puesto que diseñaron un tipo de prueba de vocabulario cuyo formato se asemeja al C-test, por lo que nos referiremos a él en los capítulos 4 y 6.

Valorar la cantidad de palabras aprendidas no es suficiente. Laufer *et al*. (2001) muestran su preocupación por conocer la profundidad de ese aprendizaje y medir la capacidad receptiva y productiva del vocabulario adquirido (vocabulario activo y pasivo). Estudios recientes van encaminados hacia la valoración cualitativa del aprendizaje del vocabulario.

El capítulo 4 de la tesis retoma este tema y se dedica íntegramente a la evaluación del vocabulario.

2.5.4.1.4. Evaluación de la expresión oral

Si ya en la década anterior (1984-94) fue fructífera y abundante la investigación sobre evaluación de la competencia oral en lengua extranjera, en estos últimos años se aprecia un énfasis significativo en los estudios centrados en los exámenes orales.

La necesidad de validar los tests orales (Shohamy 1994), el desarrollo de escalas orales para distintos tests (Chalhoub-Deville 1995), los efectos del formato y la densidad léxica del test en el subtest de interacción oral del *Australian Assessment of Communicative English Skills* (O´Loughlin 1995), y el diseño de tareas para los exámenes orales en grupo (Fulcher 1996) son algunos ejemplos de los aspectos estudiados.

A pesar de que es larga la tradición de la evaluación de la expresión oral, sobre todo mediante la entrevista oral (*Oral Proficiency Interviews*, OPIs), hoy en día se cuestiona la validez de este tipo de pruebas y abundan las investigaciones sobre la forma de medir mejor la competencia oral en una lengua.

Las entrevistas cara a cara plantean problemas de validez y fiabilidad, porque la situación impide que el intercambio lingüístico/la interacción se asemeje a una conversación normal. Por ejemplo, el turno de palabra está estructurado, el sujeto no se implica o involucra en la comunicación y las estrategias de corrección son más formales.

Lazaraton (1996) presenta un estudio de las entrevistas orales en el *Cambridge Assessment of Spoken English* (CASE) y cómo afectan las muestras de apoyo del entrevistador al resultado de la entrevista. En la línea de las entrevistas orales está el estudio de Kormos (1999) y la revisión del ACTFL *Oral Proficiency Interview* de Salaberry (2000). McNamara y Lumley (1997) investigan las variables que afectan a la evaluación de las destrezas orales en el Occupational English Test para profesionales de la salud.

Las nuevas pruebas de entrevista oral para evaluar los progresos de alumnos en programas de inmersión centran los estudios de Carpenter, Fujii y Kataoca (1995), Bae y Bachman (1998).

Los intentos de mejorar la entrevista oral han planteado formatos que varían el número de sujetos implicados en la entrevista, con dos examinadores y con dos o más examinandos. En cualquier caso, los investigadores señalan que los criterios de

evaluación (*rating performance*) deben estar siempre claros y los correctores han de ser debidamente seleccionados y formados (Wigglesworth 1993; Lumley y McNamara 1995; Lumley 1998).

Por otra parte la tecnología aporta nuevas posibilidades como el uso de laboratorios de idiomas, video conferencias, comunicación por teléfono, uso de ordenadores, grabadoras, etc., algunas de las cuales evitan el interlocutor humano.

2.5.4.1.5. Evaluación de la expresión escrita

Como forma de expresión o actuación lingüística, la evaluación de la expresión escrita se enfrenta a los mismos problemas citados en el epígrafe correspondiente a la expresión oral, es decir, la búsqueda de criterios adecuados, la garantía de objetividad y fiabilidad en las puntuaciones, y la propuesta de tareas que provoquen el tipo de comportamiento lingüístico buscado.

Pero el problema principal de la evaluación de las pruebas de expresión o producción es el diseño y aplicación de técnicas de corrección. En el caso de la expresión escrita, por ejemplo, cómo puntuar los ensayos sin que aflore la subjetividad del corrector. Por eso, durante años se ha valorado esta destreza de forma indirecta, a través de pruebas gramaticales o de vocabulario. Sin embargo, la tendencia actual, promovida por el movimiento comunicativo, es reconocer que la expresión escrita va más allá e incluye aspectos textuales de estructura del discurso. Por eso se proponen tareas semejantes a las situaciones que ocurren en la vida real, como redacción de cartas, *e-mails*, memos, ensayos, etc.

Las investigaciones más recientes insisten en el diseño de tareas bien estructuradas y escalas de corrección apropiadas. También insisten en la necesidad de que los correctores estén bien formados (Cumming 1990; Brown 1991; Weigle 1994; Sakyi 2000). Sobre la actuación del corrector y su incidencia en la evaluación destacamos la aportación española con las investigaciones de Amengual Pizarro (2003), centradas en fiabilidad de las puntuaciones holísticas en ítems abiertos.

Algunos de los últimos trabajos utilizan las nuevas tecnologías e inician el complejo desarrollo del *e-rater* (Burstein y Leacock 2001).

2.5.4.2. Según el ámbito de aplicación

2.5.4.2.1. Los exámenes nacionales o estandarizados

Siguen siendo objeto de numerosas publicaciones que los describen y estudios que los analizan. Esta actividad revela la preocupación de los investigadores por los exámenes de ámbito más amplio.

De nuevo el TOEFL protagoniza algunas contribuciones, aunque no con la frecuencia de la década anterior. Se estudia cada una de las partes de esta prueba. Hale y Courtney (1994), por ejemplo, revisan la sección de comprensión oral y la conveniencia de tomar notas para mejorar la actuación en la prueba.

Sin embargo, la mayoría de los estudios empíricos van encaminados a examinar la validez de otras pruebas estandarizadas. Henning *et al*. (1994) revisaron la eficacia del *English Comprehension Level* (ECL), examen utilizado por el Ministerio de Defensa de Estados Unidos para evaluar la competencia lingüística de los militares de otros países que pasan un tiempo de formación en aquel país. Fulcher (1997) estudió la validez y fiabilidad del test de nivel de la Universidad de Surrey. Powers *et al*. (1999), el *Test of Spoken English*. Scott *et al*. (1996) el *Listening summary translation examination (LST)-Spanish version*. Cushing Weigle (2000) valoró el *Michigan English language assessment battery* (*MELAB*). Dollerup *et al*. (1994) se ocuparon de cómo mejorar el *Sprogtest*, la prueba que se utilizaba en Dinamarca para diagnosticar la competencia de los universitarios en lengua inglesa, la lengua de sus libros de texto. En esta línea de evaluación de un examen de nivel trabajaron Wall, Clapham y Alderson (1994) para validar el test institucional de la Universidad de Lancaster. Por otra parte, Paapajohn (1999) se centró en cómo afecta la variación de los temas en la prueba de química C*hemistry TEACH test.*

En España hemos de constatar la falta de estudios coordinados sobre la prueba de Inglés de la Selectividad, a pesar del tiempo de vigencia de la prueba en nuestro sistema educativo. Recientemente se está paliando esta sequía con el volumen *Estudios y criterios para una Selectividad de calidad en el examen de Inglés* (2005), que incluye las aportaciones de varios investigadores preocupados por el tema, coordinados por Herrera Soler y García Laborda.

No obstante, Alderson (2001) advierte de la necesidad de estudiar también los tests de ámbito local o de menor repercusión, por la valiosa información que aportan a la hora de acometer reformas educativas, sobre todo si se valora su efecto rebote.

2.5.4.2.2. El Inglés para fines específicos (IFE)

Se han diseñado pruebas de evaluación cuyo contenido se refiere a contextos concretos de uso de la lengua y no a situaciones de tipo general. Sin embargo, estas pruebas no son diametralmente opuestas a las de Inglés general. Sí hay algunas diferencias que menciona Douglas (1997, 2000) en Alderson y Banerjee (2001). La diferencia fundamental es la interacción entre el conocimiento de la lengua y el conocimiento de los contenidos concretos. Se supone que su conocimiento de la lengua va ligado a un campo específico de conocimiento fuera del cual estaría en clara desventaja.

Parece evidente que el primer paso en la creación de una prueba de IFE es el análisis de la situación concreta de uso de la lengua, los temas habituales, los escenarios típicos en que se desarrolla y las características de la lengua en ese campo específico (sintácticas, léxicas). En este sentido se han movido diferentes investigaciones, con resultados a veces contradictorios, que pretendían fijar hasta qué punto el conocimiento del campo específico (*background knowledge*) condiciona la actuación en las pruebas de lengua (Jensen y Hansen 1995; Fox *et al.* 1997; Clapham 1996; Jennings *et al*. 1999; Cumming 2001).

Por otra parte, en cuanto a las tareas que se proponen en las pruebas de Inglés para fines específicos, preocupa especialmente a los especialistas el cuidado de la autenticidad, para que la actuación del sujeto en las pruebas mida exactamente su actuación en las tareas de la vida real. A este respecto destacamos la preocupación de los autores por conseguir materiales auténticos (Wu *et al*. 2001; Cumming 2001), aunque Lewkowicz (1997) demostró que no siempre es fácil distinguir los textos auténticos de los creados.

El diseño y estudio de las pruebas de IFE ha suscitado gran cantidad de interrogantes y ha propiciado investigaciones más profundas en numerosas cuestiones sobre la evaluación.

Queda aun sin respuesta la pregunta planteada por Alderson (1988) de hasta qué punto una prueba de IFE tiene que o puede ser específica, y se mantiene el reto de determinar si es necesario aplicar una prueba de IFE para conocer la competencia de un sujeto o bastaría con una prueba de Inglés general.

2.5.4.2.3. Autoevaluación

El interés por la autoevaluación va en aumento desde los años 80. Se tiende a implicar cada vez más al alumno en su propio proceso de aprendizaje.

La autoevaluación aparecía ante los ojos de muchos especialistas como un campo prometedor para la evaluación formativa (*formative assessment*) (Oscarson 1989 en Alderson y Banerjee 2001). Permitía a los alumnos confiar en sus propios juicios, valorar la evaluación como algo que abarca todo el proceso de aprendizaje y era de utilidad para los profesores. Algunos especialistas, sin embargo, dudan que sea posible para los alumnos legos en materia de evaluación de lenguas hacer una buena autoevaluación sin ayuda (Blue 1988 en Alderson y Banerjee 2001).

En las últimas décadas se han seguido desarrollando y validando nuevos instrumentos de autoevaluación (Blanche 1990; Hargan 1994; Carton 1993; Bachman y Palmer 1989).

1.5.4.2.4. La Evaluación Alternativa

El movimiento de la *Alternative Assessment* o Evaluación Alternativa[21] surgió en el contexto educativo de los Estados Unidos. Incluye todas aquellas formas de evaluación distintas de la evaluación tradicional: abarcan un período mayor de tiempo, son de tipo formativo más que sumativo y tienden a producir un efecto rebote beneficioso. La autoevaluación está, por tanto, dentro de este grupo de procedimientos.

[21] Hemos de hacer notar que este movimiento se denomina significativamente en inglés con el término *Assessment*, en lugar del más limitado *Testing*. Remitimos a la revisión terminológica del capítulo 1.

La preocupación por los aspectos éticos y educativos de la evaluación es una constante en el movimiento de la Evaluación Alternativa.

En general, estos modelos de evaluación presentan posibles inconvenientes, como el tiempo o los problemas de administración y corrección. Y ventajas, porque aportan una mejor información y se integran mejor en el proceso de aprendizaje del aula.

La revista *Language Testing* publicó un número especial monográfico en 2001, *LT* 18 (4), editado por McNamara y titulado *Re-thinking Alternative Assessment*. En él aparecen artículos significativos sobre el tema, firmados por Lynch, Shohamy, Brindley, Butler y Stevens, Rea-Dickins y Spence-Brown, además del propio editor.

McNamara (2001) reflexiona sobre el carácter social de la evaluación y propone una revisión de las prioridades y responsabilidades que han de guiar la investigación en este campo. Lynch (2001) sigue una línea semejante, desde una perspectiva crítica reconsidera aspectos relacionados con la ética, la validez y la autenticidad de las pruebas. Y Shohamy (2001), consciente del enorme poder de las pruebas, aboga por una evaluación inspirada en principios democráticos. De este modo se conseguirán modelos más éticos, educativos y válidos.

Como en otros aspectos de la evaluación queda todavía mucho por hacer, pero se camina con paso firme en busca de nuevas perspectivas para la evaluación, siempre con la intención de mejorarla y adaptarla a la realidad actual. Citamos parte del ilustrativo editorial de McNamara (2001: 332) al respecto:

> Clearly, the research presented here represents only a beginning on the vast task of renovating our research directions to reflect more closely the emerging theoretical insights into the role of assessment as a social practice, and to carry out our responsibilities as researchers to learners and teachers as much as to managers and administrators in language education.

2.5.4.3. Diseño de pruebas

A la hora de afrontar el diseño de una prueba nos encontramos con no pocas dificultades. Las distintas teorías (*approaches to test design*) hacen sus aportaciones. Son muchas las variables que intervienen en la comunicación lingüística y que deben ser tenidas en cuenta en la evaluación.

Por una parte, como hemos visto, se reclama autenticidad. Desde el conductismo se considera que los resultados de una prueba son la interpretación del comportamiento observado "meaningful interpretation of observed behaviour" (Chapelle 1998: 33). Por otra, se subraya la necesidad de partir del análisis de las necesidades (Munby 1978), o de las tareas (Bachman y Palmer 1996).

La Lingüística Aplicada y la investigación en adquisición de segundas lenguas indican que el comportamiento lingüístico depende también del contexto en que se desarrolla. Por tanto, al diseñar una prueba no sólo deben afectan las características del individuo (*traits*) sino también el contexto: "Performance is a sign of underlying traits in interaction with relevant contextual features. It is therefore context-bound" (Alderson y Banerjee 2002: 100).

No podemos concluir el repaso a la actualidad en Evaluación de la Lengua sin incluir los esfuerzos que dedican los investigadores a la búsqueda, creación, experimentación y posterior validación de técnicas de evaluación. En este aspecto, destacan los que se relacionan de forma más o menos directa con las pruebas de cierre y, por su interés para la investigación que sustenta esta tesis, los relacionados con el C-test. Siguiendo las directrices de investigaciones anteriores intentan profundizar en las técnicas iniciadas en la década de 1984-94. Tras sus pasos encaminamos nuestra investigación.

Con el fin de comprobar las ventajas que Klein-Braley y Raatz atribuían a la nueva técnica, Jafarpur (1995) comparó el C-test con los *clozes* tradicionales. Finalmente concluyó que esta técnica adolece de de los mismos problemas que los *clozes* y recomendó seguir investigando sobre ellos. Por otra parte, el estudio de Allan (1995) investigó la validez de los cuestionarios en los exámenes de elección múltiple de comprensión escrita y las estrategias que siguen los examinandos.

Farhady y Keramati (1996) propusieron un *cloze test* dirigido frente al test de ratio fija con omisiones aleatorias cada *n* elementos. Con el *text-driven method* las omisiones dependen de cada texto. Ya Brown (1993) había investigado acerca de los *clozes* naturales, es decir, los que no tienen en cuenta factores como la dificultad del texto, el tema, etc. Como Brown, Farhady y Keramati (1996) recomendaron dirigir el *cloze* para mejorar su fiabilidad.

Storey (1997) examinó el proceso de realización o aplicación de los tests a través del estudio de los procesos que se siguen al realizar un *cloze,* pues de las estrategias empleadas se puede inferir los procesos cognitivos utilizados. Sasaki (2000) estudió la influencia de los esquemas culturales (términos familiares) en las pruebas de cierre tradicionales. Klein-Braley (1997) comparó la actuación de un grupo de alumnos en varias pruebas de redundancia reducida. El C-test resultó ser el procedimiento más válido, fiable y económico. Y por último, Eckes y Grothjahn (2006) examinaron la validez de constructo del C-test alemán.

En cuanto a la técnica de elección múltiple para medir la comprensión, Freedle y Kostin (1999) analizaron la influencia del texto en pequeñas conversaciones del TOEFL.

2.5.5. Nuevos retos en la enseñanza de lenguas

2.5.5.1. La ética en la evaluación de lenguas

Tal y como mencionaba McNamara (1998), en la actualidad observamos una ampliación del repertorio de aspectos relacionados con la evaluación de lenguas, que nos lleva al estudio de otras disciplinas, como la ética. Ya en la década de los 80 Canale anticipaba esta tendencia y apelaba a la responsabilidad ética de los profesores en el proceso evaluador.

Language Testing (1997) también dedicó un número monográfico a la ética, *LT* 14 (3), editado por Allan Davies y en el que colaboraron Spolsky, Hawthorne, Elder, Norton y Starfield, Hamp-Lyons, Rea-Dickins, Lynch y Shohamy.

Algunos autores, como Alderson (1997), defienden que un examinador no puede evitar preocuparse por aspectos de carácter ético como es la creación de exámenes justos (validez y fiabilidad de las pruebas). Davies (1997), sin embargo, aboga por la separación de ética y validez, y propugna una profesionalización creciente en el campo de la evaluación. También Bachman (2000) se hace eco de esta idea.

Lo cierto es que todos los investigadores, conscientes del papel cada vez más importante de los exámenes en la sociedad, animan a enfocar la evaluación desde

una perspectiva ética. De nuevo, hemos de citar a Canale (1988: 75): “Once one has been involved in gathering information, one becomes *responsible* in some way to see that it is used *ethically*”. (la cursiva es mía)

La Asociación Internacional de Evaluación de Lenguas (ILTA) adoptó en 2000 un código ético para los examinadores. La propia asociación define el *Code of Ethics* como “a set of principles which draws upon moral philosophy and serves to guide good professional conduct”. En él se propician las prácticas éticas y se apela a la responsabilidad moral de los profesionales de la evaluación. Actualmente prepara un código de práctica: *Code of Practice*. Ambos serán revisados periódicamente para responder a los cambios y necesidades que plantee la profesión y la sociedad en cada momento.

2.5.5.2. Política

Es inevitable reconocer que los exámenes constituyen un instrumento de política educativa. Como tal, son muy poderosos (Shohamy 2001). En las pruebas a gran escala convergen la política nacional y la preocupación por la evaluación formativa.

Es deseable, por tanto, la colaboración entre políticos y profesionales de la evaluación para crear pruebas de calidad (Brindley 1998, 2001). Alderson (2001) expresa la misma idea cuando dice “Testing is too important to be left to testers”, pues no se puede dejar de lado a otros agentes de la educación con frecuencia olvidados, como son los propios profesores y los políticos.

La política educativa nacional también supone renovar las pruebas según los objetivos que se pretenda lograr con ellas. No podemos olvidar que la responsabilidad de los examinadores va más allá de la mera creación y aplicación de pruebas. La investigación en este campo no es muy abundante y la literatura es escasa. No obstante, a menudo aparecen determinados aspectos de política educativa en artículos que se centran en estudios sobre el impacto de las pruebas y en los dedicados a la reflexión sobre la ética en la evaluación de la lengua.

2.5.5.3. Los estándares en evaluación

Este término admite al menos tres definiciones en el campo de la evaluación:

- Los códigos de buenas prácticas: *codes of practice.*
- Los niveles de competencia en una lengua.
- Las pruebas estandarizadas o institucionalizadas.

La primera alude a los códigos éticos que todo examinador debería respetar para asegurar la calidad del proceso evaluador. En el apartado 2.5.5.1 nos hemos referido al código adoptado por ILTA en 2000. En Europa se ha publicado el código de ALTE (Association of Language Testers in Europe) con este fin.

En cuanto a la segunda definición, el Consejo de Europa (2001) ha dado un nuevo impulso a este respecto con la publicación de un marco común de referencia que fija los niveles de competencia y supone un compendio de todo lo referente a la enseñanza-aprendizaje de lenguas. Además, este marco (*Council of Europe's Common European Framework*) pretende guiar la programación, el desarrollo de criterios comunes de evaluación, la creación de textos y otros materiales, y la formación del profesorado.

Y la tercera se refiere a las pruebas estandarizadas, exámenes institucionalizados que generalmente posibilitan la obtención de un determinado título. Se aplican a un gran número de sujetos, a gran escala, y esta circunstancia incide directamente en ellas. Esta acepción nos pone de nuevo en contacto con los aspectos éticos y políticos de la evaluación (McNamara 1998; Norton 1998; Shohamy 2001).

2.5.5.4. La evaluación en edades tempranas

Las investigaciones en este campo han crecido considerablemente al proliferar el interés por comenzar la enseñanza de idiomas en la educación Infantil y Primaria. En general, se considera que los procedimientos de la Evaluación Alternativa son más adecuados que los métodos tradicionales para la evaluación de los alumnos de

edades entre los 5 y 12 años. *Language Testing* dedica todo su segundo número del año 2000 a este tema, *LT* 17 (2).

2.5.5.5. Las Nuevas Tecnologías en la evaluación

Las Nuevas Tecnologías de la información ofrecen hoy múltiples y ricas herramientas a la enseñanza de idiomas (los métodos cuentan con vídeo, realidad virtual, reconocimiento de voz, de escritura, etc.).

La evaluación asistida por ordenador ha crecido vertiginosamente en los últimos años. Es evidente que el mundo de la informática tiene mucho que aportar a la evaluación (García Laborda 2005; García Laborda y Bejarano 2005). Abre todo un mundo de posibilidades tanto para la administración de exámenes como para su elaboración, corrección, análisis, banco de datos, permite avances en la autoevaluación, etc. En palabras de Alderson y Banerjee (2001: 224) “In short, computers can be used at all stages in the test development and administration process”. Por eso sigue siendo importante la investigación en este campo: “And we need research into the impact of the use of the technology on learning, on learners and on the curriculum” (Alderson 2000c: 603 en Alderson y Banerjee 2001: 227).

Ya en 1998 se introdujo una versión informática del TOEFL. Algunos estudios comparan los resultados de los exámenes administrados de manera tradicional con los que se aplican mediante ordenador. Se obtienen ventajas, sobre todo en términos de accesibilidad y rapidez, pero también hay desventajas con los sujetos que no están familiarizados con el medio informático o los que lo rechazan (Fulcher 1999; Gervais 1997; Taylor *et al*. 1999). Se percibe una constante en los estudios: la preocupación por los posibles sesgos.

Otras vías de investigación se centran en los exámenes adaptados por ordenador o CAT *Computer-adaptative tests.* En ellos el ordenador adapta el examen a cada candidato según su actuación en las preguntas precedentes. De nuevo, esto presenta ventajas, pero también inconvenientes (Brown 1997; Laurier 1998; Chalhoub-Deville y Deville 1999; Dunkel 1999), algunos de los cuales se pueden evitar al tomar las decisiones que determinan el posterior diseño del examen.

Chalhoub-Deville y Deville (1999) consideran que los exámenes por ordenador se basan en tareas discretas (*discrete-point tasks*), de selección, como las preguntas de tipo elección múltiple, que sirven para medir los conocimientos lingüísticos, pero no las habilidades comunicativas.

En España, García Laborda (2005: 37) apunta al uso de las nuevas tecnologías incluso en puebas a gran escala, y menciona una "futura Selectividad asistida por ordenador".

Sin embargo, la evaluación mediante ordenador presenta todavía muchas limitaciones, como señalan Burstein *et al*. (1996: 245) en Alderson y Banerjee (2001: 225). "The situation is created in which a relatively rich presentation is followed by a limited productive assessment". Sin embargo se están desarrollando rápidamente sistemas para valorar incluso las habilidades productivas (*human-assisted scoring systems*: *e-rater*, *PhonePass*, DIALANG, aunque algunos sectores muestren todavía cierto escepticismo.

2.6. Perspectivas de futuro

Son muchos los autores que tomaron el cambio de milenio como punto de referencia para el análisis y revisión de lo que se ha hecho en cada disciplina y de lo que queda por hacer. Varios especialistas proyectaron su visión del campo de la lingüística aplicada en diversos artículos y libros. En ellos valoran la situación previa y alientan la investigación futura (Bachman 2000; Widdowson 1999; Pica 2000).

Lyle F. Bachman (2000) en su artículo *Modern language testing at the end of the century: assuring that what we count counts* reseña los avances prácticos y teóricos, la variedad de enfoques y herramientas que se han usado en evaluación desde los años 80, las mejoras en las técnicas y formatos de examen, los aspectos éticos, etc. Para Bachman (2000) las claves del futuro de la evaluación pasan por la creciente profesionalización y por la profundización en la investigación sobre validación.

Pica (2000) considera que el campo de la enseñanza del Inglés se encuentra en un momento de transición que revisa los enfoques anteriores y busca otros nuevos. Señala que el momento actual propicia la búsqueda de nuevos métodos.

Hasta ahora, según el autor, todos los métodos que han ido surgiendo han tenido en común la voluntad de mejorar los ya existentes y la tendencia al acercamiento profesor-alumno.

Unos años antes, Gipps (1994) había señalado el giro que sufría la evaluación, desde el modelo psicométrico centrado en los exámenes, hacia otro más abierto de evaluación educativa. La evaluación educativa se centra en el individuo y pretende conocer sus dificultades para ayudarle en el aprendizaje. Esta concepción más amplia de la evaluación abarca a los profesores, el proceso de enseñanza, los cursos, los exámenes orales y escritos, etc. Se habla ya de evaluación formativa.

La evaluación educativa distingue entre la competencia y la actuación del alumno en un momento concreto. La competencia alude a lo que el alumno podría hacer en unas circunstancias ideales, y su actuación a lo que realmente hace en una situación concreta en la que le influyen muchos factores (no sólo sus conocimientos, sino también la motivación, situación afectiva personal, familiar, actitud, carácter, nervios, tipo de prueba, etc.).

> "Thus, a student's competence might not be revealed in either classroom performance or test performance because of personal or circumstantial factors that affect behaviour" (Messick 1984). Elaborative procedures are therefore required to elicit competence. (Gipps 1994: 9)

Terminamos este capítulo con la afirmación de Gipps (1994: 1) acerca de la evaluación educativa: "Assessment is undergoing a paradigm shift, from psychometrics to a broader model of educational assessment, from a testing and examination culture to an assessment culture". Efectivamente, en nuestros días la tendencia de la evaluación es ir asumiendo cada vez más fines y un rol más completo en la educación.

CAPÍTULO 3. RASGOS DE LOS EXÁMENES O PRUEBAS

3.1. Introducción

En el capítulo 1 de la tesis hemos visto el papel de la evaluación en el campo de la enseñanza de lenguas. Uno de los instrumentos clave de que dispone el profesor para llevar a cabo la valoración del aprendizaje del alumno es precisamente el examen o prueba. Los exámenes de idiomas han de proporcionar al profesor una medida que éste pueda interpretar como representativa de la competencia del alumno en la lengua (Bachman y Palmer 1996: 23). También los propios alumnos manifiestan a menudo su deseo de disponer de un referente objetivo que les permita ser conscientes de sus progresos en la asignatura y evite la evaluación sesgada del profesor.

Gran parte de la investigación relacionada con la evaluación tiene por objeto la descripción de los rasgos de las pruebas para que sean realmente efectivas como instrumento de medida. Destaca la unanimidad entre teóricos de la evaluación, tales como Oller (1979), Hughes (1989), Gipps (1994), Bachman (1990), Bachman y Palmer (1996) al señalar las características que todo buen examen debería tener. Las únicas diferencias aparecen en la manera de clasificar tales rasgos y en el énfasis que recibe cada uno de ellos en diferentes momentos (Weir 1988). El enfoque tradicional estudiaba las cualidades de las pruebas como independientes entre sí, mientras que la tendencia actual, liderada por Bachman y Palmer (1996: 18), es considerarlas complementarias, dado que “all of which contribute in unique but interrelated ways to the overall usefulness of a given test”.

Así pues, para que las pruebas respondan al objetivo para el cual son diseñadas han de cumplir unos requisitos mínimos. Destacan dos cualidades

básicas: validez y fiabilidad. Bachman y Palmer (1996: 19) se refieren a ellas como cualidades "críticas", vitales:

> Two of the qualities -reliability and validity- are, however, critical for tests, and are sometimes referred to as essential *measurement* qualities. This is because these are the qualities that provide the major justification for using test scores -numbers- as a basis for making inferences or decisions.

Para Bachman y Palmer (1996) las dos se complementan. No obstante, como veremos más adelante (apartado 4) no faltan los que consideran que existe tensión entre validez y fiabilidad (Gipps 1994) y abogan por el deseable equilibrio.

A estos aspectos prioritarios se unen otros, como el carácter práctico e interactivo, la autenticidad y el impacto que producen las pruebas. Todos ellos son importantes e interdependientes, según manifiestan Bachman y Palmer (op. cit.: 38): "This six test qualities all contribute to test usefulness, so that they cannot be evaluated independently of each other".

Dedicamos este capítulo al estudio de los rasgos de las pruebas. Seguimos la clasificación propuesta por Hughes (1989) y Bachman y Palmer (1996).

En primer lugar examinaremos los conceptos de validez y fiabilidad, pues, como hemos mencionado, son cualidades esenciales. Desglosaremos los distintos tipos de validez y, con mayor detenimiento, la validez de constructo (Messick 1989).

En cuanto al concepto de fiabilidad, estudiaremos sus dos componentes; la actuación de los alumnos en distintas ocasiones y la fiabilidad del corrector. Después veremos la relación entre validez y fiabilidad.

A continuación analizaremos las demás cualidades de las pruebas: la autenticidad, el carácter interactivo y la factibilidad. Culminamos el capítulo con una revisión del impacto o efecto rebote. En cada apartado veremos, además, hasta qué punto el diseño del C-test reúne estos rasgos. Posteriormente, en la Perspectiva Empírica, lo comprobaremos mediante el análisis del C-test aplicado.

3.2. Validez de las pruebas

Comenzamos haciendo una aproximación al concepto de validez, partiendo de la aportación de Hughes. En los epígrafes siguientes analizaremos sus tipos.

Hughes (1989) aporta una definición general, directa y rotunda de validez: un examen es válido cuando realmente mide lo que pretende medir. Weir y Roberts (1994: 137) insisten en la misma idea: "the cardinal principle is to establish clearly what you want to find out. Validity is concerned with *measuring what you want to measure*".

A pesar de la aparente claridad de esta definición, el concepto de validez es complejo, pues como revelan Cumming y Berwick (1996: 1), se refiere a múltiples aspectos:

> Validation in language assessment is ominously important [...]. But establishing validity in language assessment is by all accounts problematic, conceptually challenging and difficult to achieve. [...] Test validation has long been recognized as an exacting process that requires many types of evidence, analyses and interpretation.

Atendiendo a su complejidad, la literatura muestra distintas concepciones; desde las que subdividen la validez en numerosos tipos distintos (Angoff 1988)[22] hasta las que la consideran un concepto unitario (Anastasi 1982; Cronbach 1988; Messick 1989).

Gipps (1994) alude a los cuatro tipos de validez que aparecen en los primeros escritos al respecto: predictiva, de contenido, de constructo y concurrente.

La aproximación de Hughes (1989), con la que iniciamos este apartado, también contempla cuatro aspectos de validez que el propio autor desglosa: validez de contenido, criterial, de constructo y aparente. En esta clasificación la validez criterial engloba a la predictiva y a la concurrente. Otros enfoques (Messick 1989) consideran también al efecto rebote como un tipo de validez, denominada consecuencial. Debido a la importancia y amplitud del tema, en este estudio dedicamos un epígrafe propio al impacto de las pruebas.

Gipps (1994) alerta del peligro de una fragmentación excesiva de la validez; pues puede suponer que, en la práctica, las pruebas se validen teniendo en cuenta sólo alguno de sus tipos. Para evitarlo, Messick (1989: 19) aboga por la concepción de validez como un concepto unitario que se basa en la validez de constructo:

[22] La revisión histórica de las concepciones de validez que hizo Angoff (1988) distingue 16 tipos de validez. Entre ellas, además de las que comentamos en este epígrafe, *convergent, discriminant, ecological, factorial, population, operational, task, temporal validity* y *validity generalization*.

"Validity is a unitary concept, in the sense that score meaning as embodied in construct validity underlies all score-based inferences".

La validez no es tanto una propiedad de las pruebas en sí mismas, como del significado o interpretación de los resultados derivados de ellas. Lo que se valida son las inferencias de las pruebas: "Test validation is empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied setting that might erode or promote the validity of local score interpretation and use" (Cronbach 1971 citado en Messick 1996: 246).

La concepción de Messick (1989: 13) es exigente y tan amplia que abarca al resto de los tipos de validez. Llega a tener en cuenta las implicaciones sociales y educativas del uso de las pruebas, es decir, su impacto:

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales suppport the *adequacy* and *appropriateness of inferences* and *actions* based on test scores or other modes of assessment.

Por su reconceptualización de la teoría de la validez, Messick ha sido considerado en Estados Unidos como autor clave en la materia (Gipps 1994). Alderson y Banerjee (2001) y Davies (2003) respaldan la idea unitaria de validez de Messick porque equilibra todos los rasgos de las pruebas y permite "to view them as linked together in the practical test situation and to reserve coherence, as Messick did, for construct validity" (Davies 2003: 362).

A continuación haremos un breve análisis de los tipos de validez, sin olvidar que en la literatura no existe una única clasificación. En primer lugar, veremos la validez de constructo, puesto que, según la teoría de Messick, engloba a los demás tipos. Después, nos centraremos en la validez de contenido. Las separaremos para su estudio, aunque algunos autores, como Underhill (1987: 106 citado en Fulcher 1999) consideran que ambos tipos de validez están íntimamente ligados: "Construct validity is not an easy idea to work with, [...] In practice, there may be little difference between construct and content validity". Más adelante abordaremos la validez criterial y aparente.

3.2.1. Validez de constructo

Podemos definir esta cualidad como la capacidad de un examen o prueba para medir exactamente lo que pretende.

Según Hughes (1989: 26) una prueba tiene validez de constructo "if it can be demonstrated that it measures just the ability which it is supposed to measure" e indica que con "constructo" se refiere a "any underlying ability or trait which is hypothesised in a theory of language ability".

De modo semejante la definen Bachman y Palmer (1996: 21): "The term construct validity is therefore used to refer to the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure".

Consideran que la validación del constructo es un proceso continuado que justifica la interpretación de los resultados de las pruebas: "It is important for test developers and users to realize that test validation is an on-going process and that the interpretations we make of test scores can never be considered absolutely valid" (op. cit.: 22).

En el apartado anterior apuntamos a Messick (1989) como principal ideólogo de una nueva concepción de validez considerada como concepto unitario basado en la validez de constructo[23]. Según Messick (1989: 16) la validez de constructo ha de basarse en la recopilación de evidencias que demuestren la representatividad y relevancia de la prueba con respecto al constructo que se desea medir: "Construct validity is based on any evidence that bears on the interpretation or meaning of the test scores".

También Moss (1992: 233) fija el propósito de la validez de constructo en términos similares: "The essential purpose of construct validity is to justify a particular interpretation of a test score by explaining the behaviour that the test score summarises" (citado en Gipps 1994).

Se hace necesario contar con un marco teórico explícito para la validación. El de Messick (1989) recoge las distintas dimensiones de la validez (Tabla 3.1).

[23] Messick (1989: 248) diferencia seis aspectos de la validez de constructo que son implícitos a la validez como concepto integrado o unitario: "These are content, substantive, structural, generalizability, external and consequential aspects of construct validity".

Tabla 3.1. Facets of validity (Messick 1989: 20)

	Test Interpretation	**Test Use**
Evidential basis	Construct Validity	Construct Validity+Relevance/utility
Consequential basis	Value Implications	Social consequences

La validez de constructo aparece como base para interpretación de las pruebas, así como para el uso de las mismas, combinada con la relevancia de la prueba según su propósito y la utilidad del examen en su contexto de aplicación.

En cuanto a las consecuencias de la interpretación de las pruebas, destacamos la mención de las implicaciones educativas y sociales (el impacto de la prueba), puesto que supone una expansión del concepto de validez y la relaciona con los aspectos éticos de la evaluación.

Esta apreciación de Messick sirve para que algunos autores (Frederiksen y Collins 1989[24]; Gipps 1994; MacNamara 1997) distingan otro tipo de validez, que denominan consecuencial: "what has come to be known as "consequential validity" is a key issue in ethical considerations" (Gipps 1994: 63).

MacNamara (1997) incluye lo que denominamos impacto o efecto rebote como un conjunto de aspectos dentro de la validez consecuencial. Como ya hemos indicado en la introducción de este capítulo, en esta tesis preferimos dedicar todo un apartado al impacto de las pruebas, tanto en el contexto del aula y los programas (*washback*) como en el sistema educativo (*systemic validity*). En él retomaremos esta teoría de validez de constructo como concepto unitario de Messick (1989, 1996), que relaciona directamente a la validez con el impacto.

Gipps (op. cit.: 61) valora la aportación de Messick y Cronbach en los siguientes términos: "[They] have taken the discussion of validity beyond a conception based on the functional worth of the testing: construct validity is needed not only to support test interpretation, but also to justify test use".

[24] Frederiksen y Collins (1989, en Gipps 1994: 27) hablan de *systemic validity* como forma específica de validez consecuencial. Se refiere al impacto de las pruebas en el sistema educativo al que pertenecen: "A systemically valid test is one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure".

Si queremos determinar la validez de constructo del C-test, debemos seguir varios pasos, que aparecen detallados en la parte experimental de la tesis. En primer lugar debemos fijar en las especificaciones previas los límites del constructo o dominio que pretende medir. El C-test es una prueba ambiciosa que pretende medir el constructo de la competencia general en lengua inglesa. Los estudios de Klein-Braley (1985) sobre validación del constructo en el C-test serán comentados más adelante.

Tomando como marco el concepto unitario de validez desarrollado por Messick (1989) vemos que todas las cualidades de las pruebas se interrelacionan. Messick propone que se intente mejorar la validez ya desde el diseño de la prueba, pues un buen diseño producirá un efecto rebote beneficioso. También entran en el proceso de validación las prácticas de preparación de los alumnos ante la prueba, tales como la familiarización y la reducción de la ansiedad.

3.2.2. Validez de contenido

La validez de contenido de una prueba viene dada por la relevancia y representatividad de las estructuras que incluya. Y en última instancia, éstas dependen del propósito de la prueba. Veamos la definición de Hughes (1989: 22) y a continuación las consideraciones de Bachman *et al*. (1996: 125):

> A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned.
>
> Content considerations are widely viewed to be essential in the design of language tests, and evidence of content relevance and coverage provides an important component in the validation of score interpretations.

No obstante, Cronbach (1971), Messick (1989) y Bachman *et al*. (1996) coinciden en señalar que la información acerca del contenido de las pruebas (relevancia y representatividad) no es base suficiente para la interpretación de los resultados, puesto que no tiene en cuenta la actuación real de los sujetos.

Al igual que en la validez de constructo, un momento clave es la especificación de los límites del dominio que se quiere medir y la selección de tareas relevantes y representativas del mismo. Carroll (1980), Weir (1988), Hughes (1989) y Alderson *et al*. (1995) recomiendan que en los momentos iniciales del diseño de un examen se especifiquen las destrezas o estructuras que ha de cubrir la prueba. Así será más fácil la selección de elementos para su inclusión. La validez de contenido se basará en la comparación de las destrezas especificadas con el contenido del examen.

La precisión de la prueba como instrumento de medida viene dada en gran parte por su validez de contenido: "The greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure" (Hughes 1989: 22). El propio autor nos alerta del peligro del diseño de pruebas cuyo contenido no sea relevante, sino simplemente fácil de medir.

Ya hemos comentado que numerosos autores (Kelly 1978 y Moller 1982 en Weir 1988; Underhill 1987) han destacado la estrecha relación entre validez de constructo y de contenido. Kelly (1978: 8) llega a considerar la validez de contenido como "an almost completely overlapping concept" con respecto a la de constructo.
Como hemos visto en el apartado anterior, también el marco de la validez de constructo como concepto unitario concebido por Messick (1989, 1996) incluye al contenido: "The content aspect of construct validity includes evidence of content relevance and representation as well as technical quality" (Messick 1996: 248).

Fulcher (1999a) destaca la importancia de la validez de contenido en el campo del Inglés para Fines Académicos (EAP). De nuevo, una prueba es válida si las tareas que propone constituyen un ejemplo representativo del dominio o constructo lingüístico que se pretende conseguir (Bachman y Palmer 1996), es decir, si reflejan, tanto en su contenido como en su formato, el curso de EAP correspondiente. La dificultad es precisar el dominio de la lengua.

La literatura coincide en que, a efectos prácticos, al crear una prueba conviene asegurarse de que ésta tenga un contenido representativo (Anastasi 1982). Y para ello es fundamental tener claros los objetivos, como referente previo a la toma de decisiones tales como la inserción del contenido, elección de las destrezas que se van a medir, del tipo de texto, del formato, asignación de tiempo, etc.

Si nuestro objetivo es medir la competencia general de un alumno en la lengua meta, podremos tomar en consideración la aplicación del formato C-test, dentro de las pruebas de cierre. A continuación, deberemos elegir los textos apropiados como base de la prueba y el punto de comienzo de las omisiones. Otras decisiones afectan al tiempo asignado, las instrucciones de realización, el criterio de corrección, etc. En el contexto del aula, el conocimiento de los programas educativos, de los alumnos y de su proceso de aprendizaje, facilitan al profesor la toma de decisiones al respecto.

3.2.3. Validez criterial

La validez criterial se refiere a la validación de una prueba con respecto a otra independiente pero que mida la misma capacidad, tomada como referencia. La validez viene dada por la correlación (desde 0 hasta 1) entre los resultados de ambas pruebas. En palabras de Hughes (1989: 23), este tipo de validez muestra "how far results on the test agree with those provided by some independent and highly dependable assessment of the candidate's ability".

En la literatura se especifican dos tipos de validez criterial: concurrente y predictiva.

La validez concurrente viene dada por la correlación de la prueba con la que se toma como referencia cuando ambas se realizan al mismo tiempo o en un breve intervalo (Davies 1983; Hughes 1989). Oller (1979) llega a decir que la fiabilidad de una prueba se puede considerar como un caso especial de validez concurrente.

La validez predictiva, por otra parte, se refiere al grado en que los resultados obtenidos en una prueba pueden predecir la actuación del alumno en un examen o situación futura. Por ejemplo, la predicción de las posibilidades de que un alumno pueda seguir un curso o nivel determinado en el futuro, partiendo de su actuación en una prueba. Este tipo de validación es fundamental en el *placement testing*, y requiere hacer un seguimiento posterior que revele si los alumnos fueron ubicados en el nivel que les correspondía (Hughes 1989).

Los resultados obtenidos en la PAAU determinan qué estudiantes acceden a la Universidad en España. Cabe plantearse si realmente es un buen instrumento para

predecir la actuación de los estudiantes en la carrera elegida. Un seguimiento posterior podría dar la clave al respecto, pero son tantos los factores implicados que se hace costosa y difícil la extracción de conclusiones fiables y válidas (Gipps 1994). Recientemente, Sanz y Fernández (2005) han llevado a cabo un estudio de la validez predictiva de la prueba de Inglés de Selectividad y el Quick Placement Test (QPT) que la pone en entredicho.

Hughes (1989), Alderson *et al.* (1995), McNamara (1997) y Fox (2004) también reconocen la dificultad de este tipo de estudios. Mencionan el hecho de que siempre queda la incógnita de cuál habría sido la actuación de los que no la superaron, en caso de haberlo conseguido. El uso de lo que se denomina *truncated samples* puede afectar negativamente al coeficiente de validez predictiva.

Por último, Hughes (1989) y Fox (2004) cuestionan el papel de la competencia lingüística en el éxito académico basándose en los resultados de diversas investigaciones (Criper y Davies 1988; Light *et al.* 1987; Graham 1987, citados en MacNamara 1997; Spolsky 2002 y Lumley 2002 en Fox 2004). Sus reflexiones confirman que en el caso de la validez predictiva resulta muy complicado decidir el criterio que se toma como referencia. Además, el lapso de tiempo dificulta el seguimiento y favorece la aparición de factores externos que pueden distorsionar los resultados previstos. Este campo permanece abierto a nuevas investigaciones que intenten aportar luz al respecto.

En el contexto del aula, un nuevo tipo de examen demostraría su validez concurrente al contrastarlo con otros instrumentos de evaluación utilizados en la clase (otras pruebas ya conocidas, las calificaciones anteriores, incluso la observación sistemática, etc.) y que hayan demostrado su fiabilidad.

Éste es el proceso de validación que se ha llevado a cabo con el C-test y que queda reflejado en la parte experimental de la tesis. Se ha utilizado para ello el paquete estadístico SPSS. Los resultados de las correlaciones con otras pruebas (PAAU, calificaciones en la 2ª Evaluación) pretenden establecer la validez concurrente de la prueba para alumnos españoles de Bachillerato. De los resultados cuantitativos se han deducido las implicaciones pedagógicas pertinentes.

3.2.4. Validez aparente

Una prueba tiene validez aparente si parece medir lo que pretende. Por tanto, este tipo de validez depende de la aceptación de la prueba por parte de los alumnos, profesores o autoridades educativas. Según Hughes (1989: 27), si un examen no convence a los alumnos, su actuación en él no sería la misma que en otras pruebas y, posiblemente, no reflejaría sus conocimientos o grado de competencia:

> A test which does not have face validity may not be accepted by candidates, teachers, education authorities or employers. It may simply not be used; and if it is used, the candidates' reaction to it may mean that they do not perform on it in a way that truly reflects their ability.

En esto radica su importancia, a pesar de que la literatura (Anastasi 1982; Hughes 1989) tache a la validez aparente de concepto poco científico: "is not validity in the technical sense" (Anastasi 1982: 136).

En la práctica docente los profesores son conscientes de la importancia de este tipo de validez. La introducción de nuevos formatos o técnicas de examen puede provocar el rechazo o la desconfianza de los alumnos y de los propios profesores, sobre todo si consideran que no es un instrumento válido de medida. Ésta era una de nuestras preocupaciones a la hora de presentar el C-test a los alumnos, conociendo las afirmaciones de Bradshaw (1990) y Jafarpur (1995), que tachan al C-test de falta de validez aparente. Su reacción ante la prueba podía determinar el éxito o fracaso de la misma. Como recordaremos, la validez aparente es también objeto de una las hipótesis de trabajo (nº 5) que plantea esta tesis.

Para evitar la falta de validez aparente al introducir técnicas de evaluación nuevas, Hughes (1989) recomienda que se expliquen concienzudamente y sin precipitación. También Messick (1996) propone la familiarización como preparación para las pruebas. Intentamos seguir estas pautas para ofrecer a los alumnos una impresión positiva y confiada al presentar la prueba. En un primer momento se les entregó un modelo de C-test ya resuelto, que explicaba el diseño de la prueba y la tarea propuesta. En el caso de las pruebas piloto, no obstante, nos basamos en las impresiones subjetivas de los propios alumnos y del investigador. Sin embargo, en la investigación definitiva se utilizó un cuestionario retrospectivo de opinión para determinar la validez aparente del C-test (véase el capítulo 12).

3.3. Fiabilidad

Según Bachman y Palmer (1996) la fiabilidad es, junto a la validez de constructo, una cualidad fundamental de las pruebas. Ambas son complementarias. La fiabilidad de una prueba viene dada por la *consistencia* de su medida de la actuación del alumno y de la corrección por parte del profesor (Weir 1988; Hughes 1989; Gipps 1994; Bachman y Palmer 1996).

> Reliability is clearly an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure. (Bachman y Palmer 1996: 20)

Hughes (1989: 35) añade: "If a test is not reliable then we know that the actual scores of many individuals are likely to be quite different from their true scores".

Messick (1996: 250) relaciona fiablidad y generalización: "Generalizability as reliability refers to the consistency of performance across the tasks, occasions, and raters of a particular assessment, which might be quite limited in scope".

En la literatura se reconoce que no es posible eliminar totalmente cierta inconsistencia de las pruebas (Gipps 1994; Bachman y Palmer 1996). El carácter humano de la actividad lo impide (Hamp-Lyons 1990). No obstante, Bachman y Palmer (1996) aseguran que un buen diseño reduce la inconsistencia. Para que una prueba sea fiable es necesario que sea realista, y que tenga en cuenta tanto la definición del constructo como la naturaleza de las tareas que propone al alumno.

Gipps (1994) señala que la homogeneidad (valorar una sola habilidad o destreza) contribuye a la consistencia interna de las pruebas. Por otra parte, considera que la estandarización de las pruebas no es apropiada para asegurar la fiabilidad.

Hughes (1989: 29) resume que, en la práctica, una prueba es fiable si garantiza la obtención de unos resultados similares sea cual sea el momento de su realización, es decir, con independencia del momento en que se aplique.

> What we have to do is construct, administer and score tests in such a way that the scores actually obtained on a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability, but at a different time.

Lo que propone Hughes no es tan sencillo como puede parecer, a pesar de que los medios estadísticos actuales faciliten la tarea. De hecho hemos visto que muchos autores señalan la imposibilidad de crear pruebas totalmente fiables.

En el siguiente apartado analizaremos los distintos métodos para valorar la fiabilidad de las pruebas y las dificultades que supone su aplicación en el aula.

3.3.1. Medidas cuantitativas de la fiabilidad

Cuantitativamente podemos hallar el coeficiente de fiabilidad de una prueba (entre -1 y 1). Generalmente se hace mediante el método *test-retest*[25], es decir, aplicando el mismo test al mismo grupo de alumnos en dos ocasiones. Pero este método presenta inconvenientes, como la posible desmotivación del alumno al repetir un examen, o la dificultad para que el lapso de tiempo entre ambas administraciones no suponga un sesgo en los resultados (Hughes 1989; Gipps 1994).

Una segunda posibilidad es crear versiones paralelas de la misma prueba y aplicarlas al mismo grupo o a otro grupo de población similar. En este caso la principal dificultad es garantizar que las dos versiones sean realmente paralelas (Gipps 1994).

Para evitar los problemas del método anterior se puede aplicar otro más económico llamado *split half*[26]. La prueba se subdivide en dos mitades equivalentes y a cada sujeto se le asignan dos puntuaciones, una para cada mitad del examen. De este modo, el alumno solamente realiza una prueba.

Si queremos averiguar hasta qué punto la puntuación obtenida por un individuo concreto se acerca a su puntuación real (*true score*) calculamos el error estándar de la prueba[27]. Un examen se puede considerar fiable si la puntuación de la mayor parte de los alumnos en la prueba es semejante a su puntuación real. En los casos

[25] El coeficiente de correlación (r) entre ambas puntuaciones se calcula con la función estadística correspondiente (Hughes 1989: 158).

[26] En el método *split half* se puede utilizar la fórmula Spearman-Brown:
Reliability of whole test= 2 x coefficient for split halves / 1 + coefficient for split halves
El estudio se puede completar con las medias y desviaciones estándar de las dos mitades. (Hughes, 1989: 158)

[27] El cálculo del error estándar se hace mediante la fórmula: Standard error of measurement= Stand Dev of test x B √1-reliability of test (Hughes 1989: 159).

de discrepancia sería recomendable recopilar más información acerca de la habilidad lingüística del alumno.

Estas medidas estadísticas pueden aportarnos una gran ayuda en la práctica docente para comprobar la consistencia interna de las pruebas, a pesar de las limitaciones prácticas que hemos comentado.

En cuanto a la valoración de la fiabilidad del C-test objeto de nuestro estudio, nos planteamos la utilización de este tipo de medidas cuantitativas. La propuesta de repetición de la misma prueba habría sido rechazada por el alumnado. El diseño de un C-test paralelo nunca está asegurado, por tanto esta idea también fue desechada.

Así pues, quedaba la posibilidad de utilizar el método *split half*. Pero en este caso aparecieron otros inconvenientes; en primer lugar, los derivados del propio diseño del C-test. La segunda mitad de la prueba (ítems 51-100) introducía una novedad y complicación añadida, las omisiones no guiadas. Para obtener dos mitades equivalentes hubo que reorganizar la prueba (véase capítulo 9, apartado 9.10.1). Asimismo, al administrar el C-test comprobamos que algunos alumnos no tuvieron tiempo suficiente para completar el examen, y otros, a pesar de completarlo, mostraron un alto grado de fatiga, que pudo afectar a la resolución de la última parte de la prueba. Con todo, se decidió aplicar este método.

Además, se analizó la fiabilidad del C-test mediante el estudio del Alfa de Cronbach y las correlaciones entre el C-test, las otras pruebas aplicadas (PAAU realizada en clase y PAAU de junio de 2001) y las calificaciones en la 2ª Evaluación.

3.3.2. La fiabilidad de la corrección

Hughes (1989: 36) afirma categóricamente que la fiabilidad de las pruebas también depende del grado de fiabilidad de su corrección: “If the scoring of a test is not reliable, then the test results cannot be reliable either”. Bachman y Palmer (1996: 221) insisten en la misma idea: “one of the most effective ways of dealing with inconsistency is through the proper selection and training of raters”.

Generalmente damos por supuesto que un mismo corrector actúa de forma exactamente igual en todas las ocasiones. Pero esta presunción es precipitada, puesto que en la corrección influyen múltiples factores. Muchos derivan de la condición humana de la actividad (Hamp-Lyons 1990) y entre los más determinantes está el tipo de prueba. Es obvio que las pruebas objetivas garantizan una mayor fiabilidad del corrector ya que no requieren ningún tipo de juicio subjetivo.

Para asegurar la fiabilidad intra-corrector (*intra-rater*), el corrector puntúa la misma prueba en distintos momentos (*mark-remark procedures*). Si existe la posibilidad de que distintos correctores corrijan la misma prueba (*multiple scoring*) comprobaremos su fiabilidad inter-corrector (*inter-rater*).

Hughes (1989) alerta de que es muy probable que el coeficiente de fiabilidad del corrector (de -1 a 1) sólo alcance el 1 en el caso de las pruebas objetivas. Para las pruebas subjetivas, que implican el juicio del corrector (redacciones, cuestionarios, preguntas abiertas, etc.), es casi imposible lograr un coeficiente tan elevado. En estos casos un grado de fiabilidad aceptable sería un >0,7 ó >0,8.

La estimación de la fiabilidad del corrector mediante procedimientos intra e inter corrector resulta fundamental en pruebas de tipo subjetivo, como las redacciones (Gamaroff 2000; Herrera Soler 2000; Amengual Pizarro 2003). De todos modos, no podemos olvidar que, incluso en las pruebas objetivas, es inevitable cierta subjetividad por parte del profesor; no en la corrección pero sí en el diseño de la prueba, selección de su contenido, etc.

Otros sesgos que pueden afectar a la actuación del corrector provienen de las caracterísicas personales del alumno y de la propia prueba; desde la limpieza y claridad de la presentación (Wood 1991) hasta el género (Goddard-Spear 1983; Herrera Soler 2000a) de profesor y alumno, las expectativas (Gipps 1994), la experiencia docente (Hamp-Lyons 1989), etc.

Parece claro que una adecuada formación del profesorado en estos aspectos contribuye a reducir sesgos (Gipps 1994; Cushing 1994; Bachman y Palmer 1996), aunque algunos estudios lo pongan en entredicho (Henning 1996; Weigle 1998).

La importancia de la fiabilidad de los resultados obtenidos con una prueba queda patente en la literatura. Harlen (1994 citado en Gipps 1994) la relaciona con la calidad de la evaluación.

El C-test es una prueba objetiva. Como veremos con mayor detenimiento en la parte experimental, por sus características puede ser considerada fiable. El sesgo del corrector queda reducido al mínimo, porque la tarea propuesta al alumno es la recuperación exacta del texto original y, por tanto, el criterio de corrección es claro. No caben alternativas en las omisiones. En este aspecto, el C-test consigue mayor fiabilidad que las pruebas de cierre tradicionales. La subjetividad queda limitada a la selección de textos y, en nuestro caso, todos los textos utilizados proceden de PAAUs, como criterio unificador de nivel. En el diseño del estudio empírico de esta tesis no planteamos como objetivo el estudio de la fiabilidad del corrector, puesto que el propio diseño de la prueba no lo requiere y, además, el investigador fue también el único corrector de los C-tests aplicados.

3.3.3. Cómo asegurar la fiabilidad de las pruebas

En el apartado 3.3 de este capítulo hemos resaltado que la dificultad para crear pruebas fiables procede del propio carácter humano de la actividad evaluativa. Probablemente no se pueden crear pruebas totalmente fiables, pero sí se puede conseguir un alto grado de fiabilidad. Distintos autores (Jacobs *et al*. 1981; Hughes 1989; Gipps 1994) nos proporcionan algunas medidas o consejos para lograrlo.

Jacobs *et al*. (1981) se refieren en concreto a los pasos que aseguran una corrección fiable de redacciones. Su primera propuesta es la adopción de un enfoque holístico.

Gipps (1994) propone un "control de calidad" en los procesos evaluativos que asegure la consistencia y reduzca los sesgos. En el Reino Unido se denomina *moderation*. Incluye el uso de medios estadísticos[28] (Harlen 1994 citado en Gipps 1994), intervención de la inspección educativa, comparación entre escuelas, grupos de discusión, etc. Todos los sistemas educativos introducen alguna de estas medidas de control, pero no siempre van dirigidas específicamente a los procesos de evaluación. Recomendamos la adopción de este tipo de iniciativas, pero su

[28] Gipps (1994) distingue siete tipos de moderación: *Statistical Moderation Through Use of Reference Tests or Scaling Tecniques, Moderation by Inspection, Panel Review, Consensus Moderation, Group Moderation, Approval of Institutions*, e *Intrinsic Moderation*.

alcance afecta más a la organización general del sistema educativo que a la actuación docente.

En el caso de la Selectividad española sería interesante la creación de un sistema de control específico de calidad de la prueba, como apunta Gipps (1994) y proponen Fernández y Sanz (2005: 25). Sin embargo, la Ley Orgánica de Educación (LOE 2006) todavía no hace mención expresa de estos aspectos, aunque probablemente lo hará en su futuro desarrollo.

El acercamiento de Hughes (1989) es eminentemente práctico. Ofrece a los profesores un repertorio de sencillos consejos para crear pruebas fiables. No son nuevos; la mayoría han ido apareciendo a lo largo de los apartados anteriores. Los conocemos, pero a veces se olvidan en la práctica docente.

No está de más recordar estas recomendaciones que sí son realmente aplicables en el aula de Lenguas Extranjeras. Las que comentamos en primer lugar van encaminadas a proporcionar consistencia a la actuación del alumno en la prueba. Se refieren al diseño de la prueba y a su administración.

En cuanto al diseño de las pruebas, Hughes (1989: 37) indica que una prueba fiable debe incluir un número suficiente de preguntas, no ha de ser demasiado breve ni tan larga que sature o agobie al alumno. No es conveniente que las preguntas planteadas dejen demasiada libertad al alumno. Incluso en las pruebas de redacción se deben proponer tareas precisas y controladas. Además las preguntas no deben ser ambiguas, sino claras y que no permitan interpretaciones.

Con respecto a su administración, recomienda cuidar la tipografía, el orden y la claridad, en definitiva, el aspecto externo del examen. Como hemos visto en el apartado 3.2.4 al comentar la validez aparente, es importante que el alumno esté familiarizado con la técnica y formato del examen, y que el profesor prepare al alumnado para realizarla facilitando instrucciones claras, orales y escritas. Con ello se evita desviar la atención del alumno hacia aspectos no pertinentes y es probable que mejore su actuación. Influyen incluso las condiciones de administración del examen: la adecuada duración de la prueba, el lugar, la luz, condiciones acústicas, temperatura, silencio, etc.

Las que desglosamos a continuación se refieren a la fiabilidad del corrector. Según Hughes (1994: 40) sería conveniente plantear ítems objetivos, pues reducen la subjetividad del profesor (véase apartado 3.3). Un instrumento que sirve de ayuda

es el manejo de plantillas de corrección que detallen lo que se considera correcto y/o aceptable en cada pregunta, así como el baremo asignado a cada parte del examen. Si se trata de varios correctores, es necesario un acuerdo en los criterios de puntuación. Y siempre que sea posible, la doble corrección reduce el error.

Estas últimas ideas no van dirigidas al aula; por su elevado coste, en la práctica, sólo se pueden aplicar en el contexto de exámenes a gran escala, como las PAAU. E incluso en las PAAU la doble corrección tiene un carácter excepcional, puesto que solamente se lleva a cabo previa solicitud expresa del alumno, una vez conocidos los resultados obtenidos en la primera corrección.

Lo ideal sería que los correctores supieran exactamente cómo corregir la prueba, que tuvieran un entrenamiento previo y hubieran demostrado su solidez como correctores.

Igualmente, lo deseable sería no conocer a los que se presentan al examen en el caso de exámenes oficiales (en la PAAU este aspecto sí se cumple, se identifica al alumno mediante un código de barras para asegurar el anonimato y la objetividad del corrector) o de diagnóstico, para que las expectativas del profesor no influyan en los resultados. En el contexto del aula el profesor conoce a sus alumnos y aunque ello puede afectar a la corrección, sobre todo en las pruebas de tipo subjetivo, normalmente las repercusiones de las pruebas son menores.

3.4. Tensión validez-fiabilidad

Hemos insistido en que validez y fiabilidad son las dos cualidades básicas o esenciales de las pruebas. En los apartados anteriores hemos analizado sus respectivas características. No obstante, Bachman (1990: 241) asegura que no siempre es fácil distinguirlas: “the point at which we “draw the line” may be somewhat arbitrary”.

Hughes (1989) y Weir (1988, 1993) señalan que una prueba sólo puede ser válida si es fiable: “To be valid a test must provide consistently accurate measurements. It must therefore be reliable“ (Hughes 1989: 42). Por el contrario, puede ser fiable y no válida, si no mide el constructo que pretende medir.

Como hemos comentado en epígrafes previos, Davies (1978), Weir (1988), Gipps (1994) y Hughes (1989), entre otros, consideran que existe tensión entre validez y fiabilidad. Con frecuencia es necesario sacrificar parte de una de ellas a favor de la otra, aunque lo recomendable sería el equilibrio entre ambas.

Veamos un par de comentarios al respecto. Hughes (1989: 42) anuncia: “There will always be some tension between reliability and validity. The tester has to balance gains in one against losses in the other”. También Gipps (1994: 76) recalca la necesidad de buscar el equilibrio cuando comenta: “What is needed, of course, is an appropiate balance between the two because they are in tension”.

Este equilibrio se logra cuando la definición del constructo que mide la prueba es clara. Para Harlen (1995), es la calidad de la prueba lo que equilibra la tensión entre validez y fiabilidad. Según Nuttall (1987), citado en Gipps (1994), el concepto de generalizability es el nexo entre validez y fiabilidad. La teoría tradicional de la evaluación se basa en la generalización de los comportamientos. Nuttall alude a la necesidad de ambas cualidades para que los resultados de las pruebas sean generalizables.

En esa línea, Linn et al. (1991) aseguran que el concepto de fiabilidad debería ser ampliado: “We need also to enquire whether we can generalize from the specific assessment task to the broader domain of achievement” (citado en Gipps 1994: 77).

Aunque tradicionalmente la validez se ha considerado más importante que la fiabilidad (Guildford 1965 en Weir 1888; Gipps 1994), la tendencia actual, liderada por Bachman y Palmer (1990: 239), las entiende como complementarias: “The investigation of reliability and validity can be viewed as complementary aspects of identifying, estimating, and interpreting different sources of variance in test scores”.

Los citados autores insisten en integrar también al resto de los rasgos de las pruebas: “The most important consideration to keep in mind is not to ignore any one quality at the expense of others” (op. cit.: 38).

3.5. Autenticidad

Bachman y Palmer (1996) describen la autenticidad y el carácter práctico como cualidades críticas de las pruebas a las cuales no se ha dado la importancia que merecen, a pesar de que la preocupación por la autenticidad sí está presente en las investigaciones sobre evaluación de la lengua.

Según Bachman (1990), fue Carroll (1961) quien sembró la semilla de la autenticidad en la evaluación al pedir para las pruebas de lengua el "total communicative effect of an utterance". Este deseable efecto comunicativo implica "funcionalidad" (*illocutionary purpose*) como base de la autenticidad.

Varias décadas después, Fulcher (1999a: 222) ofrece una definición general de autenticidad: "The degree to which sampling is successful is frequently expressed as the degree to which the test is "authentic". [...] The term "authenticity" in language testing has therefore come to mean the degree to which the outside world is brought into the testing situation".

Bachman (1990) distingue en la literatura dos acercamientos al concepto de autenticidad. El primero se denomina "*real-life approach*" y se basa en la precisión con que la actuación en la prueba predice la actuación en situaciones comunicativas de la vida real[29]. El segundo es el "*interactional/ability approach*", que intenta medir la lengua como habilidad mental. Se basa en la interacción entre el examinando, la tarea que le plantea la prueba y el contexto.

Bachman y Palmer (1996) revelan que la importancia de la autenticidad radica en que establece una relación entre la tarea concreta que propone la prueba y el dominio al que se refiere. Es decir, la autenticidad indica la correspondencia entre el uso real de la lengua (TLU) y las tareas concretas propuestas en un examen. Según los citados autores (1996: 23), si existe esta correspondencia se puede considerar que el examen es relativamente auténtico: "We define *authenticity* as the degree of correspondence of the characteristics of a given test to the features of a TLU task".

Hemos comentado que los rasgos de las pruebas no son independientes. Bachman y Palmer (1996) señalan la relación entre autenticidad y validez de

[29] El enfoque "*real-life*" se relaciona con la dicotomía entre pruebas directas e indirectas. Bachman (1990) define las pruebas de lengua como "indicadores indirectos de las habilidades que interesan al examinador".

constructo, puesto que una parte de la validación de constructos está basada en la posibilidad de generalizar la interpretación de los resultados de las pruebas.

Para Messick (1996: 234):

> authentic assessments pose engaging and worthy tasks (usually involving multiple processes) in realistic settings or close simulations so that the tasks and processes, as well as available time and resources, parallel those in the real world.

La autenticidad afecta también a la validez aparente (aspecto fundamental en el *real-life approach*). Una prueba siempre es una propuesta artificial de tareas que simula con mayor o menor éxito situaciones reales de comunicación.

Bachman y Palmer (1996: 24) apuntan la relación entre autenticidad y percepción del alumno (*face validity*) con respecto a la prueba: “It is this relevance, as perceived by the test taker, that we believe helps promote a positive affective response to the test task and can thus help test takers perform at their best”. El hecho de que el alumno considere un examen relevante y adecuado (por el tema, tipo de tareas, etc.) ayuda a predisponerle positivamente hacia él e, indudablemente, influye en su actuación.

Como primer paso para el diseño de pruebas auténticas, Bachman y Palmer (1996) proponen identificar los rasgos que definen las tareas en el dominio de la lengua que se quiere evaluar.

Al igual que con otras cualidades de las pruebas, no es fácil establecer los límites de la autenticidad. Shohamy y Reves (1985) señalan que la propia situación de la prueba y la relación examinador-examinando son artificiales; en palabras de Bachman (1990: 319), constituyen una amenaza: “a potential threat to authenticity” pero el profesor puede minimizarla creando “a testing environment that will promote authetic interaction”.

Messick (1989) alerta del peligro de que la validez quede reducida a la autenticidad, sería una visión simplista. Recordamos que su propuesta es un marco unitario de validez. Una prueba no es válida sólo porque parezca auténtica, aunque la autenticidad se haya convertido, según sus propias palabras, en señal de “buena práctica” en la evaluación.

En el contexto de la enseñanza del Inglés para Fines Académicos (EAP) la autenticidad de las pruebas adquiere un papel aún más destacable, según Fulcher (1999a), relacionada con la validez de contenido y aparente.

Spolsky (1985: 39) resume del siguiente modo la importancia de la autenticidad en la evaluación de la lengua: “In sum, the criterion of authenticity raises important pragmatic and ethical questions in language testing. Lack of authenticity in the material used in a test raises issues about the generalizability of results”.

No resulta sencillo determinar el grado de autenticidad de una prueba objetiva, como el C-test. Como se verá en capítulos posteriores con mayor profundidad, el C-test es una prueba de redundancia reducida. Podemos considerar que propone una tarea habitual en la comunicación lingüística diaria y, por tanto, auténtica. Consiste en suministrar la información que se pierde en el acto de comunicación haciendo uso de la gramática de expectativas para superar así los ruidos en el canal (Stevenson 1977 citado en Klein-Braley 1985).

Otro aspecto que deberíamos controlar es la autenticidad de los textos en que se basa la prueba (Raatz 1985; Klein-Braley 1985). En nuestro caso decidimos crear el C-test a partir de textos aparecidos en PAAU recientes. Estos textos, en principio, son auténticos. Pero, dependiendo del nivel de competencia lingüística del alumnado al que van dirigidos, quizá este rasgo debería sacrificarse a favor del tema, por ejemplo. En este punto entraríamos en la compleja discusión de la autenticidad de materiales.

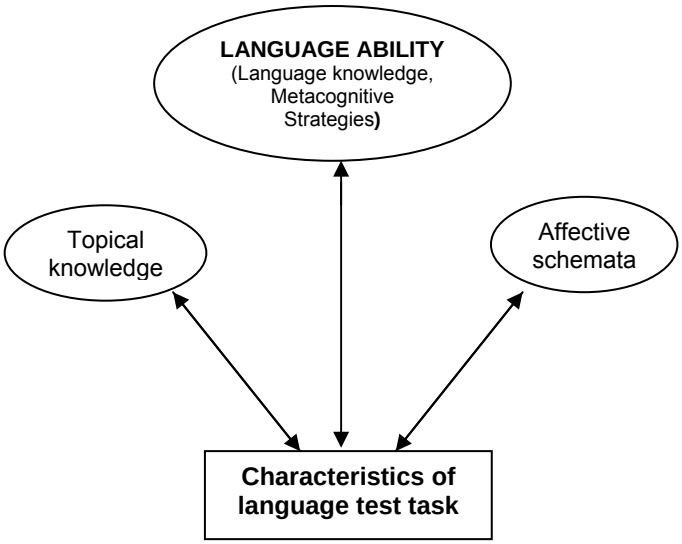
3.6. Carácter interactivo

Según Bachman y Palmer (1996: 25) el carácter interactivo de una prueba viene dado por “the extent and type of involvement of the test taker’s individual characteristics in accomplishing a test task”.

Este rasgo supone el grado y tipo de implicación que un examen demanda de parte del alumno. Por un lado, el examen mide sus conocimientos y competencia lingüística, por otro está presente la implicación afectiva del alumno (interés despertado, motivación, etc.) y su conocimiento del tema (*topical knowledge*).

El siguiente gráfico representa las interacciones entre habilidad lingüística, conocimiento del tema y esquemas afectivos.

Figura 3.4. *Interactiveness.* Bachman y Palmer (1996:26)



Como ocurría con la autenticidad, Bachman y Palmer relacionan el carácter interactivo con la validez de constructo: “it is this quality that provides the vital link with construct validity” (op. cit.: 26).

De manera que se genera una relación de dependencia: “Authenticity, interactiveness, and construct validity all depend upon how we define the construct “language ability” for a given test situation” (op. cit.: 29).

Ambos conceptos, autenticidad y carácter interactivo son relativos, y para determinar su grado en una prueba hemos de tener en cuenta las características de los alumnos, de la tarea y del constructo de la lengua que se intenta medir. Cada situación de evaluación requerirá unos niveles específicos de autenticidad y carácter interactivo, siempre en equilibrio con las demás cualidades de las pruebas.

El C-test ha demostrado implicar al alumno en la tarea propuesta. Para resolverlo ha de aplicar sus conocimientos de la lengua meta. Por otra parte, en el C-test es importante el tema del texto base. Sería difícil determinar hasta qué punto ayuda cada uno de estos aspectos a la resolución de la tarea.

Al valorar la implicación afectiva del alumno, como hemos indicado, nos referimos a la validez aparente de la prueba. Si bien se ha cuestionado la validez

aparente del C-test (Weir 1988; Bradshaw 1990; Jafarpur 1995), en este estudio hemos apreciado que, en términos generales, el carácter fragmentario (*puzzle-like*) de la tarea supone un elemento motivador, casi un reto para el alumno. En consonancia con otros estudios (Klein-Braley 1997) podemos decir que la prueba despertó el interés de nuestros alumnos.

3.7. Factibilidad

Tanto para los profesores como para las autoridades educativas es importante que una prueba permita, por sus características, su aplicación en la situación para la que haya sido creada. Por ello, aunque la naturaleza de esta cualidad es diferente de las anteriores, debe ser tenida en cuenta en el diseño de pruebas, pues afecta a todas las decisiones. En palabras de Weir (1988: 37): “A valid and reliable test is of little use if it does not prove to be a practical one”.

Según Bachman y Palmer (1996: 35), para determinar la factibilidad de una prueba deberíamos evaluar tanto los recursos que requiere su diseño como los que implica su aplicación y compararlos con los recursos de que realmente disponemos: “for any given situation, if the resources required for implementing the test exceed the resources available, the test will be impractical and will not be used”. Aportan una definición de factibilidad que relaciona los recursos empleados en el desarrollo de la prueba con el uso de la misma: “We can define practicality as the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities” (op. cit.: 36).

Cuando Bachman y Palmer (1996) hacen referencia a los “recursos disponibles” incluyen todo tipo de recursos humanos y materiales, que finalmente se traducen en un coste económico concreto para cada situación de evaluación:

- Recursos humanos: profesores, creadores de pruebas, correctores, administradores de las pruebas, e incluso personal técnico, etc.
- Recursos materiales: espacio (aulas), materiales (papel, bibliografía, etc.), recursos técnicos (ordenadores, cintas de video y audio, CD-rom, DVD, proyectores, etc.).

- Tiempo: desde que se inicia la creación de la prueba hasta que se completa su aplicación y corrección.

Lógicamente, si una prueba resulta práctica nos veremos más inclinados a utilizarla, e incluso a investigar para mejorarla. Ahora bien, el carácter práctico ha de ir unido al resto de las cualidades de las pruebas, como nos recuerda Hughes (1989: 47): "Other things being equal, it is good that a test should be easy and cheap to construct, administer, score and interpret".

El C-test es una prueba cuyo carácter práctico es indiscutible. Así se ha reconocido en la literatura (Carol y Chapelle 1990; Connelly 1997; Klein-Braley 1997). Es destacable su economía en términos de tiempo. Su creación no requiere un tiempo excesivo. Sólo implica tomar decisiones en cuanto a los textos que servirán como base para las omisiones. La administración es sencilla, sobre todo si el alumnado está familiarizado con la técnica. Finalmente, la corrección no plantea problemas; el criterio es claro y su carácter objetivo hace que se corrija rápidamente. Como se demostrará posteriormente, el C-test resulta una prueba muy rentable para medir la competencia global en Lengua Extranjera.

3.8. Impacto

Comenzamos este apartado con una aproximación al concepto de impacto. La definición más general lo describe como el efecto de las pruebas en la enseñanza y el aprendizaje. Veremos las distintas denominaciones que aparecen en la literatura para designar a este fenómeno comúnmente aceptado. Una de ellas, validez consecuencial, lo relaciona directamente con la validez de constructo (Messick 1989). Continuaremos con algunas referencias en torno a las escasas investigaciones empíricas sobre el efecto rebote.

Puesto que el impacto de las pruebas se manifiesta en los individuos (alumnos y profesores) y en la sociedad, veremos de qué forma afecta a cada uno de ellos.

El efecto de las pruebas puede ser positivo o negativo. Expondremos brevemente el repertorio de consejos de Hughes (1989) y Bailey (1996) para lograr un efecto rebote beneficioso.

3.8.1. Definición del concepto

Otra cualidad innegable de las pruebas es el efecto que producen "on teaching and learning" (Hughes 1989: 1; Shohamy *et al*. 1996: 298), o dicho de otro modo: "on society and educational systems and upon the individuals on those systems" (Bachman y Palmer 1996: 29) .

Un examen no es un ente abstracto, sino algo real que se realiza con un propósito, en un contexto y con unos individuos concretos. Por tanto, como asegura Bachman (1990: 279), afecta a todos los elementos que de una u otra forma intervienen en él: "Tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or of society at large".

El efecto de las pruebas puede ser de distinto signo, tal como expone Buck (1988: 18)*: "*This washback effect can be either beneficial or harmful*"*. Y la responsabilidad del profesor es conseguir que los efectos sean beneficiosos (Hughes 1989; Hamp-Lyons 1997).

La literatura reconoce la existencia e importancia del impacto, las pruebas de evaluación realmente influyen en la enseñanza y aprendizaje. Tanto, que en 1996 la revista *Language Testing* dedicó un volumen monográfico a su estudio: *LT* 13 (3). A pesar de todo, la investigación empírica sobre la naturaleza y mecanismos del impacto es todavía escasa (Bailey 1996).

En primer lugar hagamos las precisiones terminológicas pertinentes. Este fenómeno se conoce con diversos nombres: Baker (1971) usa el término *test impact* (impacto de las pruebas), Hughes (1989) y Bailey (1996), entre otros, prefieren denominarlo *washback* o *backwash* (que normalmente traducimos como "efecto rebote"), Messick (1996) habla de *consecuential validity* (validez consecuencial) y Frederiksen y Collins (1989) de *systemic validity* (validez sistémica).

Las dos primeros términos; *impacto* y *efecto rebote*, son los más frecuentes en la literatura para describir los efectos de las pruebas. Generalmente aparecen como sinónimos, aunque hemos de hacer constar que autores como Hamp-Lyons (1997: 298) y Davies (2003) precisan ambos términos con mayor rigor. Aportamos la distinción de Davies (2003: 361): "impact is taken to be the superordinate while washback refers to the narrower situation of the language classroom".

Impacto y efecto rebote serán los que más utilicemos en esta tesis, indistintamente, y como sinónimos de “influencia”. No consideramos necesario mantener la precisión de Davies (2003) puesto que no encontramos una postura unificada en la literatura. El contexto servirá para determinar en cada caso el ámbito de influencia de las pruebas. Por otra parte, en español los términos “impacto”, “efectos”, “consecuencias” e “influencias” nos parecen léxicamente muy próximos y siempre de fácil comprensión.

A continuación abordamos la procedencia de las otras dos denominaciones; validez consecuencial y sistémica. Ambas tienen su origen en el marco de validez de constructo como concepto unitario propuesto por Messick (1989; 1996) y expuesto en el apartado 3.3 de este mismo capítulo. El propio autor explica que: “In the context of unified validity, evidence of washback is an instance of the consequential aspect of construct validity” (Messick 1996: 254).

Una de las características de esta concepción de validez es su amplitud, pues da cabida a las consecuencias de la interpretación de las pruebas y sus implicaciones educativas y/o sociales: “Consequences associated with testing are likely to be a function of numerous factors in the context or setting and in the persons responding as well as in the content and form or the test” (op. cit.: 251).

Según Messick, si una prueba es válida lo más probable es que produzca efectos beneficiosos.

Partiendo del marco de Messick, Gipps (1994) y MacNamara (1997) hablan de validez consecuencial y un subtipo de ésta, la validez sistémica (Fredericksen y Collins 1989). La validez sistémica se refiere las consecuencias de las pruebas en el sistema educativo en que se desarrollan y aplican. MacNamara (1997) incluye al efecto rebote (*washback*) como un conjunto de aspectos dentro de la validez consecuencial.

Davies (2003: 363) considera que en los últimos años se ha sobrevalorado la preocupación por el impacto de las pruebas. Llega a denominar “testing heresy” a la excesiva preocupación social por cuestiones éticas, de impacto y políticas relacionadas con los exámenes, y aconseja una vuelta a lo fundamental “to restore language to the centre of language testing”.

Como indicamos en su momento, podríamos haber incluido el efecto rebote como subapartado de la validez, siguiendo a los autores mencionados. Pero la

importancia y amplitud del tema nos anima a dedicar todo este apartado al impacto de las pruebas, tanto en el contexto del aula y los programas como en el sistema educativo.

3.8.2. El impacto de las pruebas en el enfoque comunicativo

A veces, las pruebas producen un efecto rebote negativo, sobre todo las estandarizadas y externas. Esto se debe a la discrepancia entre las mismas y el enfoque real de la enseñanza de idiomas.

En la actualidad sigue vigente el enfoque comunicativo en la enseñanza de lenguas. El aprendizaje de una lengua tiene como objetivo la capacitación del alumno para comunicarse efectivamente en ella. Sin embargo, muchas pruebas estandarizadas miden la competencia lingüística tradicional y no la competencia comunicativa.

Morrow (1991) citado en Bailey (1996: 112) confirma la necesidad de coherencia entre enseñanza y evaluación: “this conscious feedback loop between teaching and testing, in terms not only of content but also of approach, is a vital mechanism for educational development”.

Los trabajos del *Ontario Institute for Studies in Education* (OISE) sobre desarrollo de pruebas de evaluación reconocen que el objetivo fundamental de las pruebas comunicativas ha de ser promover un efecto rebote beneficioso: “communicative tests should be explicitly designed to bring about positive washback” (Bailey 1996: 261).

3.8.3. Investigación empírica sobre el impacto

La investigación empírica sobre el impacto va encaminada a aclarar los mecanismos de funcionamiento del efecto rebote de las pruebas. En el apartado 3.8 hemos constatado la relativa escasez de tales estudios, que radica en la dificultad para su realización.

Bailey (1996) reconoce que no es fácil aislar el efecto rebote para su estudio, puesto que suele estar íntimamente ligado a otros aspectos de la enseñanza y aprendizaje. Por ello, como veremos, Alderson y Wall (1993) aconsejan el método de la observación directa y la triangulación[30] para los estudios del impacto.

Bailey (1996) cita un trabajo sin publicar de Hughes (1993) que distingue tres aspectos en el proceso de enseñanza: participantes, proceso y producto final. Según el modelo de Hughes, las pruebas pueden afectar a todos ellos.

No podemos evitar hacer mencionar los trabajos empíricos de Alderson y Wall (1993), Shohamy (1993), Shohamy *et al*. (1996), Alderson y Hamp-Lyons (1996), y el más reciente de Andrews *et al*. (2002), ampliamente reconocidos en este campo. La clave para estas investigaciones fue la observación directa de las clases. También se analizaron los materiales y se recopiló información de los participantes a lo largo de varios años, mediante el uso de entrevistas y cuestionarios.

Alderson y Wall (1993a: 120-21) establecieron un listado previo de quince posibles hipótesis relativas al efecto rebote. Llegaron a la conclusión de que se necesitaba seguir investigando directamente en las aulas e incluir las áreas de "motivation and performance, as well as educational innovation" (Bailey 1996: 263).

El estudio empírico de Alderson y Wall (1993b) en Sri Lanka supuso un hito. Describe la influencia de un nuevo examen de inglés en las clases de inglés de secundaria. La Universidad de Lancaster había recibido el encargo de evaluar la validez y fiabilidad de la prueba, así como su efecto rebote en las clases de Inglés.

El estudio parte de la descripción minuciosa del contexto educativo en que se aplicaba la nueva prueba y de lo que implicaba para el alumno su superación. Además se revisaron las características intrínsecas de la prueba; su validez, fiabilidad, etc.

Con este examen las autoridades educativas pretendían introducir el enfoque comunicativo en Sri Lanka, lo que debería suponer algunos cambios en los contenidos y en la metodología. Los investigadores eran conscientes de que los efectos no serían totalmente positivos ni negativos. Finalmente, el estudio confirmó la existencia de efecto rebote en el contenido de la enseñanza y en el diseño de las

[30] La triangulación es un método etnográfico que consiste en utilizar dos o más perspectivas (teorías, investigadores, informantes, datos, etc.) para investigar un fenómeno determinado. En los estudios sobre *washback* supone incluir, al menos, la percepción de alumnos y profesores acerca de los efectos de las pruebas.

pruebas, pero no en la metodología seguida por los profesores. Este resultado contrasta con el obtenido por Alderson y Hamp-Lyons (1996) al analizar el efecto rebote en los cursos de preparación de TOEFL. Según Alderson y Hamp-Lyons el TOEFL afecta a lo que se enseña y a como se enseña.

Alderson y Wall (1993b: 220) insisten en que una prueba por sí misma "cannot reinforce an approach to teaching that the educational system has not adequately prepared its teachers for". De nuevo reflejan la necesidad de una preparación adecuada del profesorado, ya comentada en otros apartados de esta tesis.

Citamos textualmente la conclusión final del estudio de Alderson y Wall (ibíd.):

> Testers need to pay much more attention to the washback of their tests, but they should also guard against oversimplified beliefs that "good" tests will automatically have "good" impact. Washback needs to be studied and understood, not asserted.

Los resultados del estudio de Shohamy *et al.* (1996) en el sistema educativo israelí evidencian el efecto rebote de las pruebas de carácter nacional (*high-stakes tests*):

> it had a considerable effect on classroom activities and on time allotment; it also influenced both content and methodology. Ample new commercial teaching material was published and marketed, designed specifically for the test... (Shohamy 1997: 346)

Según Shohamy *et al.* (1996) el efecto rebote de una prueba es un fenómeno complejo porque depende de numerosos factores: el estatus de la lengua meta, el propósito de la prueba, su formato y las destrezas que mide. Y no tiene porqué ser estable, sino que puede cambiar con el tiempo.

El trabajo de Andrews *et al*. (2002) pone de relieve una característica más del efecto rebote, su *unpredictability.* Considera que los cambios que se producen en el aula a partir de la introducción de una prueba son impredecibles, debido en gran parte a las diferencias individuales entre profesores y alumnos.

3.8.4. El impacto en los individuos: alumnos y profesores

Como hemos visto en el apartado 3.8.1, el efecto de las pruebas opera en dos niveles: el micro nivel de los individuos y el macro nivel de la sociedad o el sistema educativo (Bachman y Palmer 1996)[31].

En este epígrafe examinaremos algunos aspectos de la influencia de las pruebas en los individuos, principalmente alumnos y profesores. En los apartados anteriores hemos constatado su existencia, a pesar de las diferencias individuales. Después, en el siguiente punto, haremos lo mismo con los efectos en la enseñanza y el sistema educativo.

El alumno, o de forma más general, cualquier persona que realiza una prueba puede verse afectado por la propia experiencia de su preparación y realización, por los resultados obtenidos en ella y por las consecuencias derivadas de dichos resultados. Los efectos son mayores en las pruebas selectivas estandarizadas de ámbito nacional, como la PAAU española. Su preparación puede suponer largo tiempo de enseñanza dirigida específicamente a entrenar las destrezas que permitan superarla. Para Bachman y Palmer (1996: 31): “teaching may be focused on the syllabus of the test for up to several years before the actual test, and the techniques needed in the test will be practiced in class”.

La realización de la prueba puede afectar al alumno de múltiples formas. Por ejemplo, su contenido (información o temas nuevos) puede despistar al alumno o llevarle a confusión, las características de la prueba pueden permitirle o no el uso de estrategias, etc.

Bachman y Palmer (1996: 32) sugieren que se implique al alumno en el diseño de las pruebas, así aumentará la motivación y por tanto será más fácil que el efecto rebote sea positivo: “one way to promote the potential for positive impact is through involving test takers in the design and development of the test, as well as collecting information from them about their perception of the test and test tasks”. Shohamy (1997) respalda esta idea de trabajo conjunto para llegar a modelos de evaluación justos y democráticos.

[31] Recordamos su definición del efecto rebote como la influencia de las pruebas “on society and educational systems and upon the individuals on those systems” (Bachman y Palmer 1996: 29). Se corresponde con la distinción entre *impact* y *washback* propuesta por Davies (2003).

También la información sobre los resultados de la prueba y las decisiones que conllevan afectan a quienes la han realizado. Por ello, es conveniente que sea una información completa, relevante y significativa para que los efectos sean positivos, una vez más.

En cuanto a las decisiones, hemos de procurar que sean justas, pues está claro que las consecuencias pueden ser importantes en la vida de los alumnos (sobre todo en el caso de exámenes externos que condicionen el acceso a determinados estudios o programas).

El otro grupo de individuos que se ve afectado directamente por las pruebas es el de los profesores. Generalmente el colectivo es consciente de ello; cada profesor sabe hasta qué punto su actuación en el aula está condicionada por los exámenes, sobre todo los externos: "if they find that they have to use a specified test they may find "teaching to the test" almost unavoidable" (Bachman y Palmer 1996: 33).

En el siguiente epígrafe analizamos el fenómeno conocido como "enseñar para el examen". Veremos que, en determinadas circunstancias, el profesor antepone la preparación para una prueba a sus propios valores y concepciones de la enseñanza.

3.8.5. El impacto de las pruebas externas en la enseñanza: enseñar para el examen

Aunque toda prueba tiene unos efectos en el proceso de enseñanza-aprendizaje, debido a sus consecuencias, son las pruebas estandarizadas externas a gran escala (*high-stakes tests*) las que más ponen de manifiesto el efecto rebote (Shohamy *et al.* 1996). En estos casos es fácil constatar el impacto en el sistema educativo y en la sociedad, además de los lógicos efectos en profesores y alumnos.

Los alumnos saben lo que supone su actuación en el examen (calificación, obtención de un título, asignación a un grupo o nivel, posibilidad de acceder a la universidad, consecución de un empleo, etc.) y los profesores deberían ayudarles a enfocar la preparación para afrontarlo.

En nuestro país se administran pruebas estandarizadas de Inglés como Lengua Extranjera (PET, First Certificate, TOEFL, EOI, etc.) cuya superación

permite la obtención del título correspondiente. Con frecuencia se desarrollan cursos específicos dedicados a su preparación. Se podrían analizar los efectos que producen en todos los elementos implicados, pero tal análisis no tiene cabida en el presente trabajo.

Sí queremos hacer notar el impacto de la PAAU en el contexto de las Enseñanzas Medias y en la sociedad española en general. Es una prueba nacional, oficial y estandarizada cuya administración se produce fuera del medio escolar (en instalaciones universitarias). De los resultados obtenidos en ella depende en gran medida el futuro del examinando. El impacto social es evidente, ya que la prueba actúa como puerta para el acceso a la Universidad de los candidatos.

Sus repercusiones son tales que los cambios propuestos por las distintas administraciones educativas provocan todo tipo de reacciones. Así ocurrió cuando se anunció la sustitución de la PAAU por la Prueba General de Bachillerato (PGB) o Reválida, regulada después por el R. D. 1741/2003.

Tanto las editoriales como los profesores de Enseñanza Secundaria quedaron expectantes ante las novedades. Citamos la fundamental para la asignatura de Inglés: “El ejercicio correspondiente a la lengua extranjera tendrá una parte oral y otra escrita”.

Es fácil deducir que la entrada en vigor de la nueva prueba habría supuesto cambios importantes en el currículo de la asignatura, en los materiales, probablemente en la metodología de la clase de inglés, etc. Los profesores de la asignatura habrían ampliado el enfoque de las clases para preparar adecuadamente la prueba, cuya implantación quedó finalmente frenada por las circunstancias pollíticas, que prolongaron la vigencia de las actuales PAAU.

En el contexto educativo del Bachillerato los efectos de la PAAU son patentes. Nuestra experiencia en las aulas nos permite apreciar que la existencia de este examen afecta a todos los agentes implicados: a los programas educativos, el profesorado, los métodos utilizados en el aula, los materiales, los alumnos, etc.

Las clases de Inglés de Bachillerato, especialmente las de segundo curso, se enfocan hacia la superación de la PAAU. En otras materias ocurre de forma similar, puesto que la PAAU no es una prueba específica de lengua inglesa.

Este fenómeno, que los docentes reconocemos en el aula, se conoce en la literatura como "*teaching to the test*"[32] (Gipps 1994) o "*test-like teaching*" (Shohamy 1997) y podría considerarse, en cierto modo, como un tipo "peculiar" de impacto. En el apartado anterior hemos visto que el examen condiciona claramente la actuación del profesor, en un sentido u otro. Es bien conocido en los Estados Unidos y en el Reino Unido, aunque a veces se evite la denominación anterior y se considere simplemente "*preparation for examinations*".

Puede ser una actividad útil (Linn 1981 en Gipps 1994) siempre que las habilidades o destrezas que se entrenan sean transferibles a otras situaciones. En caso contrario la prueba pierde todo su valor como instrumento de medida. Lo interesante es que se enseñe el constructo objeto de examen (*skills and knowledge measured by the test*), no las respuestas a las preguntas concretas. Además *enseñar para el examen* incluye otras prácticas, como las dirigidas a aumentar la motivación de los alumnos o a reducir la ansiedad ante el examen.

Según Smith (1991) citado en Gipps (1994), *enseñar para el examen* es una reacción de los profesores ante la perdida de autoestima que supone la obtención de malos resultados en una prueba importante para los alumnos.

Son ilustrativas las palabras de Shohamy (1997: 346) refiriéndose a la investigación de Shohamy *et al*. (1996) sobre el efecto rebote de una prueba en Israel: "Most teachers reported high anxiety, fear and pressure to cover the material as they felt that the success or failure of their students reflected on them".

Smith (1991) llegó a la conclusión de que el impacto de la prueba no viene dado por su forma sino por el uso político o social de los resultados obtenidos en ella. Debemos tener cuidado para que el interés de las pruebas estandarizadas no nos haga perder la perspectiva y olvidar los propósitos educativos de la evaluación.

Tampoco podemos ignorar que las pruebas son también un instrumento de política social (Shohamy 1997, 2001), símbolo de orden y control (Gipps 1994; Kunnan 1999) y agentes de cambio (Qi 2005). En ello radica la preocupación por la ética de las pruebas. Las de ámbito nacional se pueden utilizar para introducir

[32] El fenómeno descrito como "enseñar para el examen" también se denomina en los Estados Unidos "*test score pollution*". Los resultados de las pruebas se pueden ver contaminados (mejorados) por el entrenamiento previo (Linn 1981). Haladnya (1991) encuentra tres fuentes de contaminación con "*pervasive effects*": la forma de preparación, las condiciones de administración y otros factores externos que los profesores no pueden controlar (familiares, lengua materna, etc.)
Un ejemplo bien estudiado es el *"Lake Wobegon" Effect* (Gipps 1994: 47).

cambios en el sistema educativo, hecho que no siempre va unido a los cambios en el currículo o a una adecuada formación del profesorado. El citado estudio de Shohamy *et al*. (1996) sobre el impacto de dos pruebas nacionales en Israel corrobora la afirmación anterior. Es una forma poco ética de utilizar las pruebas para conseguir otros fines, totalmente ajenos a los educativos.

> Policy makers in central agencies use tests in these ways to manipulate educational systems, to control curricula, to create new knowledge, to define knowledge and to impose new textbooks, to communicate educational agendas and new teaching methods. (Shohamy 1997:346)

Finalizamos este epígrafe con la reflexión de Gipps (1994: 57) acerca del poder de la evaluación y las perspectivas de futuro:

> These detailed accounts of the impact of testing on curriculum, teaching, school systems, pupil motivation and teacher's practice should leave us in no doubt as to the *power of testing*, particularly high-stakes testing, to affect teaching and learning.
> A broader conceptualization of assessment within the educational assessment paradigm is certainly part of the way forward.

Visto el impacto de los exámenes en la enseñanza, claro reflejo del poder de las pruebas, no nos queda más que proponer que todo examen, especialmente los de tipo oficial y estandarizado, sea fiel reflejo de unos objetivos educativos y que reúna en equilibrio todas las cualidades deseables para una prueba. De este modo el inevitable impacto en la enseñanza será siempre beneficioso.

3.8.6. Cómo conseguir que el efecto rebote sea beneficioso

Hemos visto que toda prueba, hasta la más sencilla que se aplica en clase, afecta a la enseñanza y el aprendizaje, es decir, a las personas que se ven implicadas en ella. El reto para los profesores es conseguir que ese efecto sea beneficioso (Hughes 1989). Destacamos el interés para la práctica docente de las recomendaciones que detallamos en este epígrafe.

Hughes (1989) aconseja tener en cuenta unas ideas básicas en el diseño de las pruebas para promover *achieving beneficial backwash*.

Para empezar, deberíamos asegurarnos de que el examen mida aquellas destrezas que nos interese potenciar y de que proporcione un buen número de tareas, basadas en textos lo más auténticos posible.

El contenido no debe ser totalmente predecible. Además debe tener como referencia unos objetivos conocidos también por los alumnos, han de saber qué se les pide.

Es conveniente que se evalúen directamente las destrezas que interesa aprender (*direct testing*), por ejemplo, si se pretende aprender a redactar en lengua inglesa, la prueba deberá incluir una redacción. También es importante que los profesores tengan la asistencia necesaria e incluso se les facilite formación específica en evaluación.

Bailey (1996) aporta las siguientes sugerencias para lograr efectos beneficiosos a partir de las pruebas:

- Fijar objetivos de aprendizaje,
- potenciar la autenticidad de las pruebas,
- introducir formas de autoevaluación que promuevan la motivación y la autonomía del alumno,
- aportar información detallada de los resultados de la evaluación.

Vemos que coinciden con otros autores, como Bachman y Palmer (1996)[33], cuyas propuestas han ido apareciendo a lo largo de este capítulo.

Hughes (1989: 47) recomienda a los profesores o responsables de las pruebas que se planteen la siguiente pregunta: “what will be the cost of not achieving beneficial backwash?”, porque el riesgo de que un examen no aporte efectos beneficiosos supone su fracaso y en ningún caso merece la pena.

En definitiva, este repertorio de consejos concretos de Hughes y Bailey nos lleva de nuevo a las otras cualidades de las pruebas. Podemos decir que cuando un examen reúne unos mínimos de validez, fiabilidad, carácter práctico e interactivo y

[33] Ver las referencias de Bachman y Palmer (1996) al efecto rebote en los apartados 3.6 (sobre la autenticidad) y 3.8.4 (el impacto de las pruebas en los individuos) de este capítulo.

autenticidad, es muy probable que también produzca efectos beneficiosos en la enseñanza.

Concluimos nuestro recorrido por los rasgos o cualidades que toda prueba debe reunir para ser realmente útil, insistiendo en las palabras de Bachman y Palmer (1996: 40), que propician el equilibrio:

> In designing and developing a test, we try to achieve the optimum balance among the qualities of reliability, construct validity, authenticity, interactiveness, and impact for our particular testing situation. In addition, we must determine the resources required to achieve this balance, in relationship to the resources that are available.

Con ellas cerramos este capítulo que consideramos fundamental, pues dentro de la teoría de las pruebas, conocer estos rasgos nos ayuda a optimizar los recursos de que disponemos en el diseño de pruebas de idiomas. Este acercamiento nos permite reflexionar sobre las causas de que algunas pruebas fracasen o tengan éxito.

En el capítulo 6 nos ocuparemos de manera más específica y concreta del análisis de los rasgos del C-test. Ya hemos adelantado que en la literatura encontramos opiniones contradictorias; muchos trabajos destacan su validez como instrumento de medida (Klein-Braley 1985, 1997; Connelly 1997; Rashid 2002; Eckes y Grotjahn 2006), mientras que otros la ponen en entredicho (Jafarpur 1995). No obstante, la prueba ha demostrado de forma evidente su carácter interactivo y su factibilidad. Si su uso se enfoca correctamente es muy probable que produzca un efecto beneficioso, tanto al utilizarlo en la evaluación como en las clases de idiomas (como instrumento de aprendizaje).

Más adelante, en el capítulo 9 de la Perspectiva Empírica, analizaremos la validez y fiabilidad de la prueba. Su validez aparente se estudiará en el capítulo 12, a partir de los datos del cuestionario de opinión.

CAPÍTULO 4. LA EVALUACIÓN DEL VOCABULARIO

4.1. Introducción: el vocabulario en la enseñanza de Lenguas Extranjeras

Aunque en algunos momentos se infravaloró el papel del vocabulario tanto en la práctica docente como en la investigación lingüística, en los últimos años es evidente un cambio de orientación. Alderson y Bachman (en Read 2000) reconocen su riqueza e importancia dentro de la Lingüística Aplicada; McCarthy (1990) resalta la necesidad del vocabulario para que se lleve a cabo la comunicación; Schmitt y Meara (1997) señalan que la competencia lingüística es mucho más que el manejo de la gramática, y Richards (en Schmitt 2000) insiste en el papel del vocabulario en la adquisición de segundas lenguas y en la formación del profesorado de idiomas.

> After many years of neglect, the study of vocabulary in applied linguistics is now flourishing. (Alderson y Bachman en Read 2000:ix)

> No matter how well the student learns grammar, no matter how successfully the sounds of L2 are mastered, without words to express a wider range of meanings, communication in a L2 just cannot happen in any meaningful way. (McCarthy 1990: viii)

> In the last twenty or so years, there has been a growing realization that total language proficiency consists of much more than just grammatical competence. (Schmitt y Meara 1997:18)

> Lexical knowledge is central to communicative competence and to the acquisition of a second language. [...] Understanding of the nature and significance of vocabulary knowledge in a second language, therefore needs to play a much more central role in the knowledge base of language teachers. (Richards en Schmitt 2000: xi)

Los recientes trabajos de Meara (1996), Schmitt y McCarthy (1997), Laufer y Nation (1995, 1999), Singleton (1999), Schmitt (2000), Read (2000) y Nation (2001), entre otros, han contribuido al desarrollo y auge de este campo. Cada uno de ellos aborda la cuestión del vocabulario desde una perspectiva diferente. Haremos referencia a sus trabajos continuamente a lo largo de este capítulo.

El manejo del vocabulario es un prerrequisito para el uso efectivo de la lengua (Read 2000). Según Laufer y Hulstijn (2001: 1): "virtually all second language learners and their teachers are well aware of the fact that learning a second language (L2) involves the learning of large numbers of words". Los propios docentes son conscientes del papel del vocabulario en la enseñanza de la lengua extranjera y demandan una base teórica que informe su trabajo en el aula.

El C-test, objeto de nuestro estudio, ha sido considerado a efectos prácticos como prueba específica de vocabulario (Chapelle 1994; Read 2000; Schmitt 2000), a pesar de que no fue ésta la intención de sus creadores. El diseño, como variación de las pruebas de cierre, pretendía ser un método fiable y válido para medir la competencia general en la lengua (Klein-Braley 1985, 1997) y así ha sido reconocido en la literatura (Hughes 1989; Chapelle y Abraham 1990; Dörnyei y Katona 1992; Connelly 1997; Babaii y Ansari 2001; Wolter 2002; Eckes y Grotjahn 2006). Pero lo cierto es que la prueba ha sido también un instrumento importante en la investigación sobre adquisición del vocabulario en una segunda lengua (Singleton y Little 1991; Chapelle 1994). Su aportación, sus posibilidades, e incluso la controversia que le ha acompañado, nos animan a profundizar en su estudio.

Dedicamos un capítulo de esta tesis al rol del vocabulario en el aprendizaje de Lenguas Extranjeras para tener un marco de referencia previo que nos ayude a valorar y clasificar al C-test como prueba de evaluación. En él se abordan algunos de los conceptos que se utilizan después para el análisis de los textos del C-test en el capítulo 9 de la Perspectiva Empírica.

En primer lugar, revisaremos la naturaleza y características del vocabulario. Comenzaremos por acotar el concepto de palabra, indagaremos después en los factores que determinan su aprendizaje y en la clasificación de las palabras atendiendo a distintos criterios. Así, distinguiremos entre *types* y *tokens*, entre términos funcionales y léxicos, analizaremos después las unidades léxicas de más de una palabra, y veremos los tipos de términos según su frecuencia en la lengua.

Culminaremos esta primera parte del capítulo con las últimas definiciones del constructo del vocabulario.

En segundo lugar, abordaremos cuestiones relativas a su adquisición o aprendizaje. Apuntaremos las diferencias entre la adquisición de la lengua materna y la de una segunda lengua. Enumeraremos algunos rasgos de la adquisición del vocabulario, como su carácter gradual, el papel de la memoria y la dualidad entre incorporación implícita y sistemática de vocabulario. Además presentaremos estrategias para su aprendizaje. Con ello nos internaremos en el tema central de esta tesis: la evaluación.

Indagaremos en lo que implica la evaluación del aprendizaje de vocabulario. Haremos un seguimiento histórico de las investigaciones sobre vocabulario. Acabaremos centrándonos en el siglo XX y en las últimas tendencias de *Testing Vocabulary* (cómo y qué miden las pruebas de vocabulario, tipos, C-test). Los exámenes de vocabulario han tenido una doble vertiente; se han utilizado tanto con fines académicos como en la investigación. En cada caso el diseño depende del propósito de la prueba. Veremos también algunos ejemplos de pruebas estandarizadas de vocabulario.

Por último, analizaremos las pruebas de vocabulario más comunes en el contexto de la enseñanza de lenguas extranjeras. Las clasificamos en tres grandes grupos: pruebas de elementos discretos, holísticas y de cierre.

Entre las pruebas de elementos discretos citamos las de elección múltiple, las asociaciones, la traducción y las listas de reconocimiento de vocabulario. En las de tipo holístico hacemos referencia a la redacción. En cuanto a las de cierre, haremos una clasificación básica para situar al C-test o prueba C. A continuación, el capítulo 5 profundiza en las pruebas de cierre y el 6 se centra en el C-test.

La riqueza y amplitud del campo de la evaluación del vocabulario es obvia, no obstante, por razones de espacio, simplemente esbozamos los aspectos que resultan más pertinentes para el estudio que aborda esta tesis.

4.2. Naturaleza del vocabulario

Este apartado parte del concepto de palabra y sus tipos, atendiendo a distintos criterios de clasificación. Aborda también los grados y tipos de conocimiento de una palabra y cuestiones relativas a la adquisición o aprendizaje del vocabulario.

4.2.1. Concepto de palabra

Desde un punto de vista operativo, en esta tesis consideraremos la "palabra" como elemento unitario que puede expresar un concepto. Dejamos para estudios posteriores otras aproximaciones teóricas al concepto de palabra, conscientes de la dificultad que implica su definición, como señala Bogaards (2000: 491): "As is well-known, the concept of "word" has never been very clear in linguistic theory, although many different definitions have been given".

Carter (1987: 4) propone una definición de palabra basada exclusivamente en el aspecto externo: "a word is any sequence of letters (and a limited number of other characteristics such as hyphen and apostrophe) bound on either side by a space or punctuation mark".

El criterio ortográfico es claro e intuitivo, pero insuficiente. La literatura muestra que no es tan fácil acotar los límites del concepto de palabra atendiendo a otros criterios[34].

Partiendo de la definición de Carter (1987) llegamos a otras más amplias que incluyen a collocations y chunks of words como unidades léxicas formadas por grupos de palabras que transmiten un significado y que también forman parte del vocabulario de una lengua.

Schmitt (2000) nos enfrenta ya desde la introducción de su libro *Vocabulary in Language Teaching* con los problemas que plantea el concepto de palabra. Muestra a modo de ejemplo un grupo de sinónimos del verbo "*die*". Todos comparten

[34] Alcina y Blecua (1982: 201) definen el concepto de palabra como "la secuencia de sonidos formada por uno o más morfemas que puede ser aislada por conmutación". Bello (1984) insiste más en el aspecto léxico-cognitivo "cada palabra es un signo que representa por sí solo una idea o pensamiento".

aproximadamente el mismo significado, pero algunos están formados por una sola palabra y otros son unidades de más de una (*phrasal verbs e idioms,* tales como *pass away, bite the dust,* etc.).

Es evidente que no siempre una sola palabra se corresponde directamente con un significado. Las unidades significativas de más de una palabra reciben múltiples denominaciones en inglés: *lexemes*, *lexical units* o *lexical items* (Schmitt 2000: 2), *multi-word items,* etc.

Volviendo a la palabra como elemento unitario, veremos su estructura interna[35]. Las palabras están formadas por la raíz o lexema y los morfemas. El morfema es la unidad mínima significativa. La raíz aporta significado léxico y los morfemas información gramatical. En inglés los morfemas son el plural, la tercera persona del singular en presente simple, la forma *–ing*, el pasado simple y el participio pasado, el caso posesivo y las formas que expresan los grados de comparación del adjetivo (Bauer y Nation 1993).

Así pues, una misma palabra puede aparecer en la lengua de varias formas; los nombres en singular y/o plural, los verbos en los distintos tiempos, etc.

Para el estudio del vocabulario se considera que cada palabra, esto es, su raíz más los distintos morfemas (*inflections*) posibles, es un *lemma*. Generalmente el *lemma* es la unidad que sirve como base para el recuento de palabras de una lengua en los *corpora,* como vemos en el *Brown Corpus* de Francis y Kucera (1982).

Además, tenemos que contar con la derivación. Los distintos prefijos y sufijos que añadimos a la raíz nos permiten formar familias de palabras: “if the affixes change the word class of a stem, the result is a derivative” (Schmitt 2000: 2).

Los expertos se plantean hasta qué punto las palabras que forman una familia no constituyen también una unidad. Como veremos a continuación, generalmente la estimación de la amplitud del vocabulario de los hablantes, nativos o no, se hace tomando como unidad de referencia la familia de palabras.

Fenómenos como la polisemia, sinonimia y homonimia (homófonos y homógrafos) complican aún más el concepto de palabra porque impiden la correspondencia directa forma-concepto o significante-significado. El estudio detallado de los mismos escapa a los objetivos de nuestro trabajo.

[35] Alarcos (1994: 59) define: “La palabra suele ser una combinación de dos o más signos: uno, a cuyo significante llamamos raíz y cuyo significado hace una referencia léxica, y otro, que llamamos desinencia o terminación, que alude a los valores gramaticales o morfológicos de la palabra”.

4.2.1.1. Amplitud del vocabulario

Atendiendo a la cierta ambigüedad del concepto de palabra expuesta en el apartado anterior, surgen dudas al determinar la amplitud del vocabulario que conoce un hablante. De hecho, las cifras que aportan los distintos autores difieren mucho dependiendo de su aproximación al concepto de palabra; unos consideran como unidad a cada palabra individual, cada *lemma*, otros cada familia[36].

Interesa concretar estos términos para poder conocer a efectos prácticos cuántas palabras hay en un idioma, cuántas sabe un hablante nativo y cuántas aproximadamente debe conocer una persona que aprende la lengua para manejarse en ella, ya que "Learner´s vocabulary size has serious implications for every day oral and written communication and academic success" (Lee 2003: 551).

En el caso de la lengua inglesa tomamos la estimación de Nation (2001) según la cual el *Webster´s Third New International Dictionary* contiene unas 114.000 familias de palabras. Para llegar a ella Goulden, Nation y Read (1990) actuaron de la siguiente forma: "In the process, we deleted derived forms, proper names, compound nouns, abbreviations, affixes and various other non-base items" (Read 2000: 19).

Sabemos que ni siquiera los hablantes nativos de una lengua conocen todo su vocabulario. Los estudios más recientes y fiables coinciden en señalar que el hablante inglés nativo con una educación media universitaria conoce unas 20.000 familias de palabras, sin incluir los nombres propios: "Recent reliable studies (Goulden, Nation and Read, 1990); Zechmeister, Chronis. Cull, D'Ana and Healy, (1995) suggest that educated native speakers of English know around 20,000 word families" (Nation 2001: 9).

El vocabulario de una lengua se caracteriza por su dinamismo, está en constante cambio y crecimiento (Andrés Cortés 2004). La creación de palabras nuevas atiende a los cambios de la sociedad y de la ciencia. La existencia de conceptos nuevos impulsa la generación de términos nuevos.

Según los expertos, durante los primeros años de vida el hablante inglés nativo añade a su vocabulario aproximadamente 1000 familias de palabras al año.

[36] Goulden, Nation y Read (1990: 342) indican esta disparidad en las cifras que aportan las investigaciones de amplitud del vocabulario "The most notable feature of these investigations is the enormous divergence among the results. The various estimates [...] range from 3000 words to 216000 words".

Después el ritmo decrece, pero sigue incorporando palabras nuevas al vocabulario durante toda la vida.

4.2.2. Grado de conocimiento de una palabra

Aunque a primera vista podría pensarse que saber una palabra es algo obvio, el conocimiento de una palabra no es unidimensional, sino que tiene muchos grados y facetas. Saber una palabra es algo más complejo que conocer su significado. No nos podemos limitar a la imagen tradicional de aprender vocabulario exclusivamente memorizando una larga lista de palabras y su traducción. En primer lugar, porque no siempre hay una correspondencia unívoca entre palabra y concepto. Además, porque el contexto que las rodea influye en su significado.

Richards (1976), Bogaards (2000), Nation (2001), Laufer *et al*. (2004), entre otros, estudiaron las dimensiones del conocimiento de la palabra y concluyeron que saber una palabra implica conocer su significado/s, pero también su forma y cómo funciona en la lengua, es decir, su uso, pues las palabras no son unidades aisladas.

> Some researchers (Richards 1976; Ringbom 1987; Nation 1990; 2001) claim that knowing a word involves a range of interrelated sub-knowledges such as morphological and grammatical knowledge and knowledge of word meanings. (Laufer *et al*. 2004: 203)

Para describir lo que implica saber una palabra, Richards (1976) partió de la premisa de que el hablante nativo continúa incorporando vocabulario nuevo durante toda su vida, mientras que su competencia gramatical permanece estable.

En el acercamiento de Richards se basó Nation (1990), que en su primera clasificación distinguía hasta ocho aspectos para determinar el conocimiento ideal de un término[37]: significado, forma escrita y oral, comportamiento gramatical, *collocations*, registro, asociaciones y frecuencia de uso de la palabra. Y cada

[37] En esta tesis se utiliza indistintamente *palabra* y *término*. Entendemos que palabra es la unidad de la lengua general frente al término como unidad de las lenguas de especialidad. Pero, a efectos prácticos, no consideramos pertinente la distinción en nuestro trabajo. Quizá deberíamos simplemente hablar de *unidad léxica* como entidad más abstracta (Cabré y Adelstein 2001).

aspecto con la dualidad receptivo-productivo, ya que, como veremos, no es lo mismo entender una palabra que ser capaz de utilizarla.

Posteriormente Nation (2001) reestructuró los componentes del conocimiento de una palabra en torno a los tres aspectos que conlleva; esto es, su forma, su significado y su uso, en ambas facetas: receptiva y productiva.

Tabla 4.1. Componentes del conocimiento de una palabra (Nation 2001: 26)

Components of word knowledge (Nation 1990: 31)

Form		
Spoken form	R	What does the word sound like?
	P	How is the word pronounced?
Written form	R	What does the word look like?
	P	How is the word written and spelled?
Position:		
Grammatical patterns	R	In what patterns does the word occur?
	P	In what patterns must we use the word?
Collocations	R	What words or types of words can be expected before or after the word?
	P	What words or types of words must we use with this word?
Function:		
Frequency:	R	How common is the word?
	P	How often should the word be used?
Appropriateness	R	Where would we expect to meet this word?
	P	Where can this word be used?
Meaning:		
Concept	R	What does the word mean?
	P	What word should be used to express this meaning?
Associations	R	What other words does this word make us think of?
	P	What other words could we use instead of this one?

Key: R = receptive; P = productive

No obstante, Schmitt (1998), Meara (1996) y Read (2000) comparten la preocupación por la posibilidad de la aplicación práctica de este marco a las pruebas de evaluación del vocabulario. Veamos la reflexión de Meara (1996: 46): “It might be possible in theory to construct measures of each of these types of knowledge of

particular words; in practice, it would be very difficult to do this for more than a handful of items".

Por ello se inclinan más hacia el diseño de pruebas que midan la competencia global o la amplitud del vocabulario que maneja el alumno que hacia el grado de conocimiento de las palabras concretas o aisladas, en consonancia con las premisas del movimiento comunicativo, como veremos en apartados posteriores.

4.2.2.1. *Learning burden*

Entendemos *learning burden* como la dificultad que presentan las palabras para ser aprendidas. Es obvio que no todas las palabras requieren el mismo esfuerzo para su aprendizaje. Si por algún motivo el hablante está familiarizado con ellas, la dificultad disminuye (Nation 2001: 24).

> The general principle of learning burden (Nation, 1990) is that the more a word represents patterns and knowledge that learners are already familiar with, the lighter its learning burden.

Y los modelos pueden ser tanto de la lengua nativa como de cualquier otro tipo de conocimiento (*background*) que tenga el hablante. En general, cuando la lengua materna está relacionada con la lengua extranjera, el aprendizaje de ésta requiere menor esfuerzo.

El grado de dificultad de las palabras para ser aprendidas también tiene que ver con otros aspectos, como su categoría gramatical, su fonética, etc. Schmitt (2000: 148) agrupa los factores en intraléxicos y factores que dependen de la comparación entre distintas lenguas: "Factors can be related to the word itself (intralexical factors), or they can involve how well the learner's L1 matches the L2 (crosslinguistic factors)".

En el apartado 4.3.6 de este capítulo desglosamos algunos de ellos con mayor detalle. Veremos, por ejemplo, que algunos estudios (Ellis y Beaton 1993) muestran que es más fácil retener los nombres que otros tipos de palabra, probablemente por la mayor facilidad para formar imágenes mentales a partir de ellos. Otros autores

(Laufer 1997), sin embargo, cuestionan esta idea. En cuanto al aspecto fonético, las palabras difíciles de pronunciar tardan más en aprenderse (Ellis y Beaton 1993).

Schmitt (2000) y Nation (2001) resaltan la función del profesor como "facilitador" del aprendizaje. Aplicando esta idea a la enseñanza del vocabulario, podemos decir que es tarea del profesor aligerar el *learning burden* de las palabras y facilitar su aprendizaje buscando estrategias, por ejemplo, analogías dentro de la lengua objeto de estudio y resaltando también las posibles conexiones entre ambas lenguas L1 y L2. Siguiendo a Nation (2001: 24):

> Teachers should be able to estimate the learning burden of words for each of the aspects of what is involved in knowing a word, so that they can direct their teaching towards aspects that will need attention and towards aspects that will reveal underlying patterns so that later learning is easier.

No obstante, los alumnos presentan grandes diferencias individuales en cuanto a la destreza o capacidad para incorporar vocabulario en una lengua extranjera y el docente debe conocer las características peculiares de cada uno para ayudarles en el proceso de aprendizaje de vocabulario[38].

4.2.2.2 Conocimiento receptivo y productivo

Cuando somos capaces de reconocer y comprender una palabra tenemos un conocimiento receptivo de la misma. Si además podemos utilizarla, oralmente o por escrito, nuestro conocimiento pasa a ser productivo: "It is the difference that we are all familiar with between being able to recognise a word when you hear or see it and being able to use it in your own speech or writing" (Read 2000:26).

La visión tradicional consideraba que el vocabulario receptivo pasaba después a productivo. Esta interpretación secuencial ha sido cuestionada. Hatch y Brown (1995) y Melka (1997), consideran que conocimiento receptivo y productivo son los dos extremos de un continuo. Waring (1998) sugiere que pueden incluso solaparse.

[38] Cook (1996: 95-117) desglosa las diferencias individuales en distintos tipos de motivación, edad, rasgos de la personalidad, aptitud, capacidad para aplicar estrategias de aprendizaje, etc. Nation (2001) y Laufer y Hulstijn (2001: 1) también mencionan la motivación, que "promotes success and achievement in L2 learning".

Los estudios (Read 2000; Laufer *et al*. 2004) coinciden en señalar que la producción lingüística es más difícil que la recepción o reconocimiento; pues supone un paso más en el conocimiento de una palabra: "A learner's passive vocabulary is always larger than his or her active vocabulary. This indicates that many words are first acquired passively, and that active knowledge is a more advanced type of knowledge" (Laufer *et al*. 2004: 208).

Esto se debe a diversos factores; en primer lugar, porque requiere un aprendizaje extra más preciso de la forma (oral o escrita) de la palabra. Además las actividades de tipo receptivo (*reading, listening*) generalmente se practican más que las productivas (*writing, speaking*) porque el aprendizaje receptivo se considera base suficiente para el productivo. Sin embargo, los estudios experimentales muestran claramente que es necesario un aprendizaje específico de las destrezas productivas para poder después utilizarlas.

A continuación reproducimos el cuadro de Laufer *et al*. (2004) que clasifica los grados de conocimiento léxico teniendo en cuenta las tareas que es capaz de realizar el alumno. Se basa en la dicotomía forma-significado y reconocimiento-producción.

Tabla 4.2. Types of vocabulary knowledge (Laufer *et al.* 2004: 206)

	Recall	Recognition
Active (Productive) (retrieval of form)	Active recall	Active recognition
Passive (Receptive) (retrieval of meaning)	Passive recall	Passive recognition

Autores como Meara (1990), Laufer (1998) y Laufer *et al*. (2004) utilizan los términos *passive* y *active* en lugar de *receptive* y *productive.* En la literatura se usan ambas nomenclaturas indistintamente, no obstante la denominación *active/passive* ha sido criticada. El argumento es que cualquier destreza lingüística, incluso las de tipo receptivo, implica actividad por parte del sujeto. En español suelen aparecer los términos vocabulario *activo* y *pasivo* como sinónimos de *receptivo* y *productivo*.

La distinción entre estos dos tipos de conocimiento del vocabulario tiene especial importancia en el diseño de pruebas de evaluación. Algunas pruebas pretenden medir el vocabulario receptivo y otras el productivo. Como hemos visto, varía el grado de profundidad del conocimiento de la palabra. El C-test no se limita al reconocimiento pasivo del vocabulario, como ocurre con las pruebas de elección múltiple, sino que exige la producción activa por parte del sujeto.

4.2.2.3. *Collocations*

El conocimiento de una palabra también supone saber junto a qué otras palabras suele aparecer en la lengua. Para expresarnos en un idioma memorizamos secuencias de palabras de distinto tipo y longitud. Y así, facilitamos el aprendizaje:

> Each [collocation]... must or should be learnt, or is best or most conveniently learnt as an integral whole or independent entity, rather than by the process of piecing together their component parts. (Palmer 1933: 4 en Nation 2001: 317)

El caso extremo de *fixed collocation* son los *multi-word items*. El apartado 4.2.3.3 está dedicado a las unidades léxicas de más de una palabra e incluye una clasificación de las mismas.

4.2.3. Tipos de palabras

Podemos clasificar las palabras atendiendo a diversos criterios. Pero, en esta tesis nos ceñiremos a las clasificaciones que son de especial interés para nuestro trabajo con el C-test.

Puesto que nos informa de la variación léxica de los textos, hacemos una primera distinción entre types y tokens. En segundo lugar, atendiendo al tipo de información que aportan las palabras, vemos la diferencia entre términos funcionales y léxicos. La proporción de los mismos en un texto resulta fundamental para conocer su densidad léxica. Por otro lado, revisamos los tipos de unidades significativas formadas por más de una palabra, tan comunes en la lengua inglesa. Y para

terminar, aportamos la clasificación de las palabras según su frecuencia en la lengua de Nation (2001), de ella surgen algunas implicaciones metodológicas para la enseñanza del vocabulario.

4.2.3.1. *Types* y *tokens*

Cuando nos enfrentamos a un texto y contamos las palabras que lo forman debemos que tener en cuenta la distinción entre *types* y *tokens*.

Puesto que no existe un equivalente en castellano, preferimos mantener los términos ingleses, que explicamos a continuación. *Token* es cada una de las palabras individuales que forman el texto o el discurso, es decir, en un texto el número de *tokens* equivale al número total de palabras. Aunque una palabra aparezca repetida varias veces en el texto, se cuentan todas y cada una de ellas. También se les llama *running words*. *Types,* sin embargo, son las palabras diferentes de que consta el texto.

La proporción entre *types* y *tokens* nos indica la variación léxica. Read (2000: 18) indica la utilidad de esta medida: “The relative proportions of types and tokens (known as the type-token ratio) is a widely used measure of the language development of both language learners and native speakers”.

Conocer la variación léxica de un texto escrito es parte importante de su análisis. En el capítulo 9 se manejan estos conceptos en el análisis de los textos que forman el C-test aplicado: número total de palabras, densidad y variación léxicas, etc., para comprobar cómo afectan las características textuales a la prueba.

4.2.3.2. Términos léxicos y funcionales

Una segunda distinción clasifica las palabras atendiendo al tipo de información que aportan.

Las que aportan información de tipo gramatical se denominan términos funcionales, *function words:* preposiciones, artículos, pronombres, verbos auxiliares,

conjunciones, etc. Read (2000: 18) considera que este tipo de palabras pertenecen más a la gramática que al vocabulario de la lengua. Suponen un número muy limitado de ítems en cada lengua. Sin embargo, algunas son palabras de alta frecuencia, que se repiten mucho en el discurso (véase *types* y *tokens* en el apartado anterior y *palabras muy frecuentes* en el 4.2.3.4).

Las palabras que aportan un contenido léxico, o palabras llenas, se denominan en inglés *content words* o *lexical items.* Son los nombres, adjetivos, adverbios y los verbos con contenido léxico[39].

La proporción entre palabras funcionales y de contenido léxico muestra la densidad léxica del texto. Como hemos mencionado en el apartado anterior, tanto la variación léxica como la densidad son aspectos importantes en el análisis de los textos escritos y, en el caso del C-test, se comprobará que inciden directamente en el grado de dificultad de la prueba (véase el capítulo 9).

Siguiendo un criterio operativo, la mayor parte de los estudios sobre el vocabulario de una lengua se centran en los términos con contenido léxico. También son el objetivo de las pruebas de vocabulario, como indica Read (2000: 18): "Generally speaking, when we set out to test vocabulary, it is knowledge of content words what we focus on".

Sin embargo, por su diseño, el C-test mide indistintamente la recuperación de ambos tipos de palabras: funcionales y con contenido léxico. De hecho, la hipótesis 3 de nuestro trabajo plantea qué tipo de términos se recuperan mejor y cómo afecta esto a los resultados obtenidos en la prueba.

En inglés, ciertos términos funcionales, tales como *a, to, the, and, in, that*, etc. son muy breves y tienen una gran frecuencia de aparición. Por tanto, cualquier tarea cuya base sea un texto con mayor carga de este tipo de términos ha de presentar también menor grado de dificultad para el alumno. Así pues, *a priori* parece razonable pensar que su recuperación en el C-test sea más fácil que la de los términos léxicos. Sin embargo, más adelante veremos que, además del tipo de término, léxico o funcional, el tamaño y frecuencia en la lengua inciden en la recuperación de los términos omitidos en el C-test.

[39] Alarcos (1994) clasifica las palabras en autónomas o independientes y dependientes. Las autónomas pueden cumplir por sí solas una función y coinciden con las cuatro clases de palabras con contenido léxico; esto es, verbos, sustantivos, adjetivos y adverbios. Las dependientes son las que sirven para marcar las relaciones entre ellas (preposiciones, conjunciones, etc.).

4.2.3.3. Unidades léxicas de más de una palabra

En el apartado 4.1 hemos señalado que el vocabulario de una lengua también incluye unidades léxicas formadas por más de una palabra. Moon (1997) define:

> A multi-word item is a vocabulary item which consists of a sequence of two or more words (a word being simply an orthographic unit). [...] Multi-word items are the result of lexical (and semantic) processes of fossilisation and word-formation, rather than the results of the operation of grammatical rules. (En Schmitt y McCarthy 1997: 43)

En lengua inglesa no existe una única terminología comúnmente aceptada para designar a estas unidades; que incluyen desde los *idioms* y *phrasal verbs* hasta las palabras compuestas, o expresiones fijas lexicalizadas. Nos referiremos a ellas como "unidades léxicas de más de una palabra", o bien manteniendo la expresión inglesa *multi-word items*.

Para identificar estas unidades en la lengua Read (2000) sugiere dos posibilidades; la tradicional que consiste en confiar en la intuición de los hablantes nativos, o bien la más actual que permite utilizar el *software* existente para buscarlas en los distintos corpus computerizados.

Los *multi-word items* se caracterizan por ser grupos relativamente fijos (son unidades significativas, pero su significado no equivale a la suma de los significados de las palabras aisladas que lo forman) y por resultarnos familiares, ya que algunas de ellas se utilizan mucho en la comunicación diaria.

Aunque también tienen una función pragmática, se reconocen como unidades de significado, por eso se consideran parte del vocabulario del inglés. Pero por sus características plantean dificultades al alumno a la hora de aprenderlas y al profesor al buscar los medios para evaluarlas.

Moon (1997: 57) reconoce esta dificultad: "Their non-compositionality, whether syntactic, semantic or pragmatic in nature, means that they must be recognised, learned, decoded and encoded as holistic units". Y cita la recomendación de Baker y McCarthy (1988: 32) para los profesores: "The more naturally MWUs are integrated into the syllabus, the less "problematic" they are".

Es evidente que son muy numerosas en inglés y que en muchos casos no están claramente delimitadas. Lo cierto es que los hablantes nativos las utilizan con

gran frecuencia y los que aprenden Inglés como Segunda Lengua o Lengua Extranjera las incorporan a su vocabulario productivo/activo en mayor cantidad a medida que aumenta su competencia y fluidez en la lengua. No obstante, los estudios de vocabulario tradicionalmente se han fijado más en las palabras como entidades individuales, y las pruebas de evaluación, en elementos discretos.

Hay diferentes marcos para clasificar las *multiword-items.* Con la debida cautela, puesto que algunas expresiones o categorías pueden solaparse, mostramos la clasificación que propone Moon (1997: 43-48), muy completa e ilustrativa:

1. *Compounds*: Son los compuestos por más de una palabra: *car park, dark-haired,* etc. Schmitt (2000: 99) define: "Compounds are created when two or more words are combined to make a single lexeme. This lexeme can be written as multiple orthographic words, hyphenated words, or as a single orthographic word".
2. *Phrasal verbs:* Combinaciones de verbos y adverbios o preposiciones, típicas de la lengua inglesa.
3. *Idioms:* Forman el grupo más complejo, "they have holistic meanings which cannot be retrieved from the individual meaning of the component words. [...] Idioms are typically metaphorical in historical or etymological terms" (Moon 1997).
4. *Fixed phrases:* Son otros grupos de palabras fijos, institucionalizados y frecuentes en la lengua, tales como *of course*, *excuse me*, *how do you do?.* También incluye los refranes.
5. *Prefabs:* "Prefabs are preconstructed phrases, phraseological chunks, stereotyped collocations, or semi-fixed strings which are tied to discoursal situations and which form structuring devices."

Los *prefabs* de Moon toman el nombre de *lexical phrases* en el marco propuesto por Nattinguer y DeCarrico (1992), y se subdividen en:

1 *Polywords*: grupos cortos de palabras, fijos y con una función concreta, como *for the most part*, *so to speak*, *at any rate*, etc.

2 *Institutionalised expressions*: grupos más largos como los refranes, proverbios, fórmulas sociales: *How do you do?*, *Once upon a time*, etc.
3 *Phrasal constraints:* frases que tienen una estructura básica con huecos en que se pueden insertar elementos distintos: *a [day/year/week] ago*, etc.
4 *Sentence builders*: que sirven como marco para toda una oración, como por ejemplo *I think that [...]*, *not only [...] but also [...]*, etc.

Las pruebas holísticas, esto es integradoras, y contextualizadas son las que mejor pueden medir el manejo de estas unidades significativas formadas por varias palabras (Read 2000: 21ss.).

El C-test es una prueba que participa de características de las pruebas de elementos discretos pero también presenta rasgos propios de las holísticas. No obstante, por las características de su diseño, no mide específicamente el conocimiento de las unidades léxicas formadas por más de una palabra, sino la recuperación de las palabras como entidades unitarias. Ahora bien, puesto que se trata de una prueba contextualizada, podemos decir que sí lo hace de manera implícita. Conocer este tipo de unidades facilita la labor de inferencia del alumno.

4.2.3.4. Tipos de términos según su frecuencia en la lengua

La frecuencia de uso de una palabra condiciona el almacenamiento y procesamiento léxico (Graña López 1997). Bybee (1995: 232) introduce la noción de fuerza léxica de una palabra, que viene dada por la frecuencia de su procesamiento. Cada vez que se procesa una palabra se refuerza la representación mental existente y, por tanto, su aprendizaje.

Teniendo en mente la enseñanza y aprendizaje de vocabulario, Nation (2001: 9-21) muestra un ejemplo práctico de análisis del vocabulario de un texto académico en lengua inglesa. Este análisis sirve como punto de partida para distinguir cuatro tipos de vocabulario atendiendo a su frecuencia de aparición y uso en la lengua. Y las implicaciones pedagógicas son claras.

4.2.3.4.1. Términos muy frecuentes

Las palabras de alta frecuencia de uso en la lengua incluyen tanto términos léxicos como funcionales. Los términos funcionales son un número limitado y conocido (*a*, *the*, *in*, *for*, *of*, etc.). Pero el análisis de textos revela que también determinados términos léxicos aparecen repetidos con frecuencia.

Michael West (1953) cifró el número de palabras muy frecuentes en 2000 familias de palabras. De ellas, unas 165 son funcionales y el resto léxicos o *content words*. Aunque hay otras listas de frecuencia más recientes, los datos que aportan son muy semejantes (Nation and Hwang 1995). Nation (2001) concluye que casi el 80% de las palabras de una lengua son palabras muy frecuentes.

A continuación, la tabla de Nation (2001: 17) muestra la proporción de las 2000 palabras más frecuentes en inglés en distintos tipos de texto.

Tabla 4.3. Proporción de las palabras más frecuentes en inglés en distintos tipos de texto

Text type and text coverage by the most frequent 2000 words of English and an academic word list in four different kinds of texts

Levels	Conversation	Fiction	Newspapers	Academic text
1st 1000	84.3%	82.3%	75.6%	73.5%
2nd 1000	6%	5.1%	4.7%	4.6%
Academic	1.9%	1.7%	3.9%	8.5%
Other	7.8%	10.9%	15.7%	13.3%

Esto implica que el aprendizaje de una lengua no requiere memorizar largas listas de palabras, al menos en los estadios iniciales. Por otra parte, facilitar y asegurar el aprendizaje de las palabras más frecuentes merece todo el esfuerzo por parte del profesor:

> The words are a small enough group to enable most of them to get attention over the span of a long-term English programme. This attention can be in the form of direct teaching, direct learning, incidental learning, and planned meetings with the words. The time spent on them is well justified by their frequency, coverage and range. […] In general, high-frequency words are so important that anything that teachers and learners can do to make sure that they are learned is worth doing. (Nation 2001: 16)

Con esta idea, Nation (2001) aporta ideas para la enseñanza de este tipo de palabras tan rentables en la lengua:

Tabla 4.4. Métodos de enseñanza de las palabras muy frecuentes (Nation 2001: 16)

Ways of learning and teaching high-frequency words

Direct teaching	*Teacher explanation*
	Peer teaching
Direct learning	*Study from word cards*
	Dictionary use
Incidental learning	*Guessing from context in extensive reading*
	Use in communication activities
Planned encounters	*Graded reading*
	Vocabulary exercises

Igualmente, es importante que el profesor disponga de instrumentos de evaluación prácticos, válidos y fiables que le informen del progreso de los alumnos en el aprendizaje del vocabulario muy o poco frecuente. Para ello se han diseñado distintas pruebas, como *The Vocabulary Levels Test* (Laufer and Nation 1995, 1999).

4.2.3.4.2. Términos académicos

Constituyen el vocabulario formal o especializado que se usa habitualmente; términos semi-técnicos o divulgativos. Puede suponer casi un 10 % en los textos académicos.

Para las personas que se enfrentan al uso del inglés como segunda lengua o lengua extranjera en contextos académicos, existe una lista: *Academic Word List* (Coxhead 1998). Enumera 570 familias de palabras frecuentes en el mundo académico sin estar restringidos exclusivamente a un campo de estudio.

4.2.3.4.3. Términos técnicos

El vocabulario técnico es el específico de un tema o ciencia. Depende directamente del tema del texto. Es aproximadamente un 5% del total de las palabras de un texto académico. La mayor parte de este tipo de vocabulario sólo tiene sentido en el contexto científico en que se estudia. A menudo se utilizan palabras frecuentes en otros contextos, pero que adquieren un nuevo significado en ese campo específico.

4.2.3.4.4. Términos poco frecuentes

El 5% restante correspondería a nombres propios, palabras raras, y a palabras que también se usan con frecuencia, pero que no entran en las listas de las consideradas más frecuentes.

El límite entre las palabras muy frecuentes y las de baja frecuencia es arbitrario. En contextos muy concretos nombres propios o palabras raras se pueden convertir en palabras frecuentes.

En cuanto a la enseñanza de estas palabras poco frecuentes, el profesor debe centrarse más en el desarrollo de estrategias para su manejo (*guessing from context clues*, *using word parts to help remember words*, *using vocabulary cards and dictionaries*) que en las palabras concretas. La incorporación de palabras nuevas a su vocabulario es tarea del alumno.

En el C-test las omisiones incluyen todo tipo de términos en cuanto a su frecuencia. Sólo los nombres propios, las cifras y las palabras de una sola letra quedan intactos. En los dos primeros casos la recuperación sería imposible sin un conocimiento previo del texto, y en el tercero, evidente. Con la intención de evitar en los C-tests los términos excesivamente fáciles por su alta frecuencia de aparición (casi siempre coincidiendo con palabras funcionales) y los demasiado difíciles (como los técnicos o muy poco frecuentes), Jafarpur (1999) propuso la creación de C-tests a la medida. Este intento, que comentaremos con mayor detalle en el capítulo 6, no consiguió mejorar la prueba.

4.2.4. Últimas definiciones del constructo del vocabulario

Actualmente, el constructo del vocabulario se puede entender como elemento discreto o bien integrado en la competencia lingüística general.

Bachman y Palmer (1996) consideran que el vocabulario es parte de la competencia lingüística (*embedded*). Sin embargo, Chapelle (1994) lo define como un constructo discreto. Nos detenemos en el modelo de competencia léxica de Chapelle por ser un marco centrado precisamente en el vocabulario que ofrece claves importantes para el para el diseño de pruebas. Read (2000: 35) valora su aportación de este modo: "While not seeking to isolate it from other language abilities, Chapelle has highlighted the broad role that vocabulary plays in language competence and performance".

La definición de competencia léxica -*vocabulary ability* de Chapelle se basa en el marco de competencia lingüística general propuesto por Bachman (1990) e incluye "both knowledge of language and the ability to put language to use in context" (Chapelle 1994: 163).

Para Chapelle la competencia léxica está formada por tres componentes:

1. El conocimiento del vocabulario y sus procesos,
2. el contexto,
3. y las estrategias cognitivas para su uso.

La autora distingue cuatro dimensiones del conocimiento del vocabulario:

- La amplitud del vocabulario,
- el conocimiento de las características de cada palabra,
- la organización del léxico,
- los procesos fundamentales del vocabulario.

El segundo componente es el contexto. No se refiere solo a la oración en la que aparece la palabra cuyo conocimiento queremos medir, ni siquiera al texto en que se inserta, como en los *clozes.* Incluye también lo que Bachman llama *pragmatic knowledge;* la situación cultural o social influye en el significado de las palabras. Dentro del contexto Read (2000) cita además las diferencias entre situaciones

coloquiales y formales (registros), entre las distintas generaciones, las variedades de una lengua, etc.

Chapelle (1994: 64) propone tres elementos para analizar la situación social en que se utiliza la lengua: *field, tenor* y *mode*. *Field* se refiere al tipo de actividad que desarrolla el hablante, *tenor* al estatus de los participantes en la comunicación y a la relación interpersonal que tienen, y *mode* al canal (oral, escrito).

Las estrategias metacognitivas para el uso del vocabulario constituyen el tercer componente que señala el modelo de Chapelle. Bachman (1990) lo denomina competencia estratégica. Son las estrategias que todo hablante maneja, de forma más o menos consciente, para utilizar el vocabulario en la comunicación. Cuando se aprende una lengua extranjera estas estrategias se hacen aún más necesarias. En la producción se usan estrategias de acomodación, como la simplificación léxica y el *avoidance*; esto es, evitar utilizar los términos que nos plantean problemas porque no los conocemos bien, etc. En las destrezas receptivas, cuando encontramos palabras desconocidas, se aplican otras estrategias como la búsqueda en el diccionario, la consulta a otra persona, seguir la lectura a pesar de no entender alguna palabra, deducir a partir del contexto, etc.

La experiencia nos muestra que, ante un término desconocido, aplicamos cualquiera de las estrategias mencionadas, la elección depende de las características personales, la situación, etc. No tiene porqué ser una sola, con frecuencia se entremezclan.

Queremos resaltar la importancia del manejo de una de ellas; la utilización de las claves contextuales para deducir aquellos ítems de vocabulario que desconocemos. Es una estrategia muy rentable, sobre todo para la realización de pruebas de vocabulario, especialmente en los *clozes* y el C-test. Tanto que Nation recomienda que se enseñe como tal en las clases.

4.3. Adquisición y aprendizaje de vocabulario

Antes de desarrollar este apartado haremos una aclaración previa sobre la terminología utilizada. Tradicionalmente se consideraba que la adquisición era un

proceso subconsciente relacionado con la lengua nativa, frente a la conciencia que requiere el aprendizaje de una segunda lengua. Aunque algunas voces piden que se mantenga tal distinción, al menos en el plano teórico (Thatcher 2000), la tendencia actual más respaldada es considerar que ambos procesos requieren conciencia en mayor o menor grado y, por tanto, la distinción no es pertinente (Ellis 1985; Laufer 1997).

En nuestro trabajo utilizamos indistintamente los términos *adquisición* y *aprendizaje* para referirnos a los procesos conscientes o inconscientes de interiorización de los conocimientos lingüísticos, salvo cuando sea necesario precisar. En la literatura, como en el título de este apartado, a menudo forman parte del mismo enunciado (Ellis 1985).

A pesar de que se han propuesto modelos para describir el proceso de adquisición del vocabulario, no existe aún una teoría global que permita entenderlo completamente. En este campo, la Lingüística necesita las aportaciones de otras ciencias, como la Psicología cognitiva y la Neurología. También sería enriquecedora la colaboración entre el campo de la adquisición de segundas lenguas y el de la evaluación (Shohamy 2000).

Es evidente que los seres humanos somos capaces de incorporar a nuestro vocabulario miles de palabras. A simple vista sorprende la cantidad de palabras que maneja un hablante nativo. Y aún más en el caso de las personas que aprenden una lengua extranjera. En ambos casos la adquisición de vocabulario se lleva a cabo principalmente de dos formas complementarias entre sí: mediante la incorporación sistemática de palabras nuevas (*explicit learning*) o con el aprendizaje implícito (*incidental learning),* a medida que nos las encontramos (Read 2000)[40].

No obstante, la adquisición de la lengua materna y de una lengua extranjera difieren mucho. Nation (1995) es consciente de que todavía quedan muchas preguntas sin respuesta y un amplio campo de trabajo para la investigación sobre la adquisición de vocabulario en L1 y L2. Schmitt (2000: 116) refleja la amplitud y dificultades con que se enfrentan los investigadores a la hora de formular una teoría

[40] Una mayor profundización en estos aspectos del aprendizaje de la lengua nos llevaría a la distinción entre las dicotomías *implicit/explicit* e *incidental/intencional*, es decir, entre aprendizaje implícito e incidental y explícito e intencionado. Para ello recomendamos la lectura de DeKeyser (2003) y Hulstijn (2003) en *The Handbook of Second Language Acquisition*. Tal precisión supera las pretensiones de este capítulo, que tan sólo intenta reflexionar acerca de algunas cuestiones relativas al aprendizaje del vocabulario y presentar el panorama actual.

de la adquisición de una segunda lengua: “In fact, there are so many variables that affect second language vocabulary acquisition, such as L1, age, amount of exposure, motivation, and culture, that it is very difficult to formulate a theory of acquisition that can account for them all”.

A continuación esbozamos algunas diferencias que señalan los especialistas en Lingüística Aplicada entre la adquisición de L1 y L2. Un estudio más profundo del tema, si bien resultaría interesante, escapa a los objetivos de esta tesis.

Vemos también cómo se produce la incorporación de vocabulario y el rol que desempeña la memoria en este proceso. Finalizamos con la descripción de los factores que determinan su aprendizaje y algunas estrategias para facilitarlo.

4.3.1. Diferencias entre la adquisición de L1 y L2

Se ha investigado mucho sobre la adquisición del lenguaje y su procesamiento. Han surgido teorías que intentan explicarlo, a pesar de todo aún no se conocen bien estos procesos[41]. Muchos de ellos se pueden aplicar también a la adquisición de una segunda lengua o lengua extranjera. Otros son claramente diferentes, porque factores como la edad o la madurez cognitiva también lo son (Ellis 2000: 107; Schmitt 2000: 18-19; Thatcher 2000).

En el caso de la lengua materna o L1, Ellis (2000) habla de una “disposición innata a adquirir la lengua de forma automática e inconsciente”. La exposición a la lengua comienza incluso antes del nacimiento. Después hay un periodo de tiempo en que el hablante recibe *input* constantemente pero no es todavía capaz de producir *output*. Cuando comienza a hablar parte de su discurso está formado por grupos comunes de palabras que memoriza y emite (*preformulated speech*). El niño adquiere la lengua nativa principalmente de forma incidental, gracias a ejemplos más que a través de reglas explícitas[42]. Es capaz de manejar un sistema complejo, pero no de describirlo.

[41] Desde la idea de la Gramática Universal innata de Chomsky (1986) hasta la teoría de Krashen *The Input Hypothesis* (1982).

[42] Chomsky habla de “*language growth*” en lugar de “*language acquisition*”. Roberts (2000), en Ellis (2000: 455-475) sigue la visión Chomskiana cuando considera que la adquisición de L1 es el prototipo de aprendizaje incidental o implícito, por ser un mecanismo innato en el que sobra la instrucción formal.

Un niño que aprende su lengua nativa lo hace a la vez que conceptualiza el mundo que le rodea. Thatcher (2000) indica que la adquisición de una primera lengua por parte de un niño forma parte de su desarrollo cognitivo. Según Aitchison (1987) en Schmitt (2000) el niño adquiere significados en su lengua nativa en un proceso de tres estadios: *labelling*, *categorization* y *network building*.

El aprendiz de una lengua extranjera ya tiene la experiencia de una primera lengua, conoce los conceptos y por tanto más bien realiza un proceso que los expertos denominan *relabelling*.

Teniendo en cuenta cómo se adquiere la lengua nativa han surgido distintos métodos para la adquisición de lenguas segundas o extranjeras, algunos intentan seguir el modelo de la adquisición de L1 y potenciar el aprendizaje implícito, evitando en lo posible la instrucción explícita (Krashen 1982), otros estudios la recomiendan (Long 1983). Las investigaciones más recientes sobre enseñanza de lenguas segundas o extranjeras indican la conveniencia de utilizar métodos que propicien tanto el aprendizaje implícito como el formal o explícito (Ellis 2000).

4.3.2. Carácter gradual de la adquisición de vocabulario

Schmitt (2000: 117) incide en que la adquisición de vocabulario se produce de forma gradual y progresiva: “vocabulary acquisition is incremental in nature”.

Saber una palabra es un proceso complicado. Supone conocer muchos aspectos (citados en el apartado 4.2.2) y no todos se adquieren simultáneamente. El significado básico de una palabra se adquiere al principio. También una primera aproximación a su forma, dependiendo de si la exposición a la palabra es oral o escrita. Según Schmitt (2000) quizá se llegue a percibir incluso la categoría gramatical de la palabra en el primer acercamiento.

Estos rasgos se van fijando a medida que aumenta la exposición a la palabra y además irán apareciendo otras acepciones significativas: “Vocabulary is learnt incrementally and this obviously means that lexical acquisition requires multiple exposures to a word” (op. cit.: 137).

Henricksen (1999) considera que la adquisición de los aspectos léxicos se hace siguiendo un continuo que va desde el conocimiento 0 hasta el manejo preciso del

término. El conocimiento de una palabra supone el manejo de distintos aspectos léxicos. Éstos pueden ser receptivos o productivos.

Más tarde se desarrolla la intuición acerca de algunos otros rasgos como la frecuencia, el registro, y la *collocation* de la palabra.

En esta línea se expresan Laufer *et al*. (2004: 203) al recoger las teorías que abogan por el aprendizaje gradual de las palabras: "Others assume that lexical knowledge consists of progressive levels of knowledge, starting with a superficial familiarity with the word and ending with the ability to use the word correctly in free production (Faerch *et al*. 1984; Palmberg 1987)".

4.3.3. La memoria en la adquisición de vocabulario

Las últimas investigaciones reivindican el papel de la memoria en el aprendizaje del vocabulario. Schmitt (2000: 129) sentencia: "Memory has a key interface with language learning". Graña López (1997: 28) lo expresa del siguiente modo:

> Cualquier consideración sobre el procesamiento léxico ha de partir del presupuesto difícilmente rebatible de que las palabras, al contrario que la mayoría de los sintagmas y oraciones, pertenecen al banco de datos de la memoria, y ello determina que haya dos aspectos que resulta necesario investigar: primero, cómo están organizadas o almacenadas las palabras en ese banco de datos, y segundo, cómo se usan, o de manera más precisa, cómo se recuperan, en las tareas de comprensión y producción del habla.

El objetivo es conseguir que la información léxica pase de la memoria a corto plazo a la memoria a largo plazo. Para ello es necesario establecer relaciones entre los nuevos ítems y los que ya han sido aprendidos previamente. Se pueden utilizar diversas técnicas y estrategias que faciliten el paso a la memoria a largo plazo y en definitiva, el aprendizaje.

En el proceso de aprendizaje del vocabulario se producen avances pero también retrocesos debidos al olvido. Schmitt (1998) demostró que es más fácil olvidar los términos que se conocen sólo de forma receptiva y no productiva.

El olvido de lo aprendido se denomina *attrition.* Sigue una curva típica que indica que ocurre, sobre todo, en el periodo de tiempo más cercano al aprendizaje, después se estabiliza. Por supuesto, no es exclusivo del aprendizaje de una lengua,

ni mucho menos del vocabulario. Pero sí es un dato que el profesor ha de tener en cuenta para proporcionar al alumno oportunidades de repasar lo aprendido en momentos relativamente próximos al primer aprendizaje (*recycling*).

4.3.4. Incorporación sistemática de vocabulario

Comenzamos con las definiciones de Ellis (2000) y Schmitt (2000). Ambos autores explican las diferencias entre aprendizaje implícito y explícito[43]. Podemos aplicarlas a las formas de incorporación de vocabulario: sistemática e incidental:

> Implicit learning is acquisition of knowledge about the underlying structure of a complex stimulus environment by a process which takes place naturally, simply and without conscious operations. Explicit learning is a more conscious operation where the individual makes and tests hypotheses in a search for structure. (Ellis 2000: 1)

> Explicit learning focuses attention directly on the information to be learned, which gives the greatest chance for its acquisition. [...] Incidental learning can occur when one is using language for communicative purposes, and so gives a double benefit for time expended. (Schmitt 2000: 120)

Hatch y Brown (1995: 368) insisten en la intencionalidad como rasgo peculiar del aprendizaje sistemático "...as being designed, planned for, or intended by teacher or student".

Como veremos, la simple exposición a la lengua (actividades como la lectura, ver películas, la interacción en el aula, etc.) propicia la incorporación incidental o implícita de vocabulario. Sin embargo, el aprendizaje del vocabulario en una lengua extranjera también requiere que el profesor dirija la atención del alumno de forma explícita y sistemática hacia determinadas palabras, generalmente las de uso más frecuente en la lengua.

Según Schmitt (2000: 121), las técnicas o métodos para hacerlo van desde la tradicional memorización de listas hasta otras actividades que ayuden a retener

[43] El volumen *Implicit and Explicit Learning of Languages* editado por Ellis (2000) es un buen compendio de las últimas investigaciones al respecto. También el ya mencionado *The Handbook of Second Language Acquisition* editado por Doughty y Long (2003) incluye aportaciones interesantes sobre este tema.

mejor la información: “the more one engages with a word (deeper processing) the more likely the word will be remembered for later use”.

El método tradicional de enseñanza de vocabulario implicaba la memorización sistemática de largas listas de palabras y su significado. Con la llegada del enfoque comunicativo estas prácticas quedaron bastante relegadas. En los años 90 no se consideraba recomendable la memorización de palabras aisladas, sin un contexto.

Hoy, sin embargo, los expertos no parecen respaldar totalmente esta idea. En el apartado anterior señalamos que de nuevo se valora el papel de la memoria en el aprendizaje, aunque la enseñanza no se limite a ella, sino que la recomiende siempre junto a estrategias de otro tipo. Lee (2003) propone la realización de tareas de escritura inmediatamente después la instrucción explícita de vocabulario para ayudar a la retención de las palabras nuevas. Prácticas de este tipo facilitan el paso del vocabulario receptivo a productivo.

Los últimos estudios (Lawson y Hogben 1996; Schmitt 1997) coinciden en señalar que los buenos alumnos que aprenden vocabulario en una lengua extranjera utilizan múltiples estrategias distintas para hacerlo (*bilingual dictionaries, written and oral repetition, studying the spelling, taking notes in class*, etc.). Las revisamos en el apartado 4.3.8.

4.3.5. Incorporación incidental de vocabulario

Hemos visto que el aprendizaje del vocabulario de una lengua se lleva a cabo de dos formas complementarias: mediante la incorporación sistemática de palabras sobre las que se incide de forma explícita y la incorporación incidental de otras a partir de la exposición a la lengua.

En el caso de la lengua materna, la mayor parte del vocabulario se aprende de forma incidental, no formal, como describe Ellis (2000: 2): “by engaging in natural meaningful communication”. En las primeras etapas de la vida el incremento de vocabulario se hace de forma muy rápida y posteriormente se ralentiza, pero como hemos visto, se mantiene durante toda la vida. Un niño de unos cinco años conoce

aproximadamente de 4000 a 5000 palabras sin que se le hayan enseñado formalmente. A mayor exposición a la lengua, más y mejor aprendizaje.

Esto no quiere decir que el aprendizaje sea inconsciente. Sin embargo las posturas de los especialistas en cuanto a la conciencia difieren. Schmitt (1990) considera que para aprender una palabra hay que tener clara conciencia de ella mientras que Ellis (1997) reclama esta conciencia sólo para el aprendizaje de los aspectos léxicos (no para los *collocations*, forma de la palabra, pronunciación, etc.).

Según Schmitt, tanto el aprendizaje implícito como el explícito requieren *atención* por parte del sujeto. Para el implícito sería condición necesaria y suficiente.

En cuanto al aprendizaje de una lengua extranjera, Hatch y Brown (1995) valoran la importancia del aprendizaje incidental para completar al explícito, basándose en los datos de estudios (Saragi *et al*. 1978; Nagy *et al*. 1985; Dupuy y Krashen 1993) que demuestran una distancia entre el vocabulario que se enseña directamente y el aprendido después, por ejemplo, de la lectura de un libro en la lengua extranjera.

Nation (2001) recomienda que sean los términos técnicos y los poco frecuentes, por cuestiones prácticas, los que se dejen a la incorporación incidental.

En cualquier caso, una buena programación de la enseñanza del vocabulario debe incluir tanto técnicas que propicien el aprendizaje sistemático como una exposición a la lengua suficiente para que se produzca el aprendizaje incidental. Schmitt (2000: 137) comenta: “L2 learners benefit from a complementary combination of explicit teaching and incidental learning”. Y Bocanegra (2001: 35) insiste:

> La simple exposición a la lengua no es suficiente para que el alumno incorpore a su interlengua nuevos datos de forma efectiva. Es imprescindible, pues generar un aducto útil y es aquí donde el aula adquiere un papel fundamental.

4.3.6. Factores que afectan al aprendizaje de una palabra

En apartados anteriores hemos visto que el conocimiento ideal de una palabra incluye distintos rasgos como su forma oral y escrita, su estructura interna, su significado (referencial, afectivo y pragmático), sus relaciones léxicas con otras

palabras, su funcionamiento sintáctico y las *collocations* más frecuentes. Esta multiplicidad hace que a menudo el conocimiento de una palabra sea parcial.

También hemos apuntado la existencia de factores que determinan la mayor o menor dificultad de aprendizaje de una palabra o *learning burden*. En este trabajo dirigimos nuestro interés hacia los que inciden en el aprendizaje de lenguas extranjeras. Schmitt (2000) insiste en la necesidad de que el profesor conozca estos factores y los agrupa en intraléxicos y contrastivos.

Destacamos los estudios de Ellis y Beaton (1993)[44] y Laufer (1997)[45] al respecto. En gran medida, como hemos visto, la dificultad de aprendizaje depende de los modelos de la L1.

> In essence, the process of learning a FL word is to map a novel sound pattern (which will be variable across speakers, dialects, emphases, etc.) to a particular semantic field that may (or may not) have an exact equivalent in the native language. Even this rudimentary description implicates a range of relevant variables: pronounceableness, familiarity with semantic content and clear labelling of that meaning in the native language. (Ellis y Beaton 1993: 560)

Uno de los factores es la pronunciación, tanto los distintos fonemas como los rasgos suprasegmentales (acento, entonación). El sistema fonológico de la lengua materna condiciona el aprendizaje de la lengua extranjera. Por ejemplo, los alumnos españoles de Inglés como Lengua Extranjera encuentran problemas para pronunciar determinados sonidos que no existen en español (*shop, just*) o que no son distintivos (*ban/van, ship/sheep*) en nuestra lengua.

Estas dificultades pueden aumentar la distancia entre el conocimiento receptivo de la palabra y su correcta producción oral. La estrategia que propone Levenston (1979 citado en Laufer 1997) es evitar las palabras que presentan más dificultad fonológica en las fases iniciales del aprendizaje.

La ortografía es un segundo factor. Las combinaciones de letras que resultan conocidas son más sencillas. Por otro lado, la correspondencia entre la escritura de

[44] El estudio de los factores psicolingüísticos que determinan el aprendizaje de vocabulario de Ellis y Beaton (1993) presenta algunas limitaciones. Se centra exclusivamente en los estadios iniciales del aprendizaje de una lengua extranjera. El procedimiento mide el aprendizaje de pares de palabras en respuestas tipo, no indaga en la capacidad para utilizar el vocabulario aprendido en el contexto de la oración. Los autores reconocen la necesidad de continuar la investigación en este campo.

[45] El trabajo de Laufer (1997) se refiere primordialmente a los factores intraléxicos que determinan el aprendizaje del vocabulario, no obstante, en algunos aspectos incluye la comparación entre L1 y L2.

la palabra y su pronunciación facilita el aprendizaje. En este aspecto, la lengua inglesa no aporta muchas claves.

También el tamaño de las palabras podría afectar a su adquisición. En principio los estudios indican que a mayor longitud de la palabra mayor dificultad para ser aprendida. Sin embargo, algunos autores (Laufer 1997) ponen en duda esta presuposición argumentando, por ejemplo, que esto no ocurre cuando los morfemas que forman una palabra son bien conocidos por el alumno.

En las situaciones de aprendizaje todos estos factores de dificultad se entremezclan. En inglés las palabras cortas son también más frecuentes, por tanto parece lógico pensar que no es su longitud sino más bien la exposición a ellas la que facilita su aprendizaje (Laufer 1997).

Los aspectos morfológicos de la palabra también influyen, especialmente su complejidad de inflexión y derivación. Las irregularidades dificultan el aprendizaje, mientras que la habilidad del hablante para reconocer los distintos morfemas facilita el reconocimiento y la producción de una palabra nueva.

La semejanza de forma entre palabras de una y otra lengua puede ser una ayuda pero también puede llevar a confusión y actuar como interferencia. Este fenómeno se denomina *synformy*.

También se aprecian diferencias en el aprendizaje de las palabras atendiendo a su categoría gramatical. Ellis y Beaton (1993) corroboran que los nombres se aprenden con mayor facilidad que los verbos. Los adverbios, por el contrario, son los más difíciles. Laufer (1997) no comparte esta idea y argumenta que los estudios realizados no son del todo concluyentes. Además, analiza los rasgos semánticos que afectan al aprendizaje; como el grado de abstracción, el registro, la *idiomaticity* y los fenómenos de polisemia/homonimia.

Las palabras abstractas parecen más difíciles de aprender que las concretas, pero no siempre ocurre así. Laufer (1997) explica que a menudo su dificultad atiende a otros factores. En lo relativo al registro, vemos que cuando se aprende una lengua extranjera se prefiere utilizar términos generales aplicables a varios contextos. Con las expresiones idiomáticas ocurre lo mismo. En general preferimos utilizar un sinónimo que no sea *idiom*. Y la multiplicidad de forma o significado contribuye a aumentar el grado de dificultad que presenta una palabra.

El siguiente cuadro resume las conclusiones de Laufer (1997) en cuanto a los factores de dificultad intraléxicos.

Tabla 4.5. Factores intraléxicos que afectan al aprendizaje del vocabulario (Laufer 1997 en Schmitt 2000: 148-149)

Facilitating factors	*Difficulty-inducing factors*	*Factors with no clear effect*
Familiar phonemes	Presence of foreign phonemes	
Familiar letter combinations (sland)	Unfamiliar letter combinations (ndasl)	
Stress always on same syllable	Variable stress	
Consistency of sound script relationship	Incongruency in sound-script relationship	
		Word length
Inflexional regularity	Inflexional complexity	
Derivational regularity	Derivational complexity	
Transparency of word parts (preview = look before)	Deceptive transparency (outline ≠ out of line)	
	Similarity of word forms (affect/effect)	
		Part of speech
		Concreteness/abstractness

Ellis y Beaton (1993) culminan su trabajo con unas orientaciones de carácter práctico para el profesor. Recomiendan la combinación de técnicas de “palabras clave” y de “repetición”. Las primeras son eficaces para el aprendizaje receptivo y las segundas para el productivo. No obstante, dedicamos el siguiente apartado a la descripción de estrategias para el aprendizaje.

Swan (1997 citado en Schmitt 2000: 149) alude a los factores que dependen de la relación entre ambas lenguas:

> Informed teaching can help students to formulate realistic hypotheses about the nature and limits of crosslinguistic correspondences, and to become more attentive to important categories in the second language which have no mother-tongue counterpart.

4.3.7. Pasos en el aprendizaje de una palabra

Partiendo del estudio de Brown y Payne (1994) sobre las estrategias utilizadas en el aprendizaje del vocabulario, Hatch y Brown (1995: 374) identifican cinco pasos sucesivos desde que nos encontramos con una palabra nueva hasta que somos capaces de utilizarla:

1. Encuentro con la palabra nueva, de forma explícita o incidental.
2. Captación del significante.
3. Captación del significado mediante diversas estrategias.
4. Consolidación de significante y significado en la memoria.
5. Uso de la palabra (si se desea un conocimiento productivo).

Los estadios del proceso no son estancos y se pueden subdividir.

De nuevo hemos de aludir en este punto a la labor del profesor como *facilitador* del aprendizaje. Su papel es el de proporcionar al alumno estrategias que le hagan avanzar siguiendo estos pasos hasta el conocimiento productivo de la palabra, como veremos en el siguiente apartado.

4.3.8. Estrategias para el aprendizaje del vocabulario

El carácter pedagógico de esta tesis hace que no podamos terminar este apartado sin mencionar las estrategias para el aprendizaje del vocabulario (*vocabulary learning strategies*, VLS).

Chamot y O'Malley (en Ellis 2000) distinguen dos tipos; las que el alumno desarrolla por sí mismo cuando se enfrenta a un problema lingüístico y las que los profesores enseñan de forma explícita como parte de la instrucción. Las estrategias pueden implicar procesos conceptuales, afectivos y de interacción social.

Los alumnos que intentan aprender una lengua extranjera realmente utilizan estrategias para aprender el vocabulario, quizá por el carácter discreto del constructo y porque se considera un aspecto importante de la lengua (Chamot 1987; Horwitz 1988).

Las de tipo mecánico son las que se aprecian con más facilidad en el contexto del aula, por ejemplo la memorización, toma de apuntes, etc.

El profesor puede proporcionar glosarios, explicación oral rápida, etc. Pero también el propio alumno puede aplicar estrategias: la primera evaluar si es necesario para la comprensión del texto. Si el término no lo es, ignorarlo. Si lo es, inferir su significado a partir del contexto, preguntarlo o buscarlo en el diccionario. La inferencia es la más recomendable (Read 2000: 53), e insistiremos en ella más adelante, puesto que su manejo resulta fundamental para resolver un C-test.

> Inferencing is a desirable strategy because it involves a deeper processing that is likely to contribute to better comprehension of the text as a whole and may result in some learning of the lexical item that would not otherwise occur.

Bocanegra y Franco (2003) confirman la existencia de aprendizaje estratégico en alumnos españoles de Inglés como Lengua Extranjera, y en mayor medida cuando el nivel de competencia en la lengua meta aumenta[46].

Las primeras clasificaciones de estrategias del alumno se hicieron a partir de la descripción de los buenos aprendices de lenguas (Rubin 1975; Stern 1975). Se utilizaron entrevistas, protocolos *think aloud*, observación directa, cuestionarios, etc.

Laufer y Hulstijn (2001: 5) confirman que “successful learners use sophisticated metacognitive learning strategies, such as inferring word meanings from context and semantic or imagery mediation, in this endeavour”.

A partir de dichas descripciones Chamot y O’Malley (1990, 2000) distribuyen las estrategias en tres grupos: metacognitivas, cognitivas y socio-afectivas.

Destacamos las taxonomías de Schmitt y McCarthy (1997) y Nation (2000) porque agrupan las estrategias específicas para el aprendizaje del vocabulario.

Schmitt y McCarthy organizan su taxonomía en dos grandes grupos de estrategias; las que sirven para descubrir el significado de una palabra y las que se utilizan para consolidarlo. Esta clasificación sigue el modelo de Oxford (1990) que las agrupa en memorísticas (MEM), sociales (SOC), cognitivas (COG) y metacognitivas (MET). A éstas añaden las determinativas (DET) como estrategias de descubrimiento. Las estrategias memorísticas (o *mnemonics*) relacionan la

[46] Fernández Toledo (2003) y Fonseca (2003) son otros trabajos sobre estilos y estrategias de aprendizaje en la enseñanza de lenguas extranjeras en España.

palabra nueva con conocimientos previos tratando de crear asociaciones que faciliten la producción (*recalling)*. Las sociales utilizan la interacción con el profesor o con otros alumnos para facilitar el aprendizaje. Las cognitivas son semejantes a las memorísticas, e incluyen la toma de apuntes, la repetición oral o escrita. Finalmente, las metacognitivas implican la existencia de una visión consciente del proceso de aprendizaje por parte del alumno, que valora y toma sus decisiones.

Tabla 4.6. Taxonomía de Schmitt y McCarthy (1997: 207-208)

A taxonomy of vocabulary learning strategies

Strategy Group		Use %	Helpful %
Strategies for the discovery of a new word's meaning			
DET	Analyse part of speech	32	75
DET	Analyse affixes and roots	15	69
DET	Check for L1 cognate	11	40
DET	Analyse any available pictures or gestures	47	84
DET	Guess from textual context	74	73
DET	Bilingual dictionary	85	95
DET	Monolingual dictionary	35	77
DET	Word lists	-	-
DET	Flash cards	-	-
SOC	Ask teacher for an L1 translation	45	61
SOC	Ask teacher for paraphrase or synonym of new word	42	86
SOC	Ask teacher for a sentence including the new word	24	78
SOC	Ask classmates for meaning	73	65
SOC	Discover new meaning through group work activity	35	65
Strategies for consolidating a word once it has been encountered			
SOC	Study and practice meaning in a group	30	51
SOC	Teacher checks students' flash cards or word lists for accuracy	3	39
SOC	Interact with native-speakers	-	-
MEM	Study word with a pictorial representation of its meaning	-	-
MEM	Image word's meaning		
MEM	Connect word to a personal experience	37	62
MEM	Associate the word with its coordinates	13	54
MEM	Connect the word to its synonyms and antonyms	41	88
MEM	Use semantic maps	9	47
MEM	Use "scales" for gradable adjectives	16	62
MEM	Peg Method	-	-
MEM	Loci Method	-	-
MEM	Group words together to study them	-	-
MEM	Group words together spatially on a page	-	-
MEM	Use new words in sentences	18	12
MEM	Group words together within a storyline	-	-
MEM	Study the spelling of a word	74	87
MEM	Study the sound of a word	60	81

MEM	Say new word aloud when studying	69	91
MEM	Image word form	32	22
MEM	Underline initial letter of the word	-	-
MEM	Configuration	-	-
MEM	Use Keyword Method	13	31
MEM	Affixes and roots (remembering)	14	61
MEM	Part of speech (remembering)	30	73
MEM	Paraphrase the word's meaning	40	77
MEM	Use cognates in study	10	34
MEM	Learn the words of an idiom together	48	77
MEM	Use physical action when learning a word	13	49
MEM	Use semantic feature grids	-	-
COG	Verbal repetition	76	84
COG	Written repetition	76	91
COG	Word lists	54	67
COG	Flash cards	25	65
COG	Take notes in class	64	84
COG	Use the vocabulary section in your textbook	48	76
COG	Listen to tape of word lists	-	-
COG	Put English labels on physical objects	-	-
COG	Keep a vocabulary notebook	-	-
MET	Use English-language media (songs, movies, newscasts, etc.)	-	-
MET	Testing oneself with word tests	-	-
MET	Use spaced word practice	-	-
MET	Skip or pass new word	41	16
MET	Continue to study word over time	45	87

= Strategy was not included on the initial list used in the survey

La clasificación de Nation (2000) es totalmente diferente. El autor divide las VLS en tres tipos: las que se utilizan para planificar el aprendizaje, las que suponen una búsqueda de información en fuentes distintas y las que sirven para fijar el conocimiento de las palabras.

Figura 4.2. A taxonomy of kinds of vocabulary learning strategies (Nation 2000: 218)

General class of strategies	Types of strategies
Planning: choosing what to focus on and when to focus on it	Choosing words Choosing the aspects of word knowledge Choosing strategies Planning repetition
Sources: finding information about words	Analysing the word Using context Consulting a reference source in L1 or L2 Using parallels in L1 and L2
Processess: establishing knowledge	Noticing Retrieving Generating

Resulta de gran utilidad para nuestros alumnos disponer de una serie de estrategias que faciliten la tarea de aprender vocabulario. Corresponde a los profesores de lenguas extranjeras elegir las más adecuadas para sus alumnos y entrenarlos en su manejo (Hatch y Brown 1995; Schmitt 2000).

Algunas de estas estrategias de descubrimiento o consolidación se utilizan también en la resolución de pruebas de evaluación. En capítulos posteriores veremos qué estrategias se utilizan para resolver un C-test, como la inferencia mediante el uso de las claves que ofrece el contexto.

4.4. Investigaciones sobre evaluación del vocabulario

Gran parte de la investigación sobre evaluación del vocabulario ha sido realizada por expertos en otras áreas, como la de adquisición de segundas lenguas (SLA), adquisición de vocabulario, lectura en L1, etc. Los expertos en evaluación de lenguas a menudo han dejado de lado las pruebas específicas de vocabulario en favor de las pruebas de evaluación de la competencia lingüística comunicativas e integradoras.

Desde el momento en que se reconoce la importancia del vocabulario en el aprendizaje de lenguas surge la preocupación por encontrar instrumentos adecuados para su evaluación: “Development of lexical knowledge is now regarded, by both researchers and teachers, as central to learning a language, and thus vocabulary tests are being used for a wide variety of purposes” (Read y Chapelle 2001: 3).

No es tarea fácil. Ya hemos mostrado la preocupación al respecto de Meara (1996) y Schmitt (1998) que recoge y comparte Read (2000: 27): “concerning the practical difficulties involved both in developing suitable measures and in eliciting evidence of learners’ knowledge”. Para ello hay que partir de una definición clara del constructo del vocabulario (véase 4.2.4). Read y Chapelle (2001) clasifican los distintos enfoques en tres grupos:

1. Los que consideran el constructo del vocabulario como componente discreto dentro del conocimiento de la lengua.
2. Los que cuestionan la visión anterior y entienden que el constructo del vocabulario debe estar integrado en la competencia lingüística general.
3. Los que lo utilizan como herramienta para la investigación de otros aspectos de la lengua.

De cada uno de ellos surgirán pruebas de vocabulario diferentes, pues la definición del constructo y el propósito también lo son.

En la literatura vemos que los exámenes de vocabulario se han utilizado tanto con fines académicos como científicos o de investigación. (Read y Chapelle 2001). Por tanto, un primer paso al diseñar una prueba es saber su propósito.

Las pruebas diseñadas con fines académicos pueden ir encaminadas a conocer la amplitud del vocabulario del alumno o bien a medir su profundidad. Son las dos dimensiones del conocimiento del vocabulario que aumentan gradualmente a medida que aumenta la competencia lingüística del alumno en la lengua objeto de estudio (Laufer *et al*. 2004).

Si lo que interesa en el número de palabras que ha aprendido el alumno, los retos son seleccionar el vocabulario objeto de la prueba y elegir el formato que se va a aplicar. Por el contrario, si nos centramos en el grado de conocimiento de las unidades léxicas, el problema es encontrar instrumentos que realmente indiquen cómo es ese conocimiento (parcial-preciso, productivo-receptivo, etc.).

Así lo expresa Bogaards (2000: 490):

> Depending on what exactly one wants to know about L2 lexical knowledge, one has to select the appropriate materials and adequate procedures to arrive at valid and reliable results.

Por otra parte, los instrumentos para medir el aprendizaje del vocabulario deben reunir los rasgos deseables para cualquier otro tipo de prueba de evaluación; esto es, fiabilidad, validez, carácter práctico, *washback*, etc. (Véase capítulo 3)

4.4.1. El estudio del vocabulario: Perspectiva histórica

En este apartado haremos un breve recorrido histórico por la investigación del vocabulario. Revisaremos las tendencias más importantes en cuanto a su evaluación a lo largo del tiempo. Finalmente, nos centraremos en las actuales para enmarcar en ellas la prueba que nos ocupa en esta tesis; el C-test.

Desde una perspectiva histórica, el papel del vocabulario dentro de la enseñanza de lenguas ha pasado por momentos de gran preponderancia y otros en que ha sido menospreciado.

Si bien su estudio fue muy valorado en la antigua Roma, en la época medieval y en el Renacimiento se dejó de lado en favor de la gramática. Ya en el siglo XVII aparecen tratados que reivindican el papel del vocabulario en la enseñanza del inglés (Comenius y William of Bath). Pero durante los siglos XVIII y XIX, a pesar de la publicación del *Dictionary of the English Language* de Samuel Johnson en 1755, el vocabulario se mantuvo en un papel secundario frente a la gramática.

No obstante, a finales del XIX hay que destacar los trabajos de Ebbinghaus sobre la adquisición del vocabulario desde el punto de vista psicológico. Investigó las conexiones entre las palabras en la mente y su estudio de las "asociaciones de palabras" supuso el inicio de una larga serie de investigaciones posteriores.

A principios del siglo XIX triunfaba el método *Grammar-Translation.* El vocabulario era el soporte de las reglas gramaticales e instrumento para la traducción. Cobraron gran relevancia las listas bilingües y los diccionarios. Pero no interesaba el uso de la lengua sino su análisis. A finales del siglo surgió el *Direct Method*, que aboga por la exposición a la lengua (*listening*) y evita en lo posible la traducción. Pretende imitar la forma en que se adquiere la lengua materna. En cuanto al vocabulario, la idea es que se adquiera de forma natural, en lo posible (excepto p. ej. los términos abstractos). Después apareció el *Reading Method*, que enfatizaba el aprendizaje de la lectura. Paralelamente en Estados Unidos y en Gran Bretaña surgieron métodos que partían del behaviorismo (*habit formation*).

A raíz de la experiencia de enseñanza de idiomas con soldados durante la Segunda Guerra Mundial, se desarrolla en Estados Unidos el *Audiolingualism:* "It was assumed that good language habits and exposure to the language itself, would eventually lead to an increased vocabulary" (Coady 1993: 4, en Schmitt 2000: 13).

En Gran Bretaña aparece un enfoque semejante, el *Situational Approach.* Estructuras gramaticales y vocabulario se agrupan según la situación en que se utilizan.

Ya en los años 70 llega el enfoque comunicativo con el *Communicative Language Teaching*; enfatiza los aspectos sociolingüísticos y pragmáticos de la lengua, y el vocabulario pasa de nuevo a un papel secundario.

En la actualidad, como hemos visto desde la introducción de este capítulo, se reconoce de nuevo el papel fundamental del vocabulario en el aprendizaje de lenguas extranjeras:

> One of the most important current trends of thought is the realization that grammar and vocabulary are fundamentally linked. [...] Pursuing this idea should finally put to rest the notion that a second language can be acquired without both essential areas being addressed. (Schmitt 2000: 14)

Los modelos metodológicos a menudo no sabían como enfrentarse con la enseñanza del vocabulario, que quedaba relegado a las listas bilingües o se confiaba en su adquisición por exposición a la lengua. Hasta el siglo pasado no se inicia un trabajo sistemático en el vocabulario.

En los años 30, Ogden y Richards crearon un corpus de vocabulario del inglés que sólo incluía 850 palabras (*Basic English*). Pretendían reducir al mínimo el vocabulario necesario para comunicarse en lengua inglesa. Pero este intento dio como resultado una lista claramente artificial e insuficiente.

Otro enfoque de la época que surgió como reacción al *Basic English* es el *Vocabulary Control Movement.* Este método intentó buscar criterios válidos para la selección de vocabulario con la finalidad de simplificar los textos utilizados en la enseñanza de la lectura en lengua extranjera (*graded texts*). Uno de los principales era la frecuencia de las palabras en la lengua. El producto final fue una lista de unas 2000 entradas, la *General Service List of English Words* (GSL) de West (1953).

4.4.2. La evaluación del vocabulario en el siglo XX

Si en el apartado anterior mencionamos los trabajos de Ebbinghaus sobre el vocabulario, aquí volvemos a él como uno de los primeros investigadores modernos

preocupados por su evaluación (Schmitt 2000). Su contribución fue el diseño de un modelo para la autoevaluación.

A comienzos del siglo XX era patente la necesidad de contar con pruebas que midieran de forma exacta y fiable el conocimiento del vocabulario de una lengua. En la primera mitad del siglo, con la psicometría, alcanzaron un gran desarrollo las pruebas de tipo objetivo, especialmente en los Estados Unidos. A partir de los años 30 los ensayos tradicionales dieron paso a estas pruebas estandarizadas psicométricas. Medían el reconocimiento del vocabulario mediante la asociación de palabras con su traducción (*matching activities*) y con ejercicios de elección múltiple (*múltiple-choice type items*). Son pruebas objetivas porque su corrección no requiere el juicio del examinador, a cada pregunta le corresponde una sola respuesta correcta que se puede predecir. Destaca su fiabilidad, la facilidad de su diseño y su buena correlación con otras pruebas, como las de comprensión lectora. Fueron el germen del *Test of English as a Foreign Language* (TOEFL) que apareció en los años 60 y sigue vigente en nuestros días.

En 1961 Lado propuso su modelo centrado en la evaluación individual de los distintos elementos de la lengua. A partir de entonces toda prueba de elementos discretos incluía un *test* objetivo de vocabulario.

Con el auge del movimiento comunicativo en los años 70 cambió la manera de entender las pruebas de vocabulario. Interesa medir el conocimiento de las palabras en un contexto y no aisladas.

Otra tendencia actual de evaluación del vocabulario es el diseño de pruebas integradas con elementos discretos (Schmitt 2000).

4.4.3. Panorama actual en la evaluación del vocabulario

Ya hemos visto la complejidad de la evaluación del vocabulario, que hace necesario definir el propio constructo para poder después validar las pruebas. A continuación, vemos las últimas tendencias en su evaluación.

4.4.3.1. Tendencias actuales de evaluación del vocabulario

En el modelo de Bachman y Palmer (1996), ampliamente aceptado y adoptado en nuestros días, la competencia general en la lengua comprende dos aspectos: conocimiento lingüístico y competencia estratégica.

La teoría sobre evaluación del vocabulario hoy tiende a dejar atrás las pruebas objetivas de elementos discretos porque ignoran todo lo referente al segundo aspecto: la competencia estratégica. El enfoque comunicativo, en que seguimos inmersos, propicia más la evaluación de la competencia lingüística general que la de los distintos elementos de la lengua. Este marco entiende que el conocimiento del vocabulario de una lengua no garantiza el manejo de ésta en situaciones reales de comunicación. Por tanto, no interesan las palabras aisladas, sino insertas en el contexto comunicativo. Las pruebas de evaluación plantean tareas que simulan situaciones de comunicación. La tarea pasa a ser el objetivo de las pruebas (Bachman y Palmer 1996) mientras que el vocabulario, la gramática, etc. serán de gran ayuda para resolverla, pero no determinantes.

Sin embargo, en la práctica se siguen utilizando pruebas de elementos discretos y a menudo descontextualizados. Read y Chapelle (2001) achacan la desconexión de la práctica docente con las últimas tendencias sobre evaluación a la falta de un marco que fije los objetivos claros de las pruebas y su diseño[47].

Read (2000) ofrece una visión conciliadora. Las pruebas discretas de vocabulario pueden ser complementarias de las globales o integradoras. Dependerá, entre otros, de los objetivos que pretenda el profesor y de la situación en que las utilice. El criterio del profesor, que conoce a sus alumnos y sus necesidades, será el que mejor guíe la elección del modelo y formato de evaluación en cada momento. En este sentido se expresa Bogaards (2000: 490):

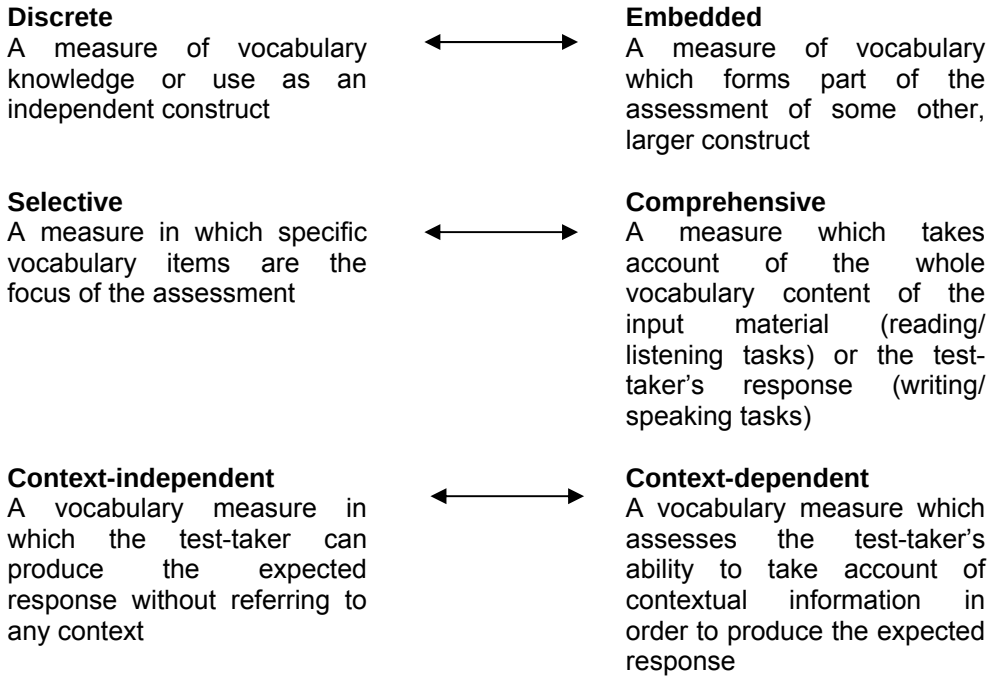
> As lexical knowledge comes in very many forms and presents a lot of different aspects, this means that there is not one single valid way to measure L2 vocabulary knowledge. Different types of tests are needed to address different aspects of the lexicon and different formats may be more or less adapted to different levels of vocabulary knowledge and to different types of questions the teacher or the researcher wants to answer.

[47] Read y Chapelle (2001) proponen un marco para la evaluación del vocabulario que se basa en el propósito de la prueba y las decisiones sobre el diseño para llegar a su validación. Toma en consideración aspectos como los usos y relevancia de la prueba, el impacto y presentación, etc.

Read (2000) hace una clasificación de los tipos de pruebas de vocabulario atendiendo a tres dimensiones: constructo, rango y contexto.

La primera tiene en cuenta el constructo: se puede evaluar el vocabulario como constructo discreto o integrado en pruebas más generales; la segunda se refiere al rango de vocabulario que incluye, si las pruebas miden una parte específica del léxico son pruebas selectivas frente a las comprensivas; y por último, dependiendo de su relación con el contexto, los exámenes pueden ser dependientes o contextualizadas e independientes de éste.

Figura 4.3. Dimensiones de la evaluación del vocabulario (Read 2000) y ejemplos de clasificación de pruebas (Read y Chapelle 2001)



Read y Chapelle (2001) entienden que el C-test es una prueba discreta, selectiva en cuanto al rango, y contextualizada. Volveremos sobre esta clasificación en el capítulo 6. En el C-test apreciamos la doble vertiente que ya hemos comentado. Es un diseño que mide de forma objetiva elementos discretos, pero a la vez es una prueba contextualizada e integradora.

A pesar de ser objetiva en cuanto a su corrección, en la Perspectiva Empírica veremos su alta correlación con pruebas subjetivas. Por otra parte, coincidiendo con

estudios previos (Klein-Braley 1985, 1997; Chapelle y Abraham 1990; Dörnyei y Katona 1992; Connelly 1997; Babaii y Ansary 2001; Eckes y Grotjahn 2006) demostraremos que no sólo mide el vocabulario, sino la competencia global en Lengua Extranjera.

4.4.3.2. Estudios recientes sobre vocabulario en España

Si bien la evaluación del aprendizaje de lenguas es un tema todavía relativamente poco explorado en nuestro país, como mencionamos en el capítulo 1, tenemos que hacer constar que no ocurre lo mismo con el campo del vocabulario.

Una revisión rápida de las últimas publicaciones de la Asociación Española de Lingüística Aplicada (AESLA) pone de manifiesto la riqueza y variedad de los trabajos de investigación relacionados con el vocabulario realizados en España.

A algunos de ellos hacemos referencia en esta tesis; como Graña López (1997), que estudia la relación entre frecuencia de las palabras y procesamiento léxico, Suau (1998), que trabajó sobre la inferencia léxica, o Cabré y Adelstein (2001) sobre terminología. Acerca del aprendizaje de vocabulario destacamos los estudios de Alcón Soler (1999) y Salazar y Alcón (2001).

Las recopilaciones *Trabajos en lingüística aplicada* (2001), *La lingüística aplicada a finales del siglo XX* (2001) y *Perspectivas recientes sobre el discurso* (2001) agrupan un buen número de artículos sobre lexicología y lexicografía.

Una parte importante de lo publicado lo constituyen los estudios sobre lingüística de corpus y computacional (Valero *et al.* 2001), así como los relacionados con la lingüística contrastiva y la traducción (Valero 1999).

4.5. Las pruebas de vocabulario

Ponemos punto final a este capítulo mostrando, al menos de manera esquemática, los distintos tipos de pruebas de vocabulario y aportando algunos ejemplos de pruebas estandarizadas.

4.5.1 Tipos de pruebas de vocabulario

Aunque las pruebas de vocabulario pueden clasificarse atendiendo a diversos criterios, las dividiremos dos grandes grupos: de elementos discretos y holísticas. Las primeras son las que tradicionalmente se han considerado pruebas objetivas de vocabulario, porque se centran en este constructo, considerado de forma aislada. Deberíamos incluir en este grupo a los *clozes*, pero preferimos dedicarles su propio apartado por su significación para esta tesis. Además, las pruebas de cierre miden la competencia lingüística global más que el vocabulario. En capítulos posteriores discutiremos ampliamente éste y otros aspectos de las pruebas de cierre.

No podemos terminar este apartado sin hacer notar la dificultad que supone separar los distintos elementos de la lengua para su evaluación. Por ello, en la práctica, a menudo la evaluación del vocabulario aparece ligada a la de otros aspectos de la competencia lingüística, como los gramaticales (Read 2000: 99).

4.5.1.1. Pruebas objetivas de elementos discretos

Hemos visto que a comienzos del siglo XX triunfa el movimiento psicométrico y con él las pruebas de evaluación objetiva, sobre todo en Estados Unidos.

Dentro de las pruebas objetivas de evaluación de lenguas extranjeras el vocabulario era un componente habitual, con preguntas de elección múltiple, listas de palabras para realizar *matching*, etc. Read (2000) señala la facilidad de diseño de estas pruebas, sus características técnicas, y el hecho de que además de medir el vocabulario indicaban aún sin pretenderlo la competencia global en la lengua.

Con Lado (1961) siguió esta tendencia de evaluar el vocabulario de forma objetiva y aislado de otros elementos de la lengua. Tanto él, en *Language Testing,* como autores posteriores, recomiendan este tipo de pruebas. Pero cuando llegó el movimiento comunicativo se dejó de lado a las pruebas objetivas descontextualizadas. A pesar de todo, es de destacar la popularidad de las pruebas de cierre en los años 70.

A continuación mostramos brevemente los formatos más comunes de las pruebas objetivas:

- elección múltiple,
- asociaciones,
- traducción,
- listas de control.

Las pruebas de elección múltiple (*multiple choice*) han gozado de gran popularidad en la evaluación de vocabulario en lengua nativa y extranjera. Su uso es frecuente incluso en la actualidad; de hecho, popularmente se llega a identificar al test o prueba objetiva con el formato de elección múltiple.

Sin embargo, se han criticado algunos aspectos de la prueba que ponen de manifiesto ciertas limitaciones. Por ejemplo, que miden el reconocimiento y no la producción, además dejan margen al alumno para elegir la respuesta correcta mediante un proceso de eliminación, como afirman Laufer *et al*. (2004: 208): "Hence, recalling a word's meaning or form can be considered as a more advanced type of knowledge than recognizing it in a set of options". Por tanto, la información que recibe el profesor acerca del grado de conocimiento del vocabulario es muy limitada si el único instrumento de medida es una prueba de elección múltiple.

Además, en las pruebas de elección múltiple el profesor no puede saber si el alumno realmente conoce la palabra correcta o las que maneja son los "distractores". El diseño de un buen test de elección múltiple requiere mucho tiempo de preparación, si bien después se ve contrarrestado por una fácil aplicación y corrección.

Las pruebas de asociación (*matching exercises*), por otra parte, son fáciles de diseñar pero también presentan limitaciones. Consisten en asociar distintos elementos con otros que tengan el mismo significado. A igual que las de elección múltiple, son sólo de reconocimiento, expresan, por tanto, un conocimiento parcial y pobre de la palabra. No aportan al examinador una idea de la profundidad del conocimiento del alumno, sino exclusivamente de su amplitud. Nation (1990) usa este formato como parte del *Vocabulary Levels Test*.

En cuanto a la traducción L1-L2, las listas de palabras aisladas para traducir a la lengua extranjera son un diseño ya muy poco utilizado en las pruebas de vocabulario. No obstante, Nurweni y Read (1999) las introdujeron en un reciente trabajo de investigación.

Por último, el procedimiento conocido como c*hecklists* o listas de control es el más sencillo. El alumno debe marcar las palabras que reconoce de una lista que se le ofrece. Sin embargo, de nuevo, el grado de conocimiento de las mismas no se puede rastrear. Por tanto, este formato puede ser útil para investigar la amplitud del vocabulario, sobre todo si se utiliza en combinación con otras técnicas, pero no su profundidad.

4.5.1.2. Holísticas o integradoras: *Comprehensive measures of vocabulary*

Son las pruebas que miden el conocimiento léxico del alumno de una forma más amplia. Son comprensivas, integradoras. Generalmente el vocabulario no es el constructo que se pretende medir, sino un elemento más de otro constructo más amplio, la competencia comunicativa oral o escrita en la lengua extranjera.

Algunas sí consideran la adquisición de vocabulario como constructo aislado, pero son medidas cuantitativas, de tipo meramente estadístico, cuya finalidad no es la evaluación educativa sino la de informar las investigaciones del aprendizaje de vocabulario en lenguas extranjeras.

Distintos estudios (Kelly 1991; Brown 1997; Laufer 1997 en Read 2000) coinciden en señalar la importancia de los conocimientos léxicos del alumno para la comprensión auditiva y lectora de los textos que se les presentan.

El ensayo es una prueba holística, cuando se propone al alumno la realización de una composición escrita sobre un tema dado, se puede medir de forma estadística la riqueza léxica. Se aplican medidas que ya hemos comentado en este capítulo, como la densidad y la variación léxica. Se puede tener en cuenta, además, la sofisticación del texto producido (analizando la proporción de términos poco frecuentes o técnicos que contiene) y el número de errores en el uso del vocabulario. Con estos datos se obtiene una estimación de la riqueza léxica del texto. Y de forma semejante se puede aplicar este marco a la producción oral.

Sin embargo, este tipo de análisis es costoso en términos de tiempo, pese a la ayuda de los medios informáticos. Por ello, suele quedar relegado a la investigación.

En contextos educativos, los profesores valoran las composiciones de sus alumnos de forma global u holística o bien con un enfoque analítico, con la ayuda de

escalas. Por la propia naturaleza de la tarea, en ambos enfoques surge la duda del sesgo que supone la subjetividad en la valoración del profesor.

Para una mayor profundización en este tema sugerimos la lectura de los trabajos de Weir (1990), Hamp-Lyons (1991), y el más reciente de Amengual Pizarro (2003) centrado en la realidad de nuestro país (redacciones de las PAAU). En éste último se plantea una cuestión interesante: la existencia de discrepancia entre las puntuaciones holísticas y analíticas de un mismo corrector en la evaluación de la expresión escrita.

4.5.1.3. Pruebas de cierre*: Clozes*

Desde la llegada del movimiento comunicativo se ha estudiado profusamente sobre las pruebas de cierre. No son propiamente pruebas de vocabulario, pero para su realización demandan del alumno un conocimiento léxico. Generalmente se consideran pruebas objetivas, sin embargo, son un grupo lo suficientemente amplio y peculiar (Read 2000) como para merecer su propio epígrafe en la clasificación.

Ya hemos señalado en apartados anteriores que sus rasgos peculiares las hacen partícipes de características propias de las pruebas objetivas de elementos discretos y también de otras que se atribuyen a las holísticas. Son objetivas en cuanto a su corrección, pero los estudios muestran su alta correlación con pruebas integradoras.

Los *clozes* comenzaron con Taylor (1953) como prueba para medir la legibilidad de los textos. Posteriormente fueron firmemente respaldados por Oller (1979) como medida eficaz de la competencia global en lengua extranjera. Enfrentan al alumno con un texto en el que se han omitido una serie de palabras, siguiendo distintos criterios, y proponen su recuperación. La discusión en torno a las bondades o no de los *clozes* ha llegado hasta nuestros días. Se sigue investigando para discernir qué miden realmente, su validez y fiabilidad.

En los próximos capítulos (5 y 6) estudiaremos detalladamente los rasgos de las pruebas de cierre, entre las que se incuye el C-test.

4.5.2. Ejemplos de pruebas estandarizadas de vocabulario

No queremos cerrar este capítulo sin mostrar alguno de los exámenes de vocabulario en lengua inglesa que se están aplicando en la actualidad. Read (2000) analiza cuatro de los más populares en su libro *Assessing Vocabulary*:

- *Vocabulary Levels Test*
- *Eurocentres Vocabulary Size Test* (EVST)
- *Vocabulary Knowledge Scale* (VKS)
- *Test of English as a Foreign Language* (TOEFL)

Son pruebas diferentes pero coinciden en algunos aspectos; todas suponen intentos serios de evaluar el aprendizaje del vocabulario, su formato es sencillo, son relativamente actuales, han gozado de una amplia difusión y prestigio, y todas ellas siguen siendo sometidas a estudios y revisiones para comprobar su validez.

La más antigua es el TOEFL, que nació en 1964 en el *Educational Testing Service* de Princeton. En los años 80 aparecieron el *Vocabulary Levels Test* y el *Eurocentres Vocabulary Size Test* (EVST). Después, ya en la década de los 90, nació el *Vocabulary Knowledge Scale* (VKS). Los tres primeros intentan medir la cantidad de vocabulario que maneja el alumno, mientras que el VKS pretende indagar en la calidad del mismo.

El *Vocabulary Levels Test* fue diseñado por Nation a comienzos de los 80 como instrumento para la programación de la enseñanza del vocabulario y se ha utilizado como prueba de diagnóstico de la amplitud de vocabulario. Se basa en la frecuencia de las palabras en inglés y se organiza en torno a varios niveles (2000, 3000, 5000 palabras, etc.) Incluye ejercicios de *matching* (palabras con su definición). Se sigue trabajando en él; versiones posteriores (Laufer y Nation 1999) introducen actividades de *blank-filling* muy cercanas al formato del C-test.

El EVST fue creado por Meara y su grupo a finales de los 80 como prueba de nivel. Mide la amplitud del vocabulario del alumno a partir de *checklists.* Su administración se hace mediante el uso del ordenador.

El VKS es un reciente intento de evaluar la calidad y profundidad del conocimiento del vocabulario. A finales de los 90 Paribakht y Wesche prepararon

este examen para su uso en la investigación de la adquisición accidental (*incidental*) de vocabulario en alumnos que aprenden inglés como lengua extranjera. Presenta al alumno una serie de palabras y una escala para que haga su propia valoración de hasta qué punto sabe cada una de ellas.

El TOEFL está ya totalmente institucionalizado. Su propósito era conocer el nivel de inglés de los estudiantes extranjeros que pretendían acceder a universidades americanas. Consta de varias secciones, una de las cuales es el vocabulario. A lo largo de su historia ha ido evolucionando en consonancia con las tendencias en evaluación. Durante años ha utilizado el formato de elección múltiple, pero a partir de las últimas revisiones tiende a contextualizarse. La tarea que propone es buscar el sinónimo de un término dentro de una frase o texto breve.

Por otra parte, ya centrándonos en España, veremos que dentro de la parte objetiva de la Prueba de Inglés de las PAAU hay una pregunta específica de vocabulario. En la Perspectiva Empírica de la tesis se estudia cómo correlaciona esta sección de la prueba con el C-test.

CAPÍTULO 5. LAS PRUEBAS DE CIERRE

5.1. Introducción

Iniciamos un capítulo fundamental para el desarrollo de esta tesis. En él identificamos al C-test como un tipo de prueba de cierre. Para comenzar, abordamos el concepto de *cloze*. Nos remontamos a la Psicología de la Gestalt para buscar el origen de esta técnica, creada por Taylor en los años 50. Revisamos sus características y los distintos tipos de pruebas de cierre. Después, profundizamos en el C-test, su diseño y características, sus ventajas y sus puntos débiles. Finalizamos con un repaso de la literatura sobre el C-test y las investigaciones recientes más significativas que informan nuestro trabajo experimental con dicho instrumento de evaluación.

5.2. Concepto de "prueba de cierre" o *cloze technique*

La aparición de la técnica "de cierre" o *cloze technique* en la década de los 50 supuso el comienzo de un nuevo procedimiento de diseño de pruebas integradoras para medir la competencia lingüística.

Fue creada por Wilson L. Taylor (1953). Consiste en la mutilación de un texto mediante la omisión de un determinado número de sus elementos. La tarea que se propone al alumno es la recuperación del texto original (Taylor 1953: 416):

> A cloze unit may be defined as: any single occurrence of a successful attempt to reproduce accurately a part deleted from a "message" (any language product), by deciding form the context that remains, what the missing part should be.

Las llamadas pruebas "de cierre" constituyen un procedimiento pragmático de evaluación de la lengua (Oller 1979: 42). Taylor acuñó el nombre de *cloze* para este tipo de pruebas. La palabra *cloze* es, según Oller, "a spelling corruption of the word *close*" (op. cit.: 341). Inventó esta denominación porque al rellenar los huecos de un texto previamente mutilado realizamos un acto de *closure*, de cierre, de modo semejante a lo que ocurre en la percepción de modelos visuales incompletos, según la Psicología de la forma o Gestalt[48]. Oller (1979: 42) lo explica así: "Taylor considered words deleted from prose to present a special kind of closure problem".

Las pruebas de cierre permiten valorar la competencia lingüística global de un alumno porque reducen la redundancia de la lengua y obligan al alumno a aplicar la gramática de expectativas de que dispone. Para aportar una respuesta correcta a una omisión el alumno necesita tener en cuenta la información lingüística procedente del propio texto, pero también ha de hacer inferencias del contexto extralingüístico.

La técnica de cierre resultó ser un procedimiento de evaluación prometedor que se ha utilizado mucho en la enseñanza e investigación de lenguas extranjeras, ya que, tras múltiples investigaciones (ver abundante literatura al respecto), ha demostrado ser un método práctico, válido y fiable, además de producir un efecto rebote positivo.

El C-test es un tipo de prueba de cierre que surgió posteriormente, y que reúne las características deseables para toda prueba de evaluación de la lengua, comentadas en el capítulo 3. No obstante, también se le achacan algunas deficiencias. Nos ocupamos de su análisis en el próximo capítulo.

5.3. La Psicología de la Gestalt

La escuela psicológica gestaltiana comenzó en Frankfurt am Main en 1910-12 con los estudios de Wertheimer (1880-1943), Khöler (1887-1967) y Koffka (1886-1941) en el campo del aprendizaje y la percepción.

[48] Oller (1979: 341) se refiere al origen del término *cloze* y apunta lo siguiente: "The term is a mnemonic or perhaps a humorless pun intended to call to mind the process of closure celebrated by Gestalt psychologists". En esta tesis se utilizan indistintamente las denominaciones "prueba de cierre" y *cloze*. Mantenemos el término inglés por su claridad y amplia difusión internacional.

La teoría de esta escuela psicológica, así como su influencia en la psicología posterior, es mucho más amplia y compleja que lo expresado en este apartado, pero por razones de espacio nos centraremos exclusivamente en los aspectos más significativos que explican cómo se relaciona la Gestalt con las pruebas de cierre.

Los psicólogos gestaltianos preferían el método experimental. Comenzaban el análisis por la totalidad para ir centrándose después en las partes que la forman. Utilizaban la técnica del análisis fenomenológico, que parte de una experiencia perceptiva.

El concepto clave es G*estalt*. Tiene tres significados: un todo o sistema (*Ganzheit*), estructura o configuración (*Struktur*) y propiedad emergente o sistemática (*Gestalt-qualität*). Para la psicología de la Gestalt un todo "es una configuración compleja cuyos componentes se encuentran relacionados entre sí [...] todos los componentes adquieren significación dentro de la estructura global" (Moya Santoyo 2002: 45).

En cuanto al aprendizaje, destacaron el aspecto creativo: la percepción es subjetiva. Subrayaron la importancia de la configuración global y describieron las reglas básicas de la percepción de los objetos (figura-fondo, semejanza, proximidad, cercanía, buena continuidad, etc.).

Uno de los principios básicos para la Psicología de la Gestalt es el de "pregnancia"; la tendencia a percibir los objetos como totalidades bien estructuradas y de la forma más simple. Este fenómeno se explica gracias a otros cuatro principios que son procesos internos: cierre, proximidad, continuidad y semejanza.

Nos interesa, sobre todo, el principio de cierre, que inspiró a Taylor para la creación de los *clozes.* Según esta regla las estructuras cerradas se perciben más fácilmente como unidades, tendemos a cerrar las configuraciones incompletas y a recordar como cerrado aquello que no lo está totalmente. Según Khöler (1972: 19) "los procesos responsables de la formación de objetos visuales tienden a formar figuras cerradas y no simples figuras lineales".

Este principio, que observamos claramente en la percepción visual, se puede aplicar también a otros campos. De manera semejante, según Taylor, tendemos a completar un texto mutilado o incompleto, porque constituye un todo o unidad. Los elementos de un texto se interrelacionan y adquieren su pleno significado en la

estructura global. La redundancia de la lengua y la gramática de expectativas nos ayudan a completar la estructura, es decir, el todo.

5.4. Los *clozes* como expresión de los principios de pregnancia y cierre

Las pruebas de cierre proponen al alumno la tarea de completar un texto previamente mutilado. Taylor destacó la similitud entre los principios que rigen la percepción, según la Psicología de la Gestalt, y la tendencia a percibir el texto como una totalidad. A continuación veremos qué mecanismos se ponen en funcionamiento para que el hablante de una lengua consiga "rellenar" los huecos que distorsionan su percepción global del texto.

La técnica de cierre parte del reconocimiento de la lengua como sistema funcional y creativo que contiene abundantes redundancias (Read 1982 citado en Weir 1990; Spolsky 1973).

Según Oller (1979: 344), este procedimiento mide la interiorización de los conocimientos gramaticales[49], ya que para predecir una palabra de un texto debemos utilizar las habilidades que subyacen a la actuación lingüística y demostrar así nuestro grado de competencia: "in fact the cloze technique elicits information concerning the efficiency of the little understood grammatical processes that the learner performs when restoring missing or mutilated portions of text".

Cuando pedimos a un estudiante que recupere una palabra omitida en un texto, estamos haciendo que ponga en práctica esas habilidades de que dispone y que subyacen a su actuación lingüística. Las posibilidades en cada punto del texto son limitadas. Para encontrarlas cuenta con información; como las claves textuales y contextuales, ha de utilizar la redundancia de la lengua y el sistema de expectativas (*expectancy system*) que tiene como hablante. Para completar un texto correctamente hay que entender el texto y el contexto extralingüístico. Es necesario tanto utilizar las claves y limitaciones sintácticas, morfológicas y semánticas que impone el sistema de la lengua, como inferir información del contexto extralingüístico. Read (2000: 55) menciona a Sternberg y Powell (1983), que

[49] Los apartados 1.6 y 1.7 del capítulo 1 desarrollan el concepto de gramática pragmática de expectativas de Oller (1979) y el principio de redundancia reducida de Spolsky (1973).

clasifican el contexto de una palabra en interno y externo. El buen conocimiento de la lengua supondrá una gran ayuda en este proceso. Fotos (1991: 315) explica:

> The principle of reduced redundancy used in Information Theory is also thought to be involved, because the cloze test reduces natural linguistic redundancies and requires the test taker to rely upon organizational constraints to fill in the blanks and infer meaning.

Taylor (1953: 418ss.) expresaba ya las ideas de gramática de expectativas y redundancia de la lengua, posteriormente desarrolladas por Oller (1979) y Spolsky (1973) respectivamente:

> Some words are more likely than others to appear in certain patterns or sequences. “Merry Christmas” is a more probable combination than “Merry birthday”.
>
> “Man coming” means the same as “A man is coming this way now”. The latter, which is more like ordinary English, is redundant; it indicates the singular number of the subject three times (by “a”, “man”, “is”) [...] Such repetitions of meaning, such internal ties between words, make it possible to replace “is”, “this”, “way”, or ”now”, should any of them be missed.

A continuación mostramos un ejemplo práctico del proceso que se sigue para completar las omisiones de un texto, teniendo en cuenta las limitaciones (*constraints*) que impone el propio sistema de la lengua (sintácticas, morfológicas, semánticas), y utilizando todo tipo de información lingüística y extralingüística.

> Judith Taylor talks about her life as a top model.
> I am sure that _____(1) people think that the _____(2) of a model is _____(3) easy and very exciting. _____(4) is true that I _____(5) to some fantastic places _____(6) I meet some interesting _____(7). And the clothes, of _____(8)... I love wearing beautiful _____(9)!

Éste es el aspecto que presenta una prueba de cierre tradicional, de ratio fija. El alumno recibe un texto, adecuado a su nivel de competencia en la lengua, con un determinado número de huecos u omisiones. Para recuperar el texto original debe rellenar esos huecos. Todo texto contiene una serie de claves que el alumno ha de localizar y utilizar para completar su tarea de forma satisfactoria.

En primer lugar ha de tener en cuenta el tipo de texto (género, tema, contexto, etc.), puesto que cada género muestra determinadas convenciones y rasgos estilísticos. En este caso, el texto es sencillamente el relato del estilo de vida de una modelo, contado por ella misma. Sólo esta información inicial nos da una idea del registro de lengua, del tipo de vocabulario y estructuras gramaticales que probablemente aparecerán en el texto (serán términos relacionados con la vida diaria, frente a otras posibilidades que encontraríamos en textos científicos, legales, literarios, etc.). Así comenzará a aplicar su gramática pragmática de expectativas.

Además, ha de valorar las limitaciones morfosintácticas y las claves que aporta la redundancia de la lengua. Se percatará de que probablemente la primera palabra omitida (1. *most*) sea un adjetivo o un determinante debido a su situación en la oración, delante de un nombre, mientras que la segunda (2. *lifestyle*) debería ser un nombre, puesto que va seguida por un complemento del nombre "*of a model*" y precedida por el determinante "*the*". La quinta omisión (5. *go*) ha de ser un verbo de movimiento precediendo a la preposición "*to*", la octava se reconocerá como parte de la expresión habitual "*of course*", y así sucesivamente.

En algunas ocasiones un hueco puede completarse correctamente con más de una palabra (por ejemplo, también sería posible completar la omisión 2 con "*life*"). Es necesario hacer inferencias extralingüísticas (contexto, cultura, etc.) para buscar la palabra más adecuada y si varias lo son, queda a juicio del corrector determinar la validez del término según el criterio de corrección que haya fijado para la prueba[50].

Mostramos ahora el texto original que corresponde al ejemplo analizado:

> Judith Taylor talks about her life as a top model.
> I am sure that most people think that the lifestyle of a model is very easy and very exciting. It is true that I go to some fantastic places and I meet some interesting people. And the clothes, of course... I love wearing beautiful clothes!

Aportar información lingüística que no aparece en los mensajes es una actividad normal y relativamente fácil en nuestra lengua materna, pero la dificultad

[50] En el apartado 5.12 revisaremos los distintos criterios de corrección para las pruebas de cierre tradicionales: desde el que considera válido todo término que se ajuste a los límites que impone el texto hasta el que sólo admite el que aparece en el texto original. El C-test reduce considerablemente las posibilidades de que exista tal disparidad de criterios.

aumenta al intentar recuperar el mensaje en una segunda lengua. A mayor dominio del lenguaje corresponderán mejores resultados, puesto que se será capaz de utilizar mejor la redundancia de la lengua (Spolsky 1973) y se localizarán mejor las claves lingüísticas y extralingüísticas del texto.

5.5. Qué miden las pruebas de cierre

Lee (1985), Fotos (1991) y Connelly (1997), entre otros autores, se hacen eco de la controversia acerca de qué miden realmente las pruebas de cierre. Como hemos visto se considera una medida integradora y pragmática que rápidamente atrajo la atención de los investigadores (Oller 1979; Alderson 1979).

En la literatura, las opiniones acerca de qué miden los *clozes* van desde los que consideran que miden lo mismo que las pruebas de elementos discretos (Farhady 1979), o que sólo miden las destrezas básicas (Alderson 1979), hasta los que ven en las pruebas de cierre una medida de la competencia lingüística global (Oller 1979, 1988; Chavez-Oller *et al.* 1985; Bachman 1982).

Fotos (1991: 332) sugiere que todas estas opiniones son correctas, pero incompletas. Introduce un nuevo punto de vista: el nivel de los estudiantes sobre los que se aplique la prueba. Según sus investigaciones, cuanto mayor sea el nivel del alumnado en la lengua "the cloze test has more language proficiency to measure".

Oller (1979) cita diversos estudios que intentaban determinar la sensibilidad de los *clozes* a las limitaciones que impone el texto (*textual constraints*) (Aborn *et al*. 1959; MacGinitie 1961; Darnell 1963; Coleman y Miller 1967) con resultados contradictorios.

También en la literatura más cercana a nuestros días encontramos resultados dispares. Algunas investigaciones (Alderson 1979; Klein-Braley 1983) consideran que las pruebas de cierre sólo miden la comprensión del texto más cercano a las omisiones y no son sensibles a otras limitaciones lingüísticas más allá de la oración. Sin embargo, el estudio de Chihara *et al*. (1977) con omisiones en textos cuyas oraciones habían sido previamente desordenadas, muestra que "items that proved to be maximally sensitive to discourse were simply those that involved meanings that

were expressed over larger segments of discourse" (citado en Oller 1979: 361), y que estos ítems no se limitaban a las palabras con contenido léxico.

Jonz (1991) y Chihara *et al*. (1994) de nuevo estudiaron los efectos de desordenar las oraciones del texto en las pruebas de cierre. Los resultados indicaron que para resolver correctamente un *cloze* se necesita información que va más allá de la oración.

Jonz (1990: 62) considera que "the standard fixed-ratio cloze procedure has a high level of sensitivity to intersentential ties and lexical selections", en consonancia con otras investigaciones tanto anteriores como posteriores (Bachman 1982, 1985; Brown 1983; Chavez-Oller *et al.* 1985; Sasaki 2000). El autor no duda en asegurar que "the cloze procedure produces tests that are generally *consistent* in the ways they measure the language knowledge of examinees" (op. cit.: 61) (la cursiva es mía). Como hemos visto, la consistencia en la medida indica fiabilidad de la prueba. Jonz (1990) también apunta una idea importante respecto a la diferencia entre hablantes nativos y aprendices de lengua extranjera al resolver pruebas de cierre: "Future studies should hypothesize that the constraints on native-speaker cloze responses would vary from those on nonnative-speaker responses" (op. cit.: 73).

Según Jonz (1990: 63), la investigación de Bachman (1982) demostraba que la puntuación obtenida en un *cloze* refleja "complex skills ranging along a hierachy of lower- to higher-order human language processing capacities". Este estudio (Bachman 1982) concluye que las pruebas de cierre reflejan un factor de competencia lingüística general "along with three specific traits: a syntactic (clause level) trait, a cohesive (interclausal and intersentential) trait and a strategic (semantic) trait".

Jonz (1990: 72) achaca a las diferencias individuales las discrepancias entre algunos resultados de su investigación y la de Bachman (1982), porque indican que "the constraints on response for any cloze item might, in fact, vary in princpled ways from one person to the next". Este aspecto requiere, en palabras del propio autor, "careful investigation".

Storey (1997) retomó esta línea de investigación, centrándose en las estrategias que cada sujeto emplea en la resolución de las pruebas de cierre.

También las investigaciones de Chavez-Oller *et al*. (1985) destacaron este aspecto: la resolución de una prueba de cierre implica la puesta en funcionamiento de los mecanismos de "higher-order language processing".

El título de la tesis plantea una cuestión acerca del C-test que ya se ha abordado en la literatura con respecto a las pruebas de cierre en general, pero sin resultados concluyentes hasta la fecha. En concreto, nos planteamos si el C-test supone una alternativa a otras pruebas en la evaluación del Inglés como Lengua Extranjera o si únicamente debe utilizarse como complemento de las mismas.

En cuanto a las pruebas de cierre tradicionales en la literatura encontramos estudios que respaldan ambas posturas. Heilenman (1983) recomienda usar los *clozes*, con cautela, como complemento a otros métodos en pruebas de nivel. Sin embargo el éxito del estudio de Shohamy (1983) con estudiantes hebreos respalda su uso en lugar de otras medidas de la competencia lingüística global.

Fotos (1991) se cuestionó si las pruebas de cierre pueden sustituir a otros instrumentos de evaluación más tradicionales, en su caso los ensayos.

La investigación desarrollada por Fotos (1991: 334), como la de Shohamy, recomienda el uso de pruebas de cierre de ratio fija en sustitución de otras pruebas integradoras. Aunque señala algunas limitaciones de la técnica, sentencia que "carefully constructed cloze tests have the potential to become useful tools of integrative language assessment in the EFL situation".

Poco después, Soyoung Lee (1996) retoma las investigaciones acerca de la validez concurrente de las pruebas de cierre de ratio fija y su correlación con los ensayos. Entre otros aspectos señala los problemas que se derivan del ensayo como instrumento de evaluación de la competencia lingüística. A pesar de ser un procedimiento integrador, el ensayo plantea controversias en cuanto al método de corrección (holístico o analítico), el sesgo del tema "writing proficiency may vary with topic", la necesidad de contar con más de un ensayo de cada sujeto para asegurar fiabilidad, etc. (Arthur 1979; Henning 1987; Amengual 2003).

A la vista de los resultados, Lee (1996: 62) se decanta por las pruebas de cierre como alternativa a otras pruebas: "This result confirms the finding of two previous studies (Fotos 1991; Hanania and Shikhani 1986) that cloze tests can be an *alternative* to essay tests" (la cursiva es mía). Además apoya su uso en la práctica docente: "as a teaching device in classroom situations".

Nuestro trabajo experimental tiene como objetivo principal dar respuesta a la pregunta, ciñéndose al C-test: por sus características como instrumento de medida de la competencia lingüística, ¿podría ser el C-test una alternativa a otras pruebas o simplemente un complemento para las mismas?

Hemos visto que los antecedentes en la literatura no muestran resultados claros al respecto. Se hace necesaria una investigación que determine la validez y fiabilidad de la prueba mediante el análisis de su naturaleza y el estudio de las correlaciones con otras pruebas estandarizadas que midan el mismo constructo.

5.6. Las pruebas de cierre como medida de la comprensión lectora

Oller (1979) menciona distintas aplicaciones de la técnica de cierre. Además de evaluar la competencia global en la lengua, las pruebas de cierre pueden medir la efectividad de la enseñanza y servir a fines investigadores. En este apartado destacamos su utilización para determinar la legibilidad de los textos y para estimar la comprensión lectora.

A lo largo de la historia se han desarrollado muchas fórmulas distintas para calcular la legibilidad de los textos (Klare 1976). Las más utilizadas se basan en la longitud media (número de palabras) de la oración[51]. Pero con ellas no se ha logrado la exactitud que se buscaba, según Glazer (1974: 405 citada en Oller 1979: 348), porque “all language elements can, in some way, be involved in the reading comprehension process”. La autora añade que el número de palabras de una oración no es del todo significativo. En algunos casos una frase más larga puede resultar más fácil de comprender, si comunica mejor una idea (Pearson 1974).

Otros métodos intentan juzgar la legibilidad de los textos con apreciaciones subjetivas. No obstante, Oller (1979: 349) considera que los resultados son sólo estimativos, porque puede haber gran disparidad entre las apreciaciones:

[51] Son las fórmulas de Dale-Chall (1948) y Flesh (1948). La de Dale-Chall tiene en cuenta además el número de palabras no familiares, la de Flesh analiza también el número de afijos y el de referencias personales.

> There is some evidence that subjective judgements of sentence complexities, word frequencies, and overall passage difficulties may have some validity. However, in order to attain necessary levels of reliability, many judges are required, and even the, the judgements are not very precise.

Ante este panorama, Taylor (1953) propuso la técnica de cierre como base para medir el grado de legibilidad de la prosa y como medida indicativa de la comprensión lectora. Taylor trabajó de forma experimental con pruebas de cierre creadas a partir de textos previamente estudiados por los alumnos y constató una mejora en los resultados.

5.7. Rasgos fundamentales de las pruebas de cierre

En el capítulo 3 hemos revisado las características, es decir, los requisitos mínimos que debería tener toda prueba de evaluación de la lengua para ser considerada un instrumento eficaz de medida.

Validez y fiabilidad se consideran las cualidades básicas. Pero no se pueden olvidar otros aspectos, como el carácter práctico e interactivo, la autenticidad y el impacto (Bachman y Palmer 1996).

Las pruebas de cierre no son una excepción. Si queremos demostrar que cumplen el objetivo para el cual son diseñadas, hemos de valorar si reúnen estas características. En concreto, veremos los rasgos de validez, fiabilidad y factibilidad.

5.7.1. Validez y fiabilidad

A pesar de la popularidad de los *clozes*, su validez y fiabilidad han sido cuestionadas. Como hemos visto, son dos rasgos esenciales, por tanto es de vital importancia determinar si las pruebas de cierre los cumplen y en qué grado.

En general, podemos decir que la literatura respalda a las pruebas de cierre como instrumento de medida integrador de la competencia lingüística en EFL (Oller 1979; Jonz 1990; Fotos 1991).

En cuanto a su validez, se ha estudiado la validez de constructo, de contenido, concurrente y predictiva de la prueba (Bachman 1982; Brown 1983, 1988). El estudio

de Lee (1985) sobre la validez de constructo de los *clozes* analiza la validez de cada una de las omisiones. Es decir, si cada una de ellas mide un solo constructo. Según Oller (1979: 347) el hecho de que las pruebas de cierre sean sensibles a "discourse level constraints as well as structural constraints within sentences" es lo que genera coeficientes de validez mayores que los obtenidos con otras pruebas pragmáticas de evaluación.

El trabajo de Lee (1996: 69), en la línea de Fotos (1991) y Hanania y Shikhani (1986), confirma la validez concurrente de los *clozes* como medida de la expresión escrita: "The common integrative nature between the essay and cloze tests is proved".

Destacamos las implicaciones prácticas de la investigación de Lee: una vez confirmada la validez de las pruebas de cierre cabe plantearse utilizar la técnica en el aula como "an effective teaching device" más que como mero formato de examen. Hinofotis (1987) y Buck (1988) sugieren, por ejemplo, su uso en ejercicios de comprensión oral, e indican que cada profesor puede hacer las variaciones pertinentes para adaptar las pruebas de cierre a sus propias necesidades educativas en el aula.

Chapelle y Abraham (1990: 139) investigaron la fiabilidad de los distintos tipos de prueba de cierre en un estudio que volveremos a mencionar más adelante. Comprobaron que, aunque algunas técnicas resultan más fáciles que otras, no hay diferencias notables que afecten a la fiabilidad: "Although statistically significant, these differences in difficulty were not sufficiently large to affect reliability substantially. The reliabilities were adequate, although not as high as desired".

5.7.2. Factibilidad

La factibilidad es uno de los aspectos en que encontramos mayor unanimidad en la literatura sobre los *clozes* (Dörnyei y Katona 1992; Connelly 1997). Los autores reconocen y alaban su economía de esfuerzo y tiempo. Quizá esta cualidad haya sido básica para impulsar su gran popularidad en el contexto de EFL.

El carácter práctico de las pruebas de cierre radica en su fácil diseño (Oller 1979). Su creación implica escasas decisiones subjetivas por parte del profesor.

Únicamente debe seleccionar el texto base, decidir el tipo de prueba (ratio-fija, variable, C-test, etc.), el punto de comienzo de las omisiones y fijar el criterio de corrección. El profesor gana tiempo en la creación y en la corrección de la prueba, e incluso en la aplicación, puesto que es aplicable a muchos sujetos a la vez y su administración no requiere un tiempo excesivo.

Debido a sus características, se puede aplicar pruebas de cierre repetidamente en las clases, no sólo como pruebas de evaluación, sino también como instrumento didáctico en el aprendizaje de la lengua extranjera (Lee 1996). El profesor determinará según su contexto y propósito cómo utilizar las pruebas de cierre en la enseñanza de lenguas.

5.8. Selección de textos para crear pruebas de cierre

La selección de textos constituye uno de los primeros pasos en la creación de pruebas de cierre. Los expertos insisten en que se haga cuidadosamente (Fotos 1991).

Oller (1979) asegura que *a priori* cualquier texto puede ser adecuado para crear una prueba de cierre, siempre que tenga la extensión suficiente para efectuar las omisiones[52]. Por supuesto, se ha de tener en cuenta el nivel del alumno en la lengua y el propósito de la prueba. El autor apela al sentido común del profesor. Aconseja evitar los textos que puedan distraer al alumno de su tarea, es decir, aquellos que "involve topics that are intrinsically disturbing or so emotionally charged that they would distract the attention of the students from the main problem set by the test –filling in the blanks" (Oller 1979: 365). Tampoco recomienda la elección de textos que requieran conocimientos técnicos, ni los que contengan temas polémicos (política, religión, aborto, etc.). Estos mismos parámetros sirven también para el diseño de C-tests.

Sasaki (2000: 108) investigó cómo afectan a la resolución de las pruebas de cierre los esquemas de contenido activados por palabras que resultan familiares al examinando. Demostró que la inclusión de términos familiares redunda en una

[52] No es conveniente diseñar *clozes* sobre oraciones aisladas: "Cloze items over isolated sentences, of course, do not qualify as pragmatic tests at all and are not recommended" (Oller 1979: 366).

mayor motivación para resolver la prueba. Además facilita al alumno la comprensión del texto y le ayuda a encontrar las claves para completar los ítems. A partir de su análisis llegó a conclusiones útiles para la selección de textos sobre los que crear pruebas de cierre. Recomienda elegir los que sean "familiar enough for the intended examinees to fully use their knowledge".

5.9. Tipos de pruebas de cierre

Oller (1979: 344) define la técnica de cierre o *cloze procedure* y aprecia su contribución a la comprensión de los procesos internos del aprendizaje de la lengua*:*

> ...the family of techniques for systematically distorting portions of test –is a method for testing the learner's internalized system of grammatical knowledge. [...] elicits information concerning the efficiency of the little understood grammatical processes that the learner performs when restoring missing or mutilated portions of text.

A lo largo del tiempo las pruebas creadas con la técnica de cierre han sufrido modificaciones. Las aportaciones de distintos investigadores interesados en la técnica han dado lugar a varios tipos de *cloze*. Alderson (1979: 225) concluyó, tras su investigación con omisiones de distinta ratio, que "the cloze procedure is not a unitary technique", puesto que si se introducen variaciones da lugar a pruebas bien distintas.

Todas las propuestas basadas en la técnica de cierre pretenden mejorar la validez, fiabilidad y el carácter práctico de las pruebas introduciendo variaciones en el tipo o frecuencia de las omisiones (sistemática cada *n* palabras, racional, etc.) y estudiando cómo afecta el sistema de corrección aplicado (palabra aceptable o exacta, etc.). El campo de las pruebas de cierre, a pesar del volumen de lo ya investigado, sigue estando hoy abierto a nuevas investigaciones.

En el apartado siguiente clasificaremos las pruebas de cierre atendiendo al tipo y frecuencia de las omisiones. Nos centraremos en el C-test como variación que pretende superar algunas deficiencias de las pruebas de cierre tradicionales.

5.9.1. De ratio fija

Se considera el método estándar (Oller 1979; Chapelle y Abraham 1990; Jonz 1990), es el más utilizado e investigado. Los exámenes se construyen a partir de un texto previamente seleccionado, en el que se omiten de forma sistemática una cada *n* palabras. Normalmente el número máximo de omisiones es de 50 (Alderson 1979; Brown 1983; Jonz 1990; Fotos 1991). Las pruebas de cierre resultantes de aplicar una ratio fija a las omisiones se han considerado pruebas pragmáticas muy adecuadas para medir la competencia en una lengua extranjera:

> The number of words correctly replaced (by the exact-word scoring procedure) or the number of contextually appropriate words supplied (by the contextually appropriate scoring method) is a kind of overall index of the subject's ability to process the prose in the text. (Oller 1979: 345)

Algunos autores lo respaldan incondicionalmente (Oller 1979; Heilenman 1983; Hinofotis 1987; Laesch y Van Kleeck 1987) por su validez y fiabilidad.

No obstante, el método de ratio fija también ha suscitado dudas, pues algunos autores (Alderson 1979; Klein-Braley 1983; Brown 1988) critican la variabilidad e inconsistencia de los clozes. Reclaman que dependiendo del punto de partida de las omisiones y de su frecuencia un mismo texto se puede dar lugar a tests de muy distinto grado de dificultad. Oller (1979), por el contrario, insiste en que no afecta dónde comiencen las omisiones: "it matters little where the counting begins" (op. cit.: 365). Y Bachman (1982) confirma la consistencia de las distintas pruebas de cierre creadas a partir de un mismo texto.

El profesor ha de decidir la frecuencia de las omisiones. Pero conviene tener en cuenta que si se omite más de la quinta parte de las palabras de un texto (1/5th) muchos de los ítems no pueden ser recuperados, ni siquiera por hablantes nativos (Oller 1979: 345). El modelo de prueba de cierre presentado como ejemplo en el apartado 5.3 ha sido creado omitiendo una de cada cinco palabras del texto a partir de un punto concreto, que sirve como introducción al tema del texto o *lead-in*[53].

[53] Algunos autores son partidarios de dejar la primera frase intacta y comenzar las omisiones en la segunda frase. También se puede dejar la última frase sin mutilar. Este procedimiento es el que recomiendan Klein-Braley y Raatz (1991) para el diseño de C-tests. Sin embargo, en el caso de las pruebas de cierre estándar de ratio fija, Klare *et al.* (1972) y Oller (1979) no lo consideran necesario, aunque tampoco perjudicial.

Por otra parte, al seguir una ratio fija, las omisiones pueden no ser representativas del texto. Además la técnica plantea problemas de corrección que veremos más adelante (si cualquier palabra aceptable en un punto concreto del mensaje debe ser considerada válida en la corrección o exclusivamente la correcta). Otro aspecto significativo es que en este tipo de *cloze* los hablantes nativos obtienen unos resultados muy dispares, no consiguen fácilmente recuperar el 100% de las omisiones.

5.9.2. De ratio variable

Esta variación consiste en crear pruebas de cierre controladas seleccionando las palabras que se van a omitir según un criterio dado, por ejemplo, limitar las omisiones a palabras con carga semántica, o sólo a términos funcionales (Bachman 1985; Fotos 1991), dependiendo del propósito de la prueba.

Weir llama a esta técnica *selective deletion gap filling* (en Hughes 1989: 66) aunque la denominación más habitual es *rational deletion cloze tests.* Chapelle y Abraham (1990: 124) exponen la fundamentación de la técnica:

> Rational cloze research and practice rests on the assumption that different cloze items can be explicitly chosen to measure different language traits. Some evidence indicates that test writers can select words reflecting distinct aspects of the learners' grammatical and textual competence (Bachman, 1982), or at least differing in difficulty in a regular fashion (Bachman, 1985).

Más adelante, señalan las ventajas atribuidas a las pruebas de cierre de ratio variable. Chapelle y Abraham (op. cit.: 124) destacan principalmente la mayor fiabilidad y mejor correlación con otras pruebas:

> …practically speaking, items selected by experienced text writers may produce tests that are more reliable and more highly correlated with other language tests, especially test measuring traits similar to those particular cloze items were chosen to measure.

Sin embargo, el estudio comparativo de Bachman (1985) entre *clozes* de ratio fija y variable no encontró diferencias significativas entre los resultados de ambos.

Klein-Braley (1997: 62) señala que este tipo de prueba "does enable the test constructor to determine what exactly is being tested", pero también dirige nuestra atención hacia un problema importante: con este sistema de omisiones se pierde el azar, y por tanto se deja de lado uno de los principios de las pruebas de redundancia reducida fijados por Spolsky (1973): "it is not in agreement with the theory of reduced redundancy testing. If the tester chooses what to test, then the random sampling model has been abandoned".

5.9.3. De elección múltiple

Algunos autores, como Chapelle y Abraham (1990), clasifican a las pruebas de elección múltiple como un tipo de prueba de cierre.

Este método fue creado por Jonz (1976). Propone al alumno un texto con cierto número de omisiones. Para cada omisión se aportan varias opciones o posibilidades con que completar el texto. Una de ellas es la correcta, y las otras se denominan *distractors*. El alumno debe identificar la respuesta correcta.

Jonz destacó la economía de administración y corrección de la prueba. Pero Klein-Braley (1997: 60) hace la misma crítica que al formato anterior "this test form can no longer claim to rely on the principle of random sampling". Por ello, de nuevo cuestiona su validez como prueba de redundancia reducida.

Por otra parte, distintas investigaciones muestran un dato que resulta casi obvio: aportar una respuesta propia es más difícil que identificarla (reconocimiento vs. producción). A lo que Shohamy (1984) añade que la facilidad de una prueba no tiene que ver con su validez y fiabilidad.

Weir (1988: 47) señala que este tipo de prueba logra ser fiable al asegurar la objetividad del corrector, pero, sin embargo, las pruebas con este formato deben ser creadas con cuidado para que la pregunta sea clara y para que el alumno no consiga la respuesta correcta simplemente por "eliminación" o "deducción":

> There is considerable doubt about their validity as measures of language ability [...] In a multiple-choice test the distractors present choices that otherwise might not have been thought of [...] What the test constructor has inferred as the correct answer might not be what other readers infer, or necessarily be explicit in the text.

En consecuencia, la elaboración de estas pruebas conlleva mucho tiempo y esfuerzo (Klein-Braley 1997).

En cuanto a su fiabilidad, los estudios aportan resultados bien distintos: Manning (1986) encontró un coeficiente de 0.80, mientras que el estudio de Klein-Braley (1997) muestra un coeficiente de fiabilidad KR-21 de sólo 0.49.

Para Weir (1988), las preguntas de elección múltiple no son un tipo de cloze, sino un método diferente de crear exámenes. En su línea, no pensamos que sea adecuado clasificar al test de elección múltiple como tipo de cloze porque la tarea propuesta no es sino el reconocimiento o identificación de la respuesta correcta, y por tanto, no mide la competencia en la lengua objeto de estudio. El hecho de que el alumno reconozca y comprenda la palabra no significa que sea capaz de utilizarla en su producción oral o escrita. No obstante, este tipo de prueba puede ser útil en ciertos momentos o situaciones que el profesor ha de determinar.

Bachman (1990: 48) recuerda que las pruebas de elección múltiple fueron rechazadas al principio, después ampliamente utilizadas y, luego, de nuevo criticadas: “and once again multiple-choice tests are being criticized as artificial and inappropriate, even though there are many situations in which the multiple-choice format is the most appropriate available”.

La popularidad de este tipo de prueba hace que a menudo se identifique la idea de prueba objetiva o test con las pruebas de elección múltiple, como refleja el Diccionario de uso del español de María Moliner (2000).

5.9.4. *Cloze-elide technique*

Es otra variación de las pruebas de cierre introducida por Manning (1986, citado en Fotos 1991). Consiste en insertar palabras incorrectas en un texto, que deben ser detectadas por los alumnos.

Como en el caso de la elección múltiple, creemos que la técnica no puede ser considerada como prueba de cierre. En primer lugar, porque no presenta al alumno un texto en el que se omite información, sino que ésta ha sido reemplazada por un error. Además, al igual que la de elección múltiple solo mide el reconocimiento, en este caso del error, pero no la capacidad para producir una respuesta correcta.

Futuras investigaciones sobre ambas técnicas llegarán a determinar qué miden exactamente, su validez y fiabilidad.

5.9.5. C-test

En los años 80 Klein-Braley y Raatz propusieron un nuevo tipo de prueba de cierre, el C-test. En él las omisiones siguen la *rule of two*, es decir, se elimina la segunda mitad de cada segunda palabra. La tarea consiste en recuperar el texto original completando la segunda mitad de una palabra sí y otra no. Al comienzo y al final del texto se mantiene una parte intacta. Como otras pruebas de cierre, el C-test pretende medir la competencia lingüística global, mediante la redundancia reducida y la aplicación de la gramática de expectativas del hablante.

Con esta innovación, Klein-Braley y Raatz pretendían dar a los clozes más objetividad y fiabilidad. Aunque veremos después con mayor detalle las ventajas atribuibles a este tipo de prueba de cierre, podemos adelantar que con el C-test se logra, efectivamente, una mayor objetividad en la corrección, ya que limita más las opciones de respuesta para cada omisión. Es casi imposible que varias palabras sean válidas en el mismo lugar y coincida su primera mitad, como ocurre en el cloze tradicional. Debido al elevado número de ítems que el alumno debe completar, las omisiones son siempre representativas del texto y los hablantes nativos adultos llegan a obtener puntuaciones perfectas (el 100%).

Sus creadores destacan ciertas aportaciones del C-test:

- Facilidad de diseño y corrección.
- Adaptación a todo tipo de textos, temas y niveles de dificultad.
- Obtención de resultados válidos y fiables incluso con materiales que no se han trabajado previamente.
- Formato atractivo al alumno: validez aparente.

Retomaremos el análisis del C-test con mayor profundidad en el siguiente capítulo. Mostraremos las características de su diseño y las investigaciones más importantes relacionadas con la prueba.

5.10. Criterios de corrección de las pruebas de cierre

Oller (1979: 367) hace notar que con todos los criterios de corrección de pruebas de cierre investigados se ha constatado la obtención de buenos resultados. Cuando el criterio de corrección es más o menos riguroso lo que se produce es un cambio en la media (*mean score*), pero "...there is little change in the relative rank order of scores", pues se mantienen las puntuaciones relativas de los alumnos con respecto al grupo.

Brown (1980) examina la validez y fiabilidad de las pruebas de cierre en relación con el método de corrección. Valora los cuatro métodos de corrección que revisamos en los apartados siguientes: palabra correcta, palabra aceptable, *clozentrophy* y formato de elección múltiple. Al igual que Oller expone que los resultados obtenidos con todos los métodos presentan una alta correlación, por lo tanto, concluye que la elección del método de corrección debe quedar a elección del propio examinador, dependiendo de la situación.

Sin embargo otros estudios (Heilenman 1983; Hinofotis 1987; Laesch y Van Kleeck 1987) atribuyen mayor validez y fiabilidad al método de la palabra exacta.

5.10.1. Palabra exacta

Una de los posibles criterios de corrección para los *clozes* es considerar como válidas exclusivamente las respuestas que se correspondan directamente con la palabra exacta del texto original. Es también el criterio más estricto y el que menos problemas plantea al corrector, puesto que su tarea se limita a verificar la exactitud incluso ortográfica del término.

5.10.2. Palabra aceptable

Otros autores (Jonz 1991) respaldan un criterio más razonable y flexible, pero que da mayor cabida a la subjetividad. Consideran correctas las respuestas aceptables para cada omisión y contexto. Proponen el *textually appropriate scoring*

criterion. Pero puede surgir el problema de determinar cuáles son concretamente las palabras que se consideran “aceptables” para cada hueco y los límites de la aceptabilidad. Correspondería al profesor o al equipo examinador, en su caso, fijar tales términos como tarea previa a la administración de la prueba.

Por las características de su formato, con el C-test se impone la aplicación del criterio corrección de la palabra exacta, puesto que es muy difícil que sean válidas varias palabras en un punto concreto del texto, que su primera mitad coincida y que la segunda mitad tenga el mismo número de letras.

5.10.3. *Clozentrophy*

Con este criterio se valoran las respuestas comparándolas con las de los hablantes nativos. Este método añade la tarea de administrar previamente cada prueba a un número de hablantes nativos y listar las posibles respuestas. Como hemos indicado para el criterio de la palabra aceptable, la prueba pierde objetividad, aunque quizá gane autenticidad.

5.10.4. Elección múltiple

Si el examinador aporta un repertorio de posibles respuestas para cada omisión la prueba se convierte en un test de elección múltiple. El alumno sólo debe reconocer la respuesta correcta y señalarla. Su tarea se limita al reconocimiento y no a la producción.

Como hemos reflejado en el apartado 5.9.3, consideramos que no es adecuado aplicar este criterio de corrección a las pruebas de cierre. La razón es que las convierte en otro tipo de prueba, cuya validez y fiabilidad habría que determinar, pero, en todo caso, distinta en su naturaleza.

Siguiendo las indicaciones de Oller (1979) y Brown (1980) es cada profesor quien debe valorar las posibilidades existentes y elegir un criterio de corrección de clozes adecuado para su contexto y objetivos.

CAPÍTULO 6. EL C-TEST

6.1. Introducción

Este capítulo se centra en el C-test como variación que intenta mejorar las características técnicas de las pruebas de cierre. En el capítulo anterior hicimos una descripción somera de la técnica. En el actual revisamos sus antecedentes y concretamos los detalles de su diseño siguiendo los parámetros de sus creadores, Klein-Braley y Raatz. Mostramos también un ejemplo práctico y analizamos las ventajas y desventajas que se le han achacado en la literatura.

6.2. Antecedentes del C-test

Klein-Braley y Raatz (1981) fueron los creadores de este nuevo tipo de prueba de cierre. Podríamos decir que el C-test surgió como desarrollo de las pruebas de cierre, aunque Klein-Braley (1984: 97) prefiere considerarlo "a different and more satisfactory operationalisation of the construct".

Así pues, el C-test nació a partir de los *clozes* (Taylor 1953), con la pretensión de superar sus puntos débiles y los problemas técnicos que diversas investigaciones habían puesto de manifiesto (Klein-Braley 1981; Alderson 1978, 1979, 1980, 1983) y de encontrar un instrumento más consistente de medida.

Incluso la denominación elegida por Klein-Braley y Raatz indica la fuerte relación entre el C-test y los *clozes*:

> The C in the name C-Test was chosen specifically as an abbreviation of the word "cloze" in order to indicate the relationship between the two test procedures**.** The

> C-Test was an attempt to retain the positive aspects of cloze tests but to remedy their technical defects. (Klein-Braley 1997: 63)

Klein-Braley y Raatz presentaron el *C-principle* por primera vez en 1981 en el Fourth International Language Symposium de Colchester. Las primeras investigaciones se hicieron con C-tests en lengua inglesa y alemana, y pronto suscitaron gran interés entre los profesionales del sector. Después, los propios creadores diseñaron y aplicaron C-tests dirigidos a alumnos de todas las edades (desde siete años hasta adultos) y tipos (nativos, aprendices de segunda lengua, de lengua extranjera).

6.3. Deficiencias de las pruebas de cierre tradicionales

En el capítulo anterior, dedicado a las pruebas de cierre, constatamos la popularidad alcanzada por las pruebas de cierre tradicionales y las ventajas (alta validez, fiabilidad, alta correlación con otras pruebas) que encuentran en ellas destacados autores, tales como Oller (1979, 1988), Bachman (1982), Chavez-Oller *et al.* (1985), Jonz (1990) y Fotos (1991).

También mencionamos que, a pesar de su éxito, algunos aspectos de la técnica fueron criticados. Partiendo de las investigaciones de Klein-Braley (1981) y Alderson (1978, 1979, 1980, 1983) sobre las pruebas de cierre tradicionales, Klein-Braley y Raatz (1981, 1984) observaron en ellas ciertas deficiencias técnicas (*shortcomings*) que les llevaron a desarrollar un nuevo diseño: el C-test.

Haremos una enumeración de algunos de los “defectos” detectados en las pruebas de cierre y puestos de manifiesto por los trabajos de Klein-Braley y Raatz (1984) y Klein-Braley (1997), entre otros:

- Aunque los *clozes* pretenden mutilar el texto aleatoriamente, no lo logran omitiendo una cada *n* palabras.
- El grado de dificultad de las pruebas resultantes a partir de un mismo texto al aplicar ratios distintas y/o variando el punto de comienzo de las omisiones no es equivalente.

- En las pruebas de cierre existentes hasta entonces, si se quiere omitir un buen número de palabras se hace necesario que el texto de partida sea excesivamente largo.
- Al utilizar un único texto no se puede asegurar que éste sea representativo de la lengua, y no es extraño que el tema produzca sesgos.
- En cuanto al método de corrección, se aprecia que con el de la palabra exacta se crean pruebas demasiado difíciles incluso para hablantes nativos, mientras que con el de la palabra aceptable se pierde objetividad.
- Los hablantes nativos no consiguen buenos resultados en las pruebas de cierre, cuando sería lógico que los solucionaran sin problemas.
- A esto hay que añadir que los estudios estadísticos sobre validez y fiabilidad de las pruebas de cierre no son concluyentes, sino que aportan resultados dispares.

Cuando se plantearon la creación de otro tipo de prueba, en primer lugar, Klein-Braley y Raatz establecieron los criterios que debería cumplir la nueva técnica (Raatz 1985; Klein-Braley 1997):

- Debía producir pruebas más breves y a la vez incluir un número suficiente de ítems (al menos 100).
- Para evitar problemas de subjetividad la técnica debía fijar la ratio, el punto de comienzo de las omisiones y utilizar sólo el criterio de corrección de la palabra exacta.
- Para no favorecer a los alumnos que conocieran bien el tema del texto se deberían utilizar textos variados.
- Además pretendían asegurar que las palabras omitidas fueran representativas del texto y que los hablantes nativos adultos obtuvieran puntuaciones casi perfectas.

Y todo ello, sin sacrificar las consabidas validez, fiabilidad y carácter práctico propias de las pruebas de cierre.

6.4. Descripción de la técnica para diseñar C-tests

Como prueba de cierre, el diseño del C-test se basa en un texto previamente mutilado que el alumno debe recuperar. En el caso de las pruebas de cierre tradicionales las omisiones corresponden a palabras completas. Klein-Braley (1997: 64) considera que la técnica de las omisiones es el motivo de que los *clozes* no funcionen satisfactoriamente, porque "it is not sampling in the way that the theory demands".

Sin embargo, en el C-test se omite la segunda mitad de cada segunda palabra, a partir de la segunda palabra de la segunda oración del texto, siguiendo una regla que Klein-Braley y Raatz denominan *rule of two*, la "regla del dos". Es lo que se conoce como *C-Principle*: "Deletions are not performed at the text level but at the word level: we no longer remove whole words from the text; we damage parts of words" (Klein-Braley 1997: 64).

Las normas para crear C-tests son muy claras: Si la palabra tiene un número impar de letras se ha de omitir la segunda mitad más una letra. Si la palabra sólo tiene una letra (como *I* y *a* en lengua inglesa) no se tiene en cuenta en el recuento. Tampoco se consideran las cifras ni los nombres propios.

La primera oración del texto se mantiene intacta, pues sirve para introducir el tema del texto (en las pruebas de cierre tradicionales esta práctica no se considera necesaria, aunque tampoco perjudicial). Las omisiones comienzan en la segunda palabra de la segunda oración. De este modo se evita la ambigüedad de los *clozes* tradicionales para decidir el punto de inicio de las omisiones. También la última oración del texto queda intacta.

Sus creadores recomiendan que se utilicen varios textos (de 4 a 6 distintos) para construir una prueba C-test. Lo normal es que el total de omisiones del C-test sea de 100 y que se distribuyan en cuatro textos de 25 omisiones cada uno (o bien cinco textos con 20 omisiones). El profesor ha de ordenar los textos por orden creciente de dificultad, y puede hacerlo simplemente de forma intuitiva.

Klein-Braley y Raatz aconsejan que se comience por seleccionar un número mayor de textos y después de probar su funcionamiento se elijan los cuatro o cinco definitivos para cada C-test. También Brown (1993: 112) propone este proceso para las pruebas de cierre tradicionales.

El examinando debe recuperar exactamente el texto original completando la segunda mitad de una palabra sí y otra no. El criterio de corrección tampoco ofrece dudas; será siempre el de la palabra exacta. Así disminuye el margen de subjetividad del corrector y aumenta para el profesor la economía de esfuerzo y tiempo que caracteriza a la prueba.

A continuación mostramos un ejemplo de creación de C-test a partir de un texto sobre el que aplicamos las reglas que hemos explicado anteriormente.

UFOS

Any object or light reportedly sighted in the sky and which cannot be immediately explained by the observer automatically receives the label Unidentified Flying Object, or UFO. Sightings of unusual flying phenomena date back to ancient times , but UFOs (sometimes called flying saucers) became a favourite dinner table topic after the first widely publicized US sighting in 1947. Many thousands of such observations have since been reported world-wide.

At least 90% of UFO sightings are easily explained. Objects often mistaken for UFOs include bright planets and stars, aircraft, balloons, kites, aerial flares, peculiar clouds, meteors and satellites. The remaining sightings can probably be attributed to other mistaken sightings or to inaccurate reporting, hoaxes or delusions.

Éste es el aspecto que ofrece un C-test cuando se presenta al alumno:

UFOS

Any object or light reportedly sighted in the sky and which cannot be immediately explained by the observer automatically receives the label Unidentified Flying Object, or UFO. Sightings o_____ unusual fly_____ phenomena da_____ back t_____ ancient ti____, but UFOs -some_____ called fly_____ saucers- bec_____ a favourite din_____ table to_____ after t_____ first wid_____ publicized US sigh_____ in 1947. Ma_____ thousands o_____ such observ_____ have si_____ been repo_____ world-wide.

A_____ least 90% o_____ UFO sightings a_____ easily expl_____. Objects of_____ mistaken f_____ UFOs include bri_____ planets and stars, aircraft, balloons, kites, aerial flares, peculiar clouds, meteors and satellites. The remaining sightings can probably be attributed to other mistaken sightings or to inaccurate reporting, hoaxes or delusions.

Algunos autores (Weir 1988; Bradshaw 1990; Jafarpur 1995) rechazan su aspecto “fragmentario” porque resta validez aparente a la prueba. Otros (Klein-Braley 1997) reclaman lo contrario o al menos no consideran que este rasgo sea tan significativo. En el apartado 6.7.1.1 retomaremos algunos aspectos que hacen referencia a la validez aparente del C-test.

La primera oración intacta permite al alumno familiarizarse con el tema del texto (Feldman y Stemmer 1987). La parte final también contribuye a dar sentido y pregnancia al "todo" que constituye el texto.

La parte omitida de cada palabra se sustituye por un guión que no indica exactamente el número de letras restantes. Pero el alumno recibe instrucciones claras sobre la tarea que debe realizar, sabe que se ha omitido la segunda mitad de la palabra.

Como podemos ver, la elevada frecuencia de las omisiones (una palabra sí y otra no) evita que sea necesario utilizar textos muy largos para obtener una buena muestra de la actuación del alumno, lo que supone una ventaja añadida.

De este modo, Klein-Braley y Raatz consiguieron inventar un nuevo tipo de prueba de cierre que supera algunas deficiencias de las tradicionales, sin renunciar al principio de redundancia reducida. Esa certeza les llevó a considerar al C-test "técnicamente superior" a los *clozes* tradicionales (Klein-Braley 1985: 76).

6.5. Aportación del C-test a los *clozes*

Ya hemos visto cómo funciona el *C-principle* en la creación de C-tests. Klein-Braley (1997) expone los puntos en que el C-test supera a las pruebas de cierre tradicionales:

- Permite crear pruebas con más omisiones en textos más breves. A mayor número de omisiones, mayor representatividad del texto.
- El método de corrección es más objetivo. Se considera válida la palabra exacta y en pocos casos coincide más de una posibilidad que se ajuste a cada omisión y al contexto. Así se ahorra tiempo y esfuerzo en la corrección.
- Los hablantes nativos los resuelven con facilidad. Por el contrario, los que no conocen la lengua no pueden obtener resultados positivos en un C-test.
- Al constar de varios textos diferentes el C-test da cabida a mayor diversidad de contenido. Reduce las ventajas de que podría gozar un alumno experto en un tema concreto.

También Connelly (1997) reconoce las ventajas del C-test sobre los *clozes* y las clasifica en dos tipos: técnicas y de carácter práctico. Según el autor, las ventajas técnicas hacen referencia a la validez y fiabilidad de la prueba. Diversos estudios mostraron la superioridad técnica del C-test frente a las pruebas de cierre tradicionales (Klein-Braley 1985; Chapelle y Abraham 1990; Dörnyei y Katona 1992). Las de carácter práctico son casi evidentes; es una prueba de fácil diseño y corrección. Es destacable su economía de esfuerzo y tiempo. En la práctica docente se agradece contar con instrumentos de evaluación tan económicos.

6.6. El C-test como prueba de redundancia reducida

Sus creadores reconocen al C-test como prueba de redundancia reducida[54]. En el capítulo anterior hemos visto que Spolsky (1968, 1973) describe este principio y lo utiliza en pruebas de evaluación de la lengua como medio para que el alumno refleje su competencia en la lengua. Además, el principio de redundancia reducida justifica el uso de pruebas integradoras y pragmáticas, frente a las de elementos discretos:

> When one considers all the interferences that occur when natural language is used for communication, it is clear that only a redundant system would work. [...] The assessment of proficiency in a language must rather be based on functioning in a much more linguistically complex situation than is provided by the one-element test. (Spolsky 1973: 168ss.)

Las pruebas de cierre son el ejemplo más popular de prueba de redundancia reducida porque cumplen el requisito que fijaba Spolsky (1973: 175): "(they) test a subject's ability to function with a second language when noise is added or when portions of a test are masked". A continuación vemos cómo define Klein-Braley (1997: 49) las pruebas de redundancia reducida, siguiendo la teoría de Spolsky:

> A test of reduced redundancy aims at obtaining a random sample of the examinee's performance. Noise is deliberately introduced into the channel. The way in which examinees perform under these conditions is believed to provide evidence for their language proficiency as a whole.

[54] Además de las pruebas de cierre, el principio de redundancia reducida se hace operativo en otras pruebas que enumera Klein-Braley (1997: 50), tales como el dictado (Oller 1971), el Noise Test (Spolsky 1971), Partial Dictation (Johansson 1973), etc.

El alumno debe superar los "ruidos" que aparezcan en el canal y completar el mensaje utilizando todos los medios a su alcance: las claves lingüísticas y contextuales (gramática de expectativas) que le permitan inferir hasta llegar a la recuperación del texto original. Y para ello utiliza diversas estrategias (véase apartado 6.8.1).

El C-test, como prueba de cierre, se considera una prueba de redundancia reducida que pretende medir la competencia lingüística global. Klein-Braley (1997: 52) explica que la técnica de omisiones del C-test proporciona una amplia muestra de la actuación del alumno y así supera a los *clozes* típicos:

> C-tests systematically damage approximately one quarter of the text, using this substitution for random deletion on the assumption that the deletions are sufficiently "dense" to catch a fairly large sample of the examinee's processing procedures and strategies**.**

Klein-Braley (1997) realizó un estudio comparativo entre varias pruebas de redundancia reducida (dos *clozes* de ratio fija, dos de elección múltiple, dos *cloze-elide*, un dictado y un C-test). Pretendía demostrar el buen funcionamiento del C-test. Para ello asignó puntuaciones a las pruebas atendiendo a criterios de validez, fiabilidad, facilidad de diseño y corrección, etc. Y confirmó la superioridad de la prueba; efectivamente el C-test "shows superior performance over the other test procedures in the categories difficulty level, reliability, validity, factorial validity" y, por tanto, "[it] is *the best representative of the reduced redundancy tests* for general language proficiency for this problem group" (op. cit.: 71) (el énfasis es mío).

Las investigaciones llevadas a cabo por Babaii y Ansary (2001) de nuevo se plantearon si el C-test es una realización válida del principio de redundancia reducida. Los resultados confirman la anterior afirmación de Klein-Braley, los autores (op. cit.: 216) concluyen que el C-test: "conforms well to the principle of reduced redundancy which fundamentally emphasizes that both a global and a local knowledge are required to supply the missing elements in a distorted linguistic message".

Antes de finalizar este apartado queremos insistir de nuevo en la estrecha relación entre el concepto de redundancia reducida y el de gramática de

expectativas (Oller 1976, 1979)[55], expresado así por Feldman y Stemmer (1987: 255): “Closely linked to the concept of redundancy is Oller’s (1976) pragmatic expectancy grammmar”.

El hablante nativo es capaz de utilizar sin ningún problema la redundancia natural de un texto y la gramática de expectativas. Sin embargo, los que aprenden una segunda lengua o lengua extranjera se enfrentan a muchos problemas de comprensión. Las palabras de Feldman y Stemmer (1987: 255) reflejan que cuanto mayor sea el nivel de competencia en la lengua extranjera, el alumno mostrará mayor capacidad para utilizar la redundancia de la lengua:

> [...] the more clues the learners are able to pick up, because of the natural redundancy of a text, and the more they are able to make use of their pragmatic expectancy grammar, the more developed is their foreign language competence and the better they will accomplish the task.

6.7. Rasgos del C-test

En el capítulo 3 identificamos validez y fiabilidad como características básicas de las pruebas. Es, por tanto, fundamental garantizar que el diseño de una prueba posee estos rasgos para que pueda ser considerada instrumento adecuado de evaluación.

Además de las investigaciones llevadas a cabo por el equipo de Klein-Braley, Raatz y Süssmilch, desde su aparición en el panorama de la evaluación de la lengua, periódicamente surgen nuevos estudios sobre el C-test que intentan determinar su validez como instrumento de evaluación de la competencia lingüística global.

El C-test es una prueba todavía relativamente reciente, que, como veremos, ha obtenido resultados contradictorios, por tanto sigue siendo un campo abierto a la investigación en Lingüística Aplicada.

[55] Véase el capítulo 5, apartado 5.4, sobre los *clozes* como expresión de los principios de pregnancia y cierre.

6.7.1. Validez y fiabilidad

Klein-Braley y Raatz pretendían conseguir un diseño que superara algunos puntos débiles de las pruebas de cierre tradicionales, pero sin perder validez, fiabilidad, ni factibilidad.

Raatz (1985) insiste en la importancia de la validez de las pruebas. Dejando aparte la validez aparente, que merece un subapartado en este capítulo, Raatz distingue entre validez pragmática, de constructo y de contenido. Recordamos que una prueba tiene validez pragmática si realmente funciona en la situación para la que se crea y aplica. La validez de constructo viene dada por la teoría que la sustenta. La de contenido se muestra si la prueba es auténtica.

En el caso del C-test, en general, las investigaciones respaldan la validez pragmática y de contenido, pero se han cuestionado la validez de constructo y aparente, como queda reflejado en el apartado 6.8.1.

Comenzaremos con los estudios que validan y respaldan la técnica.

Klein-Braley (1985: 101) estudió la validez de constructo del C-test mediante la aplicación de C-tests a distintos tipos de alumnos (de edades y niveles de competencia variados) y demostró que: “C-Tests are authentic tests of the construct of general language proficiency”.

Años más tarde, el análisis comparativo del comportamiento del C-test frente a otras pruebas también de redundancia reducida llevó a Klein-Braley (1997: 69) a corroborar sus expectativas y concluyó: ”These results show that the improvement is genuine”. El C-test obtuvo buenos resultados estadísticos en el análisis de validez, fiabilidad[56] y, sobre todo, en factibilidad: “According to these usability criteria the C-test has the best overall ranking”.

Los resultados de 1997 coinciden básicamente con los del estudio comparativo anterior realizado por Chapelle y Abraham (1990). En este caso se crearon distintas pruebas de cierre a partir del mismo texto (*clozes* de ratio fija, variable o selectiva,

[56] Klein-Braley (1997) calculó los coeficientes de fiabilidad KR-21 de las distintas pruebas, (a pesar de que no son estadísticamente independientes y los índices tienden a sobrevalorarse). El dictado resultó ser el procedimiento más fiable, seguido de la prueba DELTA y el C-test (.85), por encima de los *clozes* (.66) y la prueba de elección múltiple (.55). En cuanto al grado de dificultad, la media del C-test fue *P*=.52. Los *clozes* resultaron los más difíciles (*P*=.27) y la de elección múltiple la más fácil (*P*=.70).

de elección múltiple y C-tests). El análisis mostró que “The C-test [...] produced, on average, the highest correlations with the language tests” (op. cit.: 140).

Ikeguchi (1998) también obtuvo resultados satisfactorios en su investigación de la fiabilidad y validez de la prueba con estudiantes universitarios de Inglés en Japón.

En 1992, Dörnyei y Katona desarrollaron un nuevo estudio de validación del C-test entre estudiantes húngaros de enseñanza secundaria y de la Eötvös University de Budapest. Reafirmaron la superioridad del C-test frente a los *clozes* (Dörnyei y Katona 1992: 187) y manifestaron que su investigación “confirmed that the C-test is a reliable and valid instrument”. Los resultados fueron tan optimistas en todos los aspectos analizados que los autores (1993: 35) no dudaron en describir la técnica como “a friendly way to test language proficiency”.

Connelly (1997) obtuvo resultados semejantes en su estudio con estudiantes de ingeniería de Bangkok. Babaii y Ansary (2001: 209) aplicaron C-tests a estudiantes de ingeniería y llegaron a la misma conclusión: “with a certain degree of latitude, C-testing is a reliable and valid procedure that mirrors the reduced redundancy principle”. También los resultados de Rashid (2002) con estudiantes de dos niveles de secundaria respaldan la técnica.

En fechas más recientes, Eckes y Grotjahn (2006) se ocuparon de nuevo de la validez de constructo del C-test, en este caso en el aprendizaje de Alemán como Lengua Extranjera. Tomaron el TestDaF como criterio y confirmaron (op. cit.: 315): “Taken together, the evidence provided by our analyses lends strong support to the conjecture that C-tests are measures of general language proficiency.”

Por otra parte, Babaii y Moghaddam (2006) analizaron cómo afectan los rasgos de los textos al procesamiento que se lleva a cabo al resolver C-tests.

Sin embargo, no todas las investigaciones aportan resultados tan positivos como las anteriores. Veamos otras que cuestionan la validez del C-test.

Autores como Carroll (1987), Weir (1988), Kokkota (1988), Cleary (1988) y McBeath (1989) criticaron algunos aspectos del C-test, como su inflexibilidad. Pero es quizá Jafarpur (1995) el que realizó un estudio más profundo y con resultados más impactantes en cuanto a la validez de constructo. Encontró múltiples inconvenientes en el C-test y manifestó que no es superior a las pruebas de cierre tradicionales: “C-testing is suffering from the very same shortcomings pertaining to its prototype, the cloze procedure” (op. cit.: 209). Aunque su estudio muestra la

fiabilidad de la prueba, cuestiona la validez del C-test y, por tanto, lo desacredita totalmente.

Para determinar la validez y fiabilidad del C-test, en nuestro trabajo experimental se creó un C-test (con dos versiones) que fue administrado a alumnos españoles de segundo curso de Bachillerato, estudiantes de Inglés como Lengua Extranjera. Los resultados obtenidos fueron sometidos a un análisis estadístico y posteriormente valorados. Para estudiar las correlaciones del C-test con otras pruebas se tomó como principal referencia externa el examen de Inglés de las PAAU. En la segunda parte de la tesis, Perspectiva Empírica, se exponen los resultados obtenidos y las conclusiones que se infieren a partir de ellos.

6.7.1.1. Validez aparente

Con respecto a la validez aparente del C-test la literatura también muestra opiniones contrapuestas. Klein-Braley y Raatz (1984) afirman la validez aparente de la prueba porque los alumnos ven al C-test como instrumento "legítimo" para la evaluación de la lengua. El estudio de Klein-Braley (1997: 71) no es tan tajante, simplemente considera que los alumnos aceptan la prueba aunque no les entusiasme: "Students appear to accept the procedure as face valid, as they demonstrate by tackling the tests, even if they do not particularly enjoy doing them".

Sin embargo, otros autores (Bradshaw 1990; Weir 1988; Jafarpur 1995) señalan justo lo contrario. Weir (1988: 53) dice que la técnica es "irritating for students" y añade que "the face validity of the procedure is low". Jafarpur (1995: 209) asegura tajantemente "C-tests *do not* possess face validity". Connelly (1997: 145) matiza la rotundidad de la afirmación de Jafarpur sobre la falta de validez aparente de la prueba y la limita a "some groups of students and teachers".

A pesar de que, según Raatz (1985: 134), la actuación en un C-test es independiente de factores como la concentración y la velocidad, hemos de decir que diversos estudios posteriores (Soyoung Lee 1996; Oh 1992; Shohamy 1982) muestran que la actuación del alumno en las pruebas de cierre, incluso en sus formatos más tradicionales, se ve afectada por factores afectivos, tales como la ansiedad del estudiante. Por tanto, también en cuanto a la validez aparente hay

fuerte ligazón entre pruebas de cierre y C-test. Aunque el C-test presenta un aspecto quizá más agresivo para el alumno, debido a que la frecuencia de las omisiones acentúa su carácter fragmentario, el tipo de reacción que provoca en el examinando no es en absoluto exclusivo de este formato. En todo caso, tanto la familiarización con el diseño como la seguridad de que se trata de una prueba válida como instrumento de medida, ayuda a aliviar la ansiedad de los sujetos (Oh 1992).

En la Perspectiva Empírica de esta tesis también retomamos el tema de la validez aparente del C-test (capítulo 12), y la analizamos a partir de los resultados obtenidos en cuestionarios retrospectivos de opinión administrados a los alumnos con posterioridad a la realización del C-test.

6.7.2. Autenticidad

No podemos olvidar otra característica de las pruebas; la autenticidad. Si definimos autenticidad como “comportamiento lingüístico de la vida real” ninguna prueba de lengua, por el mero hecho de serlo, es auténtica (Klein-Braley 1985). Pero dejando aparte esta visión estricta, se pueden distinguir grados de autenticidad en las pruebas, aunque ésta sea cuestionable.

Raatz (1985: 63) apunta que sólo las pruebas integradoras pueden ser auténticas, puesto que las de elementos discretos reducen y distorsionan la realidad. Además, recuerda la importancia de este rasgo para las pruebas: “authenticity should be present in the test material”. Pero no es suficiente con los materiales, también se requiere un comportamiento auténtico en la resolución de la prueba, y en este punto aparecen los problemas.

El C-test se considera una prueba auténtica porque cumple el requisito previo (es una prueba pragmática e integradora); además, en su creación se manejan materiales auténticos (los llamados *slices of reality* procedentes de periódicos, revistas o fuentes literarias, de temas reales, etc.) o levemente adaptados al nivel del alumno. En cuanto a la tarea que ha de realizar el alumno, la que propone el C-test es relativamente común en la vida diaria: completar un mensaje distorsionado por ruidos en el canal, haciendo uso del principio de redundancia reducida y de la gramática de expectativas (Spolsky 1973; Oller 1979). Según Klein-Braley (1985: 77)

todas las pruebas de redundancia reducida son en principio “simulations of reality” puesto que “the behaviour demanded from the examinee is viewed as an approximation of linguistic behaviour needed in everyday life”. Según la autora, la autenticidad de la prueba vendrá dada por la validez de constructo: “the claim for authenticity stands and falls with the construct validation of the tests”. Como hemos visto en el apartado 6.7.1, el C-test supera también este requisito.

6.7.3. Factibilidad

Pero si hay un rasgo del C-test que no admite discusión es su marcado carácter práctico. Éste es el aspecto menos cuestionado del C-test. Incluso los más críticos con la prueba (Jafarpur 1995: 209) admiten que: “it is easy to construct and to score C-tests”. Los autores son unánimes al alabar su economía tanto en el diseño como en la corrección (Süssmilch 1984; Döryei y Katona 1992; Klein-Braley 1997; Connelly 1997; Babaii y Ansary 2001).

Por su objetividad el C-test supera a los *clozes* tradicionales (se utiliza el criterio de la palabra exacta y generalmente sólo hay una posible solución para cada palabra mutilada) y consigue facilitar la tarea de corrección del profesor: “since it takes only slightly more time than it is needed for simply reading the text. The original text becomes automated so that checking is unnecessary” (Klein-Braley 1997: 65).

A modo de conclusión, veamos la reflexión de Dörnyei y Katona (1992: 203) sobre los rasgos del C-test y en concreto sobre su carácter práctico:

> A major objective of research on language testing is to increase the cost-effectiveness of the assessment; our conclusion about the C-test is that not only is it a reliable and valid measure of general language proficiency but it is also *one of the most efficient language testing instruments* in terms of the ratio between resources invested and measurement accuracy obtained. (el énfasis es mío)

6.7.4. Efecto rebote

La literatura hace pocas alusiones al impacto del C-test en la enseñanza y el aprendizaje. Quizá sea debido a la dificultad que supone aislar el efecto rebote para

su estudio (Bailey 1996) y al elevado coste de tiempo y esfuerzo que requieren las investigaciones sobre el impacto.

Sin embargo, en el capítulo 3 destacamos la importancia del efecto de las pruebas, tanto en el micro nivel de los individuos como en el macro nivel de la sociedad o el sistema educativo (Bachman y Palmer 1996).

A pesar de la falta de estudios empíricos que respalden esta afirmación, nos atrevemos a decir que la utilización del C-test como instrumento de evaluación, por sus características, necesariamente ha de afectar en ambos niveles. Nos quedamos con el más cercano al aula de lenguas extranjeras, el de los individuos.

Lo interesante es que el efecto rebote de una prueba sea beneficioso para profesores y alumnos. Para lograrlo se recomienda la colaboración entre ambos (Bachman y Palmer 1996; Shohamy 1997) con el fin de llegar a modelos de evaluación justos[57]. Si se siguen estas premisas, con toda seguridad el C-test aportará efectos positivos a la enseñanza de lenguas extranjeras.

Desde este trabajo animamos a la realización de investigaciones al respecto.

6.8. Métodos de análisis de los procesos que subyacen a la actuación del alumno en las pruebas de evaluación de la lengua

Bachman (1990: 113) considera que la actuación del alumno en las pruebas de lengua varía según su habilidad lingüística individual y las características del método de examen: “Furthermore, the effects of different test methods themselves are likely to vary from one test taker to another”.

También se ve afectada por las características personales “...test takers’ cognitive and affective characteristics, their “real world knowledge”, and factors such as their age sex, native language, educational and socio-economic background” (ibídem), que constituyen una fuente potencial de sesgos en la evaluación. Sin embargo, puesto que los aspectos individuales no pueden ser controlados por el profesor (aunque sí analizados), dirigimos nuestra atención hacia los primeros.

[57] Entre otros aspectos mencionamos aquí la importancia de la familiarización de profesores y alumnos con la técnica (Bachman 1990).

Cohen (1984) había sido de los primeros en realizar estudios encaminados a comprender las estrategias de los alumnos y su reacción a los distintos métodos de examen. Poco después Messick (1988: 54) señaló la importancia de investigar estos procesos, tras constatar que: "individuals differ consistently in their *strategies* and styles of task performance".

Entendemos estrategia como "plan", tal y como lo describen Faerch y Kasper (1980: 60 citado en Feldman y Stemmer 1987): "a potentialy conscious plan for solving [...] a problem in reaching a particular goal". Las estrategias se caracterizan por su dinamismo y pueden interactuar unas con otras.

Resulta relativamente sencillo analizar de forma cuantitativa la actuación del alumno en una prueba teniendo en cuenta el producto final. Pero no lo es tanto si valoramos los procesos que subyacen a su actuación; lo que se denomina en la literatura *test taking processes themselves*. Grotjahn (1986) insiste en que la validación de las pruebas de evaluación de la lengua deberían incluir este tipo de análisis introspectivo cualitativo.

La mayoría de las investigaciones realizadas sobre el C-test utilizan el análisis estadístico de las correlaciones con otras pruebas como criterio de validación. Pero con este método no se obtiene información acerca de los procesos mentales del alumno al realizar la prueba. Para evitar esta carencia Grotjahn (1987) diseñó un programa de investigación que combina ambos tipos de análisis: cuantitativo y cualitativo. Recomienda los métodos cualitativos utilizados por Cohen (1984): el protocolo *thinking-aloud* y la entrevista retrospectiva. A partir de ellos se consigue la preciada información sobre el procesamiento de la información.

6.8.1. Estrategias para la resolución de C-tests: Validez de constructo

Ya en el capítulo 4, sobre la evaluación del vocabulario, incidimos en la importancia de describir las estrategias que utiliza el alumno para resolver una prueba de lengua, a pesar de no ser éste el objetivo de nuestra tesis.

Conocerlas implica desentrañar los procesos que subyacen a la actuación del sujeto. Para ello, la literatura (Cohen 1984; Grotjhan 1986, 1987; Feldmann y

Stemmer 1987) propone incluir el análisis introspectivo cualitativo en los procesos de validación.

Siguiendo estas pautas, además del análisis de la validez concurrente del C-test, utilizamos un cuestionario retrospectivo. Otros métodos (como los protocolos *think-aloud*) no pudieron ser aplicados debido al volumen de la muestra.

Feldmann y Stemmer (1987) estudiaron los procesos que tienen lugar al resolver un C-test. La prueba consiste en recuperar una información, y para ello se requiere la presencia de claves que sean estímulo para la recuperación del texto original. Según los citados autores (op. cit.: 256), el procesamiento comienza cuando el sujeto lee la primera oración del texto, que no presenta mutilaciones. En esta parte se intenta captar la estructura subyacente: “grasp the underlying schema and thus increase the redundancy of the following mutilated text at the semantic level”.

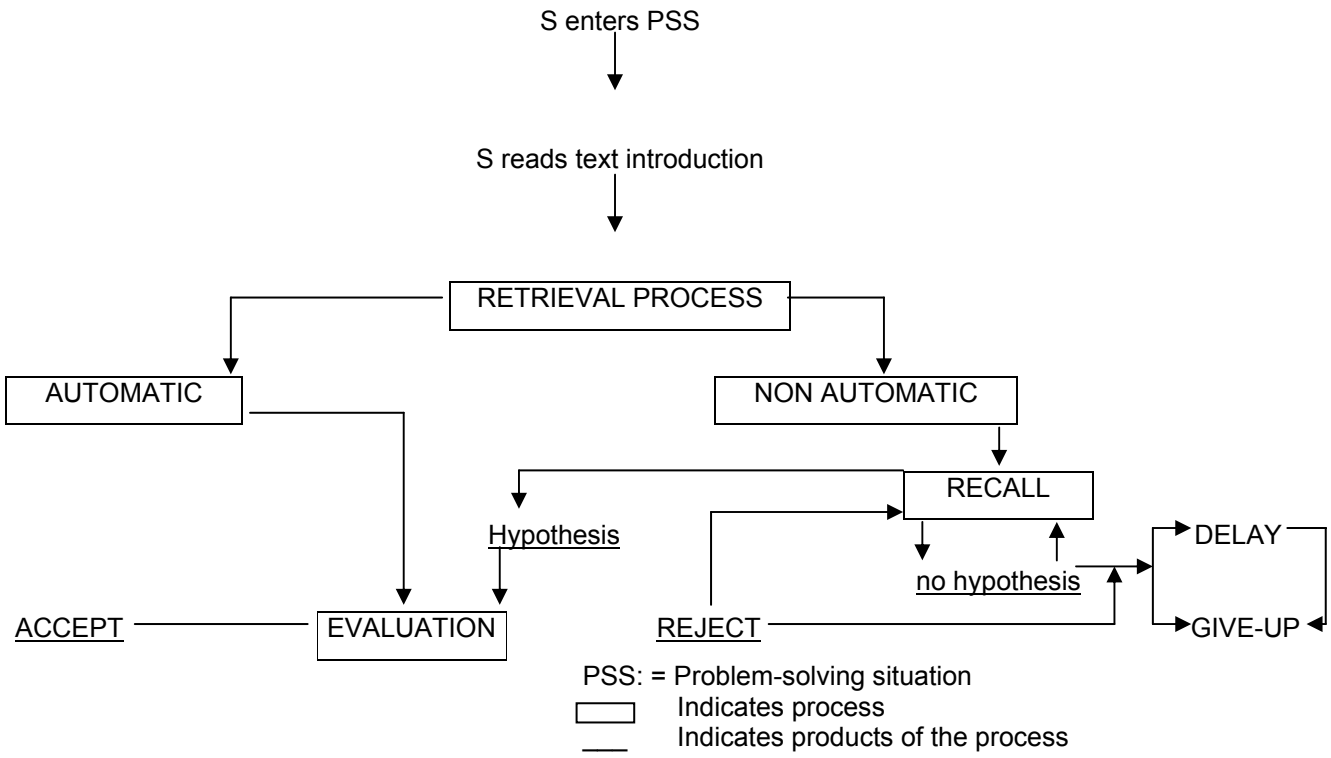
Después, el sujeto pasa a la parte mutilada; lee y relee las oraciones, trabajando generalmente de forma secuencial. Puede identificar primero las unidades más pequeñas (letras, sílabas) hasta llegar luego a las más grandes (oración, texto) en un proceso que se denomina “de abajo a arriba”, o bien a la inversa “de arriba abajo”, e incluso de forma simultánea: “A skilled reader will activate both top-down and bottom-up processing simultaneously” (op. cit.: 255).

Feldmann y Stemmer (1987) señalan que la recuperación puede ser automática o no automática. En el primer caso, la palabra surge sin pensar, mientras que en el segundo el sujeto ha de buscar otras estrategias. Una vez recuperado el término, viene la fase de evaluación, para confirmarlo o rechazarlo. Y de nuevo comienza la búsqueda, o bien llega el abandono. Cuando se duda entre dos términos, la decisión final se suele tomar “por intuición”.

Los autores identificaron toda una lista de estrategias a partir de los protocolos de los alumnos, pero es una lista abierta y susceptible de modificaciones.

El esquema que reproducimos a continuación (Fig. 9.13) resume gráficamente el comportamiento de los sujetos:

Figura 9.13. Proceso de resolución del C-test (Feldmann y Stemmer 1987: 257)



Babaii y Ansari (2001) aplicaron protocolos verbales retrospectivos en su investigación. Una vez administrado el C-test pidieron a los sujetos que verbalizaran cómo habían realizado la tarea demandada. Los resultados llevaron a los autores a identificar cuatro tipos de "claves" (*cues*) para la resolución de las omisiones:

1. Procesamiento automático (16,6%)

   Cuando al ver la primera mitad el alumno reconoce a simple vista la palabra de que se trata, por su frecuencia, etc.
2. Adyacencia léxica (54,9%)

   Cuando el sujeto se fía de las claves sintácticas y léxicas que aportan las palabras del contexto inmediato a la mutilación.
3. Claves sintácticas (22,4%)

   Si se tienen en cuenta las claves sintácticas que aportan las oraciones del texto tomado de forma global (tiempo verbal, claves referenciales, coherencia textual, etc.).

4. Otras claves externas: *Top-down cues* (6,1%)
   Como son la variable temática, el conocimiento del mundo, las características personales, *similarity chains*, etc.

A partir de estos resultados, Babaii y Ansari (2001) confirman la importancia de los conocimientos gramaticales para resolver C-tests, lo que contrasta con el anterior estudio de Dörnyei y Katona (1992: 191).

A pesar de que en la investigación que nos ocupa no cabe detenernos en el análisis de las estrategias, en nuestro estudio, sobre todo a partir de los datos del cuestionario retrospectivo, reconocemos tanto el proceso identificado por Feldmann y Stemmer (1987) como las claves descritas por Babaii y Ansary (2001).

Es probable que la mayor parte de los términos de función y alguno de los léxicos más frecuentes (e.g. “animals”) se recuperen utilizando el procesamiento automático. Junto a éste, las claves de adyacencia léxica del contexto inmediato supondrían la recuperación de gran parte de las omisiones. Sólo cuando no encuentra ayuda en el contexto inmediato el sujeto dirige su atención a las claves sintácticas y externas.

Estrechamente ligado a las estrategias está determinar qué mide exactamente la prueba y, por tanto, su validez de constructo (Feldman y Stemmer 1987; Chapelle y Abraham 1990; Dörnyei y Katona 1992; Connelly 1997; Babaii y Ansary 2001; Eckes y Grotjahn 2006). En el caso del C-test, Babaii y Ansary (2001: 216) reconocen:

> As the present investigation revealed, the C-test can in fact tap various aspects of language proficiency to varying degrees. Hence, this can be assumed to be a step forward towards establishing its construct validity. That is to say, to the extent that the C-test triggers both macro- and micro-aspects of the language, it conforms well to the principle of reduced redundancy which fundamentally emphasizes that both a global and a local knowledge are required to supply the missing element in a distorted linguistic message.

Nuestro análisis pretende simplemente señalar cuáles son algunas de las claves concretas que nuestros alumnos utilizaron para solucionar el C-test. Para ello tomamos los datos reflejados en el cuestionario retrospectivo. Las recogeremos en el capítulo 12, que analiza la validez aparente de la prueba.

6.8.2. Qué mide exactamente el C-test

Para Babaii y Ansary (2001: 212) saber qué mide exactamente el C-test es "the most controversial issue about C-testing". En términos semejantes se expresan Feldman y Stemmer (1987), Chapelle y Abraham (1990: 127), Dörnyei y Katona (1992: 188), Connelly (1997), y Eckes y Grotjahn (2006). Ciertamente, se han realizado estudios diversos con resultados contradictorios que a veces han cuestionado la legitimidad de la técnica (Jafarpur 1995). Este apartado plantea de nuevo, en definitiva, la validez de constructo de la prueba.

Raatz (1984) hizo una primera aproximación en su estudio piloto de la validez factorial del C-test con un pequeño grupo de alumnos de Enseñanza Secundaria. Estudió los coeficientes de correlación del C-test con otros tipos de prueba y con tests de inteligencia y concentración. Además de constatar la estrecha relación entre competencia lingüística e inteligencia general, afirma: "Our aim was to investigate the validity of the C-Test, and there we have been relatively successful" (op. cit.: 138).

Las investigaciones posteriores se mueven entre dos extremos: las que consideran que el C-test sólo es sensible a *low constraints* y las que opinan que mide *high constraints*.

Chapelle y Abraham (1990: 127) hicieron un estudio comparativo entre distintos tipos de *clozes*, ya que, hasta la fecha, "C-test research has failed to clarify evidence for the specific traits that this technique may measure". Parten de la idea de que para resolver un C-test se puede prescindir de las claves más alejadas de las omisiones (*long-range constraints*), pues incluso Klein-Braley (1985) había señalado que las claves utilizadas suelen estar en el entorno más inmediato de la omisión. Por eso ya al introducir su trabajo expresan: "the C-test appears to reflect more grammatical than textual competence" (ibíd.). Una vez analizados todos los datos, Chapelle y Abraham (op. cit.: 140) corroboran que "The C-test, correlating most strongly with the vocabulary test, produced, on average, the highest correlation with the language tests".

Cohen *et al*. (1984: 225) también habían encontrado que "students who did not understand the macro-context could still mobilize their vocabulary skills adequately to

fill in the appropriate discourse connector without incurring in higher-level processing".

Sin embargo, las conclusiones de Little y Singleton (1990) y las de Dörnyei y Katona (1992) contradicen a Chapelle y Abraham (1990).

Dörnyei y Katona (1992: 191) observaron la alta correlación del C-test con otras pruebas de competencia lingüística general, pero destacan que "The only area in our study where the C-test appeared to be less efficient is in the testing of grammar". Achacan este hecho a que las omisiones se producen a nivel de palabra y no de oración. No obstante, como se ha comentado previamente, el posterior trabajo de Babaii y Ansary (2001) cuestiona estos resultados.

Feldman y Stemmer (1987) también conscientes de la importancia de determinar la validez de constructo del C-test, es decir, qué mide la prueba, llevaron a cabo un estudio basado en el diseño de investigación de Grotjahn: uso de protocolos *think-aloud* y retrospectivos. Consideran que a partir de estos métodos se pueden inferir mejor los procesos cognitivos no directamente observables. Según Feldman y Stemmer (1987: 254): "Processing begins as soon as the learner starts reading the introductory part of the C-test text", y puede continuar de abajo a arriba (*bottom-up processing*) desde las unidades más pequeñas hasta las más grandes o bien en sentido contrario (*top-down processing*). Ambos modos de procesamiento pueden ser simultáneos, pero Feldman y Stemmer pretendían llegar a saber cual de ellos predomina en la resolución de C-tests.

Destacaron la importancia de la primera oración del texto, sin omisiones, para la correcta resolución de la prueba. El alumno la utiliza para captar mejor la redundancia del texto y el esquema subyacente. A partir de ese punto, según los autores, las mutilaciones se van completando de dos formas: *automatic* y *non-automatic retrieval*. La recuperación automática se produce sin dudas ni titubeos, y la no automática requiere el uso de estrategias. Feldman y Stemmer (1987: 264) describen dos tipos de estrategias: de recuperación y de evaluación[58]. Una vez identificado el término que se busca, el alumno evalúa su hipótesis y posteriormente

[58] Feldman y Stemmer (1987: 259-262) hacen una taxonomía de las estrategias. Entre las de recuperación mencionan las que se basan en la sintaxis, en añadir letras o sílabas, la repetición de ítems que han aparecido antes en el texto, la búsqueda de claves semánticas, etc. Entre las de evaluación aparecen la comprobación de significado y forma, la relectura del texto, etc. Y cuando ni el propio alumno sabe explicar porqué elige una palabra, hablan incluso de la "intuición".
Los autores insisten, no obstante, en que la lista de las estrategias es necesariamente abierta.

la acepta o rechaza. Los citados autores concluyeron que: “strategies cannot be localized unambiguously along the top-down – bottom-up continuum”.

Su trabajo es muy ilustrativo, pero los propios autores advierten de lo reducido de la muestra analizada (aplicaron sólo 10 C-tests en español y 10 en francés a alumnos alemanes estudiantes de estas lenguas) y animan a la realización de ulteriores investigaciones.

Babaii y Ansary (2001) también utilizaron protocolos verbales retrospectivos en su investigación sobre la frecuencia y tipo de claves utilizadas por los estudiantes de Inglés como Lengua Extranjera para solucionar C-tests. Su análisis les permitió identificar los cuatro tipos de claves mencionados en el apartado 6.8.1. De todas ellas, las de adyacencia léxica fueron las más utilizadas. Más adelante (capítulo 9, apartado 9.3.6), veremos que tanto las de adyacencia léxica como las interoracionales se basan en los conocimientos gramaticales del alumno.

En fechas recientes, Eckes y Grotjhan (2006) se ocuparon de la validez de constructo del C-test, esta vez en la evaluación del Alemán como Lengua Extranjera. Su investigación corrobora al C-test como instrumento que mide la competencia general en la lengua. Además, precisa: “lexis and grammar are important components of general language proficiency as measured by C-tests” (Eckes y Grotjhan 2006: 316). No obstante, los autores puntualizan que el peso de estos dos componentes (léxico y gramática) depende también del nivel de los examinandos y del grado de dificultad de la prueba.

Una de las últimas aportaciones al tema, el análisis de Babaii y Moghaddam (2006), apunta a la dificultad sintáctica y al grado de abstracción de los textos como factores que aumentan la dificultad de la prueba y obligan al sujeto a utilizar “macro-level processing” en la resolución de C-tests.

Todas las investigaciones que hemos mencionado, como el acercamiento al C-test que hacemos en esta tesis, son, en realidad, complementarias. Cada una valora un aspecto de la prueba y aporta algo a su comprensión total.

En varios momentos hemos comentado que el estudio empírico de la tesis incluye el análisis de un cuestionario retrospectivo. Los alumnos lo completaron de forma anónima una vez acabado el C-test. Se aplicó con el principal objetivo de determinar la validez aparente de la prueba. Pero en la Perspectiva Empírica veremos que aporta también otro tipo de información muy útil, nos ayuda a

comprender qué creen los alumnos que mide el C-test. A pesar de todo, en nuestro trabajo sólo se apuntan ideas. Ésta sigue siendo la gran incógnita que rodea al C-test. Sería deseable que futuras investigaciones, en la línea de Eckes y Grotjahn (2006), Babaii y Moghaddam (2006) y la que aquí presentamos, dedicaran sus esfuerzos a intentar desentrañarla.

6.8.3. *C-processing difficulty*

Lo que Klein-Braley denomina *C-processing difficulty* se relaciona directamente con el grado dificultad del C-test para los alumnos. Este aspecto se ha de tener en cuenta en la selección de textos, dentro del proceso de creación de C-tests.

En su estudio empírico para desentrañar cómo se puede predecir la dificultad de un C-test, Klein-Braley (1984) encontró también algunos aspectos interesantes que no se limitan a los textos, sino que afectan a los sujetos. Observó que la dificultad para solucionar un C-test depende en cierta medida de la edad y, por tanto, de la madurez del alumno: “C-processing difficulty decreases in a linear fashion as the L1 subjects get older” (op. cit.: 109).

El otro factor fundamental es el nivel alcanzado en la lengua objeto de estudio, desde los principiantes hasta llegar al hablante nativo, que consigue puntuaciones casi perfectas. Esta afirmación es básica para considerar al C-test como instrumento que mide la competencia global en lengua extranjera.

A la vista de los resultados Klein-Braley (1984: 111) propone una sencilla fórmula que resume su investigación:

> The observed score for any individual on a C-Test must be dependent on the one hand on the C-processing difficulty of the text and on the other on the individual’s position on the language learning continuum. A simpler formal model than the Rash model for this relationship could be an additive one:
> observed score = subject ability + text level + error.

El C-test que se aplica en la parte experimental de la tesis presenta un nivel de dificultad que se supone homogéneo y adecuado al nivel de competencia de los alumnos, pues todos los textos que lo forman proceden de exámenes de Inglés de Selectividad. La edad de los sujetos también es homogénea (cursan 2º curso de

Bachillerato). Por tanto, siguiendo las indicaciones de Klein-Braley, en principio los resultados obtenidos serán representativos de la competencia de los sujetos en lengua inglesa.

6.9. Usos del C-test

En un primer momento, Klein-Braley y Raazt mostraron abiertamente su satisfacción por el descubrimiento del C-test y sus expectativas para la técnica en el campo de la evaluación de la lengua. Parecía que podía servir casi "para todo".

Dörnyei y Katona (1992: 198) resaltan la versatilidad del C-test, que tiene "something to offer to everybody". Pero la euforia inicial dio paso a la prudencia, y Klein-Braley (1997: 72) adopta una postura más reposada que le lleva a manifestar:

> They should only be used under the supervision of the test expert who should evaluate the suitability of the C-Test in question for the specific target group in question before using the results to make any important decisions.

A pesar de las palabras de la autora, es lícito plantearse en qué situaciones y para qué propósitos resulta más adecuado el uso de las pruebas de redundancia reducida y, en concreto, el C-test. Klein-Braley considera que coinciden con los usos de cualquier prueba de competencia lingüística global, aunque el C-test aporta las ventajas de factibilidad y objetividad mencionadas en apartados anteriores.

Desde fechas muy próximas a su creación el C-test se aplicó en la práctica (Raatz 1984). En Eurozentrum (Colonia) formó parte de la prueba de nivel que se aplica a los nuevos alumnos. También en la Universidad de Duisburg el C-test ha demostrado su validez institucional como prueba de nivel.

Süssmilch (1984: 173ss.), que administró C-tests en lengua alemana a alumnos nativos e inmigrantes aprendices de alemán como lengua extranjera con resultados "muy satisfactorios", afirma que el C-test "can be used by teachers in normal classroom test procedures, but it can also be used as an aid in selection, classification and placement decisions".

Según Klein-Braley (1997) se puede utilizar como prueba de nivel al comenzar un curso, servirá para determinar el nivel general del alumno en la lengua. También

como prueba de selección para clasificar a los alumnos según su competencia lingüística. Destaca su utilidad en la toma de decisiones. Como *decision-making tests* ayudan a que el profesor decida si un alumno tiene o no el nivel suficiente para seguir un determinado programa. Y, por último, menciona su posible aportación a la investigación lingüística, puesto que se pueden aplicar a aprendices de la lengua materna, de una segunda lengua o de una lengua extranjera.

Ciertamente, en la investigación lingüística, el C-test ha demostrado ser un instrumento versátil. En el capítulo 4 comentamos su uso en las investigaciones sobre aprendizaje del vocabulario. Además, varios autores (Wolter 2002; Murtagh 2003) han utilizado la prueba como referencia externa para estudiar las correlaciones con otros tipos de examen.

Wolter (2002: 320) aplicó el C-test para comprobar las posibilidades de un *word association test* como medida la competencia lingüística en lengua extranjera. El autor justifica su elección de este modo:

> I needed to use a testing format which (1) had the ability to assess overall proficiency, (2) has been shown to be reliable and valid, and (3) can be completed in a relatively short amount of time.

Sin embargo, el C-test no resulta apropiado para diagnosticar los puntos débiles de un alumno o los aspectos en que destaca, puesto que mide la competencia "global" en la lengua. Además, Klein-Braley y Grotjhan (1995) (en Raatz y Klein-Braley 1988) advierten de que, en principio, al ser una prueba independiente del currículo, el C-test no pretende detectar los pequeños progresos en la lengua, sino más bien los logros a medio o largo plazo. No obstante, en la práctica, los profesores pueden "adaptar" la fórmula y crear C-tests relacionados con el currículo para ser aplicados de forma regular en las clases.

En la Perspectiva Empírica comprobaremos que, por sus características, el C-test puede ser muy útil en el aula de lenguas extranjeras como instrumento de evaluación, siempre que se utilice adecuadamente. Además mostraremos sus posibilidades para ser incluido en exámenes estandarizados (PAAU), dada su validez, fiabilidad y factibilidad.

6.10. Variaciones sobre la técnica del C-test

El C-test es una prueba todavía relativamente reciente y poco utilizada (en comparación con las pruebas de cierre tradicionales), que a veces ha obtenido resultados contradictorios, por tanto sigue siendo un campo abierto a la investigación en Lingüística Aplicada.

Periódicamente surgen nuevos estudios. Como hemos visto a lo largo de este capítulo, algunos intentan determinar su validez como instrumento de evaluación de la competencia lingüística global en lengua extranjera. Otros, sin embargo, sugieren modificaciones en la técnica con el fin de mejorar sus características y adaptarlo a situaciones concretas. A continuación revisamos las propuestas más interesantes.

La posibilidad de desarrollar variaciones sobre el C-test no hace sino poner de manifiesto, una vez más, la enorme versatilidad y posibilidades de la prueba.

6.10.1. La “regla del tres”

Las investigaciones de Süssmilch (1988: 173) con C-tests aplicados a alumnos cuyo nivel en la lengua era muy dispar pusieron de manifiesto las dificultades de los que tenían un nivel más bajo para resolver la prueba. Llegó a la siguiente conclusión: “extremely easy tests are needed for the early stages of L2 learning”.

Para conseguir C-tests más fáciles ideó una variación que facilitaba la tarea al reducir el número de omisiones. Su propuesta supone el abandono de la “regla del dos” a favor de lo que podríamos llamar “regla del tres”, ya que aumenta la ratio de las omisiones (n=3).

El C-test que propuso estaba formado por seis textos con un total de sesenta omisiones: “The texts were shorter and had only ten deletions which affected every third word”. De este modo logró un C-test adecuado para principiantes: “...modification of the C-principle enables the construction of suitable tests for these subjects” (ibíd.).

6.10.2. C-tests “a la medida”

La aportación de Jafarpur (1999) parte de la observación de la disparidad de los ítems de un C-test en términos de dificultad. Algunas omisiones poseen unos valores aceptables en cuanto al grado de discriminación y facilidad, pero otros son excesivamente fáciles o demasiado difíciles para el alumno.

Para evitar esto propuso la creación de *tailored C-tests* controlando las características estadísticas de cada ítem. Así pues se abandonaría la “regla del dos” y se seleccionarían las omisiones individualmente.

Ya Grotjahn (1987) y Kamimoto (1993) habían sugerido dejar las omisiones sistemáticas para mejorar la técnica. Sin embargo, los resultados de la investigación llevada a cabo por Jafarpur (1999: 83) no mostraron la esperada mejoría, como él mismo reconoce: “Taken together, the results obtained from this study indicate that tailoring does not improve the statistical characteristics of the C-test”.

6.10.3. L-Test

Kokkota (1988) propuso un nuevo procedimiento de omisión de letras (LDP) que pretende superar la escasa flexibilidad del C-test. El L-Test integra características de los *clozes* de ratio variable (Bachman 1982) y del C-test.

Según Kokkota (1988: 118): “the parameters of LDP are between those of cloze procedure and C-Test, which are its extreme modifications”, por eso el L-Test aventaja tanto a las pruebas de cierre como al C-test.

Para llegar al L-Test, Kokkota estudia la relación entre el número de letras de una palabra y su grado de dificultad. Por una parte, es consciente de que la regla del dos produce en el C-test muchos ítems excesivamente fáciles (generalmente términos funcionales). Por otra, ve que cuanto más larga es una palabra, más fácil es su recuperación (Kokkota 1986). Finalmente deduce que: “by increasing or decreasing the number of undeleted letters (NUL) in an item-word we should be able to control the rate of redundancy reduction in a text” (op. cit.: 115).

El L-Test típico es un texto de entre 250 y 350 palabras de extensión en el cual aparecen unas 60 omisiones con una distancia interítem de cuatro, cinco o seis

palabras (una serie de normas rigen la distancia entre las omisiones y el número de letras omitidas en cada palabra mutilada).

Según su creador "L-Test item difficulties tend to fall between those of cloze-tests and C-tests" (op. cit.: 116). Pero la principal ventaja que señala Kokkota es la posibilidad de que el profesor ajuste la dificultad de los ítems variando el número de letras omitidas.

Con excepción de los del propio autor no se han realizado hasta la fecha estudios consistentes que demuestren fehacientemente las ventajas del L-Test como variación del C-test. Sería necesario determinar sus rasgos, principalmente validez y fiabilidad. Sin embargo, resulta obvio que el diseño de este tipo de prueba requiere mucho más tiempo y dedicación, con lo cual su economía y factibilidad disminuyen notablemente. Quizá este sea el motivo de su escaso éxito.

6.10.4. *The Productive Vocabulary Levels Test*

Laufer y Nation (1995, 1999) desarrollan un prolífico trabajo en el campo del aprendizaje del vocabulario[59]. Conscientes de la importancia del vocabulario en la enseñanza del Inglés, han diseñado un modelo de prueba de vocabulario cuyo formato se asemeja al del C-test: el *Productive Vocabulary Levels Test*.

Parten del *Vocabulary Levels Test* de Nation (1983) para diseñar una nueva prueba de "controlled productive vocabulary ability". Esta prueba pretende medir la adquisición de vocabulario. Por lo tanto, difiere del C-test tanto en su propósito como en el formato, a pesar de su parecido. Los propios autores, Laufer y Nation (1999: 37), lo expresan así:

> The test format bears some resemblance to the C-test [...] although for vocabulary-sampling purposes in this study it is not used in a paragraph but a sentence, and the cues are not always half a word.

Utilizan como base la oración en lugar del texto. En cada oración se omite la parte final de una palabra, que el alumno debe recuperar.

[59] Véase el capítulo 4, sobre las pruebas de vocabulario.

De este modo baja considerablemente la frecuencia de las omisiones, pues sólo hay una omisión selectiva por cada oración. Como en el C-test, se aporta la primera parte de la palabra, pero no se sigue el mismo criterio (omitir la segunda mitad y en las palabras cuyo número de letras es impar, la mitad más uno). En este caso el criterio es bien distinto: "The number of letters for each word was decided on by the elimination of possible alternatives to the tested word. [...] it was thought better to provide the minimal number of letters that would disambiguate the cue" (op. cit.: 37).

Las palabras mutiladas pertenecen a cinco niveles de frecuencia, desde las palabras más frecuentes en la lengua hasta las menos utilizadas (2000, 3000, 5000, University Word List (UWL) y 10.000 word levels).

El *Productive Vocabulary Levels Test* resultó ser válido, fiable y práctico como prueba productiva de vocabulario.

A continuación vemos el aspecto que presenta la prueba:

Figura 6.1. The Productive Vocabulary Levels Test (Appendix 1, Laufer y Nation 1999: 46)

The Productive Vocabulary Levels Test: Parallel Version I (Version C)
Complete the underlined words. The example has been done for you.

He was riding a bicycle.

The 2,000-word level

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.

6.10.5. Otras propuestas

Intentando buscar soluciones razonables para el problema de los ítems que no discriminan adecuadamente, Cleary (1988) propuso cambiar el sentido de las omisiones y hacerlas en la primera parte de la palabra: "left-hand deletions". Boonsathorn (1990) y Prapphal (1994) continuaron trabajando con este tipo de

prueba, conocida como X-Test, y comparando sus rasgos de validez y fiabilidad con la del C-test. Pero los estudios al respecto resultan aún poco concluyentes

Sigott y Kobrel (1996) sugirieron incrementar la dificultad de los textos aumentando las omisiones de 1/2 de cada palabra a 2/3, o bien manteniendo sólo la primera letra. También respaldaron la propuesta de Cleary (1988) como otra opción para conseguir aumentar el grado de dificultad de la prueba.

6.11. Interpretación de los resultados obtenidos en un C-test

Klein-Braley (1988: 98), una de las creadoras de la prueba, define al C-test como prueba normativa, *norm-oriented test*. Como tal pretende tener un nivel de dificultad medio, lo que supone que la puntuación media es la recuperación del 50% de las omisiones. No obstante, añade: “If necessary we can afford to let the mean difficulty slide up to 60%”, porque sus investigaciones constatan la fiabilidad del C-test incluso cuando el nivel de la prueba no se corresponde exactamente con el del alumno, y le resulta muy difícil o fácil.

Como hemos visto en el apartado 6.4 del presente capítulo, según sus creadores, para garantizar el nivel medio de dificultad es conveniente seleccionar bien los textos. También es deseable que éstos aparezcan en orden de dificultad creciente (el primero ha de ser más fácil, un *icebreaker* que no deje duda al alumno sobre lo que la prueba pide de él). Además, la prueba se suele introducir con un modelo para familiarizar al alumno con la técnica.

Siguiendo estos parámetros se deberían obtener unos resultados coherentes. Así se hizo en la investigación empírica que justifica esta tesis. En la Perspectiva Empírica mostramos cómo se interpretaron los resultados de nuestros alumnos en el C-test atendiendo a distintas variables y criterios.

6.12. Líneas de futuro

El C-test, como prueba de nivel, forma parte de uno de los principales proyectos de UNIcert® (University Foreign Language Certification System en Alemania) que sigue el Marco Común Europeo de Referencia (Eckardt y Voss 2006) y del proyecto ALTAIR de la Universidad de Bolonia (Tamburini y Paci 2002).

Hemos de destacar la actividad que desarrolla la Universidad de Duisburg, cuna de este diseño, donde actualmente funciona un proyecto de investigación sobre el C-test: C-test Research Project. Puede consultarse en http://www.uni-duisburg.de/FB3/ANGLING/FORSCHUNG/home.html.

La técnica del C-test no es ajena a la aportación de las Nuevas Tecnologías al mundo de la evaluación. Es más, por sus características, resulta muy apropiada para su utilización con el soporte informático. Algunas instituciones ya han introducido un C-test en sus páginas web. Destacamos algunas, como la del Centro de idiomas de la Universidad de Barcelona, UAB Idiomes, que propone la práctica con C-tests: http://si.uab.es/suab244w/ada/ctests/ctests.html, o la página sobre Web-Based Language Testing http://www2.hawaii.edu/~roever/wbt.html, que resalta las ventajas del uso de ordenadores en la evaluación.

Hoy, veinticinco años después de su creación, se han diseñado y aplicado C-tests en más de veinte idiomas y múltiples contextos (véase bibliografía en http://www.c-test.de). Es una prueba fructífera, que sigue presente en la investigación sobre evaluación de la lengua (LT) y que tendrá todavía mucho que decir en el futuro.

CAPÍTULO 7. ESTUDIOS PILOTO

7.1. Introducción

Dada la naturaleza y literatura expuesta en torno al C-test (Klein-Braley y Raatz 1984; Klein-Braley 1985, 1997; Dörnyei y Katona 1992; Jafarpur 1995; Connelly 1997; Babaii y Ansary 2001; Rashid 2002; Eckes y Grotjahn 2006; Babaii y Moghaddam 2006, etc.), a lo largo de los años nos hemos ido planteando diversas cuestiones con respecto a su aplicación y funcionamiento con alumnos españoles de FP, COU y 2º de Bachillerato, que estudian Inglés como Lengua Extranjera.

Con objeto de encontrar respuestas a estas preguntas decidimos llevar el C-test a las aulas en dos estudios piloto que describimos a continuación y que fueron el germen de la Perspectiva Empírica de esta tesis.

Mackey y Gass (2005: 43) resaltan la importancia crucial de este tipo de estudios piloto (*small-scale trials*) previos a la investigación principal:

> Pilot testing is carried out to uncover any problems, and to address them before the main study is carried out. A pilot study is an important means of assessing the feasibility and usefulness of the data collection methods and making any necessary revisions before they are used with the research participants.

En este capítulo detallamos los pasos seguidos en el diseño y aplicación de dos pruebas piloto; una con alumnos de 5º de Formación Profesional y la otra con alumnos de COU.

Revisamos los resultados obtenidos, las conclusiones a las que nos condujeron estos primeros análisis y su incidencia posterior en el diseño de la investigación empírica que justifica esta tesis. Los dos trabajos que exponemos a

continuación supusieron una primera aproximación al C-test como instrumento de evaluación del Inglés como Lengua Extranjera[60].

7.2. Prueba piloto I

7.2.1. Objetivos del estudio

En 1998 comenzamos a trabajar con el C-test. Aplicamos este tipo de prueba de cierre a 25 alumnos de 5º de Formación Profesional. El diseño del estudio piloto pretendía valorar la validez, fiabilidad y eficacia de la técnica del C-test. Además apuntaba posteriores líneas de investigación, tales como:

- La incidencia del conocimiento previo de los textos,
- la recuperación de los términos léxicos y los funcionales,
- la reacción ante este tipo de prueba, para ellos totalmente novedosa.

7.2.2. Sujetos

Las pruebas se aplicaron a un grupo de 25 estudiantes españoles del IES Humanejos de Parla (Madrid). Cursaban 5º curso de FP, rama Administrativa, durante el curso escolar 1998/99, con 3 sesiones semanales de Inglés. El grupo no conocía la técnica del C-test, aunque se les informó detalladamente de las características de la prueba.

[60] Los resultados obtenidos a partir de los dos estudios piloto fueron expuestos en congresos de AESLA y, los tres últimos, posteriormente publicados:

- Esteban, M., Herrera, H. y M. Amengual (2000) Niveles de correlación entre el C-test y las pruebas de Inglés de Selectividad. Comunicación al XIX Congreso Nacional de AESLA. Universidad de León.
- Esteban, M., Herrera, H. y M. Amengual (2001) ¿Puede el C-test ser una alternativa a otras pruebas en la enseñanza del inglés como segunda lengua? *La lingüística española a finales del siglo XX. Ensayos y propuestas*, Tomo I. AESLA. Universidad de Alcalá.
- Esteban, M. y H. Herrera (2003) El C-test: instrumento apropiado para la evaluación de la competencia en inglés como lengua extranjera. *Las lenguas en un mundo global*. AESLA. Universidad de Jaén.
- Esteban, M. (2005) Niveles de correlación entre el C-test y la prueba de Inglés de Selectividad. En Herrera Soler, H. y J. García Laborda (Coord.) *Estudios y criterios para una Selectividad de calidad en el examen de Inglés.* Valencia: Ed. UPV.

7.2.3. Materiales

Para la primera aplicación de esta técnica en el aula se crearon cuatro C-tests de cincuenta omisiones cada uno, a partir de cuatro textos de un nivel de dificultad adecuado para los alumnos. Los textos versaban sobre temas tratados en las unidades de su libro de texto, *Themes for 1º Bachillerato* (ed. Burlington Books).

Elegimos dos temas: "las comidas" y "los conflictos generacionales". En el diseño se tuvo en cuenta que dos C-tests se refieren a las comidas y los otros dos al segundo tema, los conflictos generacionales. A la variable temática añadimos la de texto conocido o desconocido, al incluir en los C-tests dos textos previamente trabajados en clase frente a dos nuevos. Así, los *C-tests* 1 y 2 compartían el tema de "las comidas", pero el C-test 1 era conocido para los alumnos y el 2 totalmente nuevo. De igual forma, los C-tests 3 y 4 trataban sobre "los conflictos entre padres e hijos", siendo el 3 conocido y el 4 desconocido.

Cada texto consta de más de 100 palabras, y en 50 de ellas hay mutilación. Desaparece la segunda mitad de cada segunda palabra, exceptuando los nombres propios, las cifras y fechas. Las formas verbales aparecen completas, no contractas. Con objeto de facilitar la tarea al alumno, un guión reemplaza a cada letra eliminada. Además cada texto comienza y acaba con una parte intacta (Klein-Braley 1985).

A continuación mostramos uno de los C-tests aplicados, el resto puede consultarse en el Apéndice:

C-TEST 1: BREAKFAST AROUND THE WORLD

Breakfast is an important meal because it gives you energy to start the day. When (1) y-- do (2) n-- have (3) - good (4) break----, you (5) f--- hungry (6) a-- eat (7) ca---, biscuits (8) o- sweets (9) be---- lunchtime. (10) Th-- type (11) o- food (12) i- bad (13) f-- you (14) bec---- it (15) i- not (16) v--- nutritious (17) a-- has (18) l--- of (19) su--- and (20) f--.

Breakfast (21) i- not (22) t-- same (23) i- every (24) coun---.For(25) ex-----, many British (26) peo--- have (27) to--- or (28) ce---- and (29) - cup (30) o- tea. (31) Ot---- prefer (32) - traditional (33) break---- of (34) ba--- and (35) eg--. In (36) ot--- Northern European (37) coun-----, for (38) ex----- Germany (39) a-- Sweden, (40) peo--- eat (41) c--- meat (42) a-- cheese (43) w--- bread (44) a-- coffee. (45) I- Nigeria (46) h-- soup (47) i- very (48) co----. Many Brazilians (49) e-- different (50) trop---- fruit and cold meat for breakfast.

However, in many parts of the world, people only eat a small dish of rice for breakfast.

Podemos apreciar, sin embargo, que ni en las omisiones ni en el formato general de los C-tests aplicados se siguieron de forma estricta las indicaciones de sus creadores. Estos aspectos fue revisados en estudios posteriores.

A pesar de las indicaciones de los creadores acerca del número de omisiones (Klein-Braley y Raatz 1984; Klein-Braley 1985, 1997), en nuestro caso, la extensión de los textos elegidos permitía crear C-tests de 50 omisiones sin que resultaran excesivamente largos o tediosos. Además hemos de tener en cuenta que el tema de los textos había sido tratado en actividades diversas en la clase de Inglés. De este modo, además, podíamos administrar dos C-tests en una sola sesión, manteniendo constante la variable del tema e introduciendo la de texto conocido *versus* desconocido.

7.2.4. Procedimiento

Así pues, para realizar este estudio contamos con cuatro C-tests de 50 omisiones cada uno. Se administraron de dos en dos, así cada alumno completó en una sesión dos pruebas (C-test 1 y 2) cuyo tema era el mismo. Y en la segunda sesión los C-tests 3 y 4, que también compartían tema. En ambos casos uno de los textos había sido trabajado previamente en clase y el otro era nuevo. Los trabajados en clase formaban parte de alguna de las lecciones del libro de texto y por tanto suponíamos que el alumno estaba suficientemente familiarizado con ellos. No obstante, en ningún momento se anunció su aparición en la prueba ni el tipo de técnica que se iba a utilizar.

Las pruebas se distribuyeron en dos sesiones normales de clase de Inglés del tercer trimestre del curso escolar 97/98 y los alumnos dispusieron de 45 minutos para completar dos C-tests (un total de 100 omisiones).

En los momentos anteriores a la administración se facilitaron las instrucciones pertinentes, hasta asegurarnos de la adecuada comprensión de la tarea por parte del alumnado. Se explicó también el sistema de corrección que se iba a aplicar. El criterio elegido era claro y estricto: para cada omisión sólo la recuperación de la palabra exacta sería considerada válida.

Para comprobar la validez concurrente y la fiabilidad del C-test decidimos analizar las correlaciones de las pruebas con las calificaciones de los alumnos en las evaluaciones previas a su administración (1ª y 2ª del curso 1998/99).

Mediante el estudio estadístico comparativo de los resultados de las pruebas pretendíamos analizar la variable "conocimiento previo del texto".

Con respecto a la variable del tipo de palabra omitida (términos léxicos y funcionales), para cada C-test se hizo un listado que incluía todas las palabras mutiladas divididas en dos grupos; las funcionales y las de contenido semántico. Con el programa estadístico SPSS se analizó la recuperación de los dos tipos de palabra en cada C-test.

7.2.5. Resultados y discusión

Una vez corregidas las pruebas, todos los datos obtenidos se sometieron a análisis estadístico con el programa SSPS 8.1 para *Windows,* teniendo en cuenta las variables objeto de estudio.

En primer lugar se analizaron los promedios obtenidos en cada C-test y se estudió la correlación entre los resultados de los C-tests y las calificaciones previas de los alumnos en la asignatura de Inglés.

Los histogramas confeccionados sobre las tablas de frecuencia mostraron que el C-test 1 presentaba la media más alta: 37 respuestas correctas de un total de 50*,* incluso 3 estudiantes obtuvieron 48 aciertos. En el C-test 2 la media bajaba ligeramente, hasta 31,2 aciertos en escala de 0 a 50, y únicamente un alumno consiguió 44 puntos, máxima puntuación del grupo. Resultado que parece lógico teniendo en cuenta que los alumnos desconocían el segundo texto.

Los C-tests 3 y 4 presentaban una media muy similar, incluso algo superior en el texto nuevo: 31,7 y 31,9 puntos respectivamente, en la misma escala. En este caso, nos planteamos la incidencia de los rasgos de los textos y su grado de dificultad, puesto que aunque el texto 4 era nuevo, presentaba mayor redundancia. A partir de los datos confirmamos que la elección de los textos para crear C-tests ha de hacerse de forma muy cuidadosa.

Desde una perspectiva holística subrayamos que los histogramas presentaban unas curvas razonablemente normales. Desde una perspectiva analítica cabía apuntar que la curva obtenida a partir del C-test 1 mostraba una distribución bastante normal; con la curva ligeramente sesgada y una curtosis normal, que podríamos considerar mesocúrtica.

Por otra parte, los valores modales de la curva correspondiente al C-test 2 llamaban la atención, pues se podría considerar una curva bimodal, con uno de ellos bajo la media y el otro solapándose con la media. La razón para explicar esto podría ser que los alumnos no conocían el texto y, por tanto, les resultó más difícil. El hecho de ser una curva bimodal apuntaba a que la clase tendía a polarizarse en torno a dos grupos, al menos en este texto nuevo para ellos. Las curvas correspondientes a los C-tests 3 y 4 mostraban tendencia a la centralidad. De manera más manifiesta, el C-test 4, con la mayor parte de las puntuaciones en el intervalo medio, lo que subraya la clara tendencia a un comportamiento homogéneo. Las puntuaciones revelaron que el C-test 4 no resultó demasiado fácil ni tampoco excesivamente difícil para el grupo.

Se estudió estadísticamente la recuperación de las palabras de significado léxico y gramatical. Para agrupar las palabras de los textos en léxicas y funcionales seguimos las pautas de clasificación de Quirk y Greenbaum (1973) y Aarts y Aarts (1986), que distinguieron entre las partes del discurso abiertas y cerradas, “closed-system and open-class items” y “major and minor word classes”, respectivamente.

Tradicionalmente las palabras se han agrupado en clases o partes del discurso que comparten una serie de características, principalmente morfológicas y sintácticas, ya que el criterio semántico es menos fiable. Muchas palabras del inglés, si están aisladas, no pueden adscribirse a una clase concreta sino que presentan lo que se denomina *multiple membership*. Los lingüistas advierten de que la distinción entre partes del discurso abiertas y cerradas debe hacerse con cautela.

En general, podemos decir, no obstante, que los términos funcionales pertenecen a clases cerradas e incluyen a los artículos, demostrativos, pronombres, preposiciones, conjunciones y verbos auxiliares. Los léxicos pertenecen a clases abiertas; son nombres, adjetivos, adverbios y verbos con carga léxica.

La proporción de términos léxicos y funcionales de los cuatro textos utilizados en el estudio es la siguiente:

C-TEST 1		
Palabras afectadas por la mutilación:	De contenido léxico	24 (48 %)
	Funcionales	26 (52 %)
C-TEST 2		
Palabras afectadas por la mutilación:	De contenido léxico	27 (54 %)
	Funcionales	23 (46 %)
C-TEST 3		
Palabras afectadas por la mutilación:	De contenido léxico	18 (36 %)
	Funcionales	32 (64 %)
C-TEST 4		
Palabras afectadas por la mutilación:	De contenido léxico	22 (44 %)
	Funcionales	28 (56 %)

A las palabras pertenecientes a clases cerradas afectadas por la mutilación en el C-test 1 se les dio el nombre de FUNCT 1, a las que tenían carga semántica LEXIS 1, y así sucesivamente. El número total de cada categoría entre los cuatro tests se denominó FUNCTT y LEXIST.

Comparando la recuperación de los términos gramaticales y los léxicos vimos que los gramaticales se recuperaron con mayor facilidad, lo que coincide con otros estudios (Farhady y Keramati 1996). Pero la diferencia no era significativa en los tests estudiados.

Hemos de tener en cuenta la redundancia de los textos. Al ser textos relativamente sencillos, en ellos abundaba la repetición de palabras relevantes desde el punto de vista léxico, como nombres de comidas en los C-tests 1 y 2. Posiblemente esto facilitó su recuperación y acortó diferencias. Aún así, los términos gramaticales resultaron más fáciles de recuperar, ya que son un número limitado y con gran frecuencia de uso en la lengua (Klein-Braley 1985: 91).

Los resultados más interesantes de este estudio piloto fueron los que constatamos al analizar las correlaciones. Todos los tests comparados entre sí mostraron correlación significativa. De forma semejante ocurrió al compararlos con las calificaciones de los alumnos en la asignatura de Inglés en la 1ª y 2ª Evaluación del curso 1997/98. Entre el par C-test 1 y C-test 2 se constató la mayor correlación 0,845 (que supone un coeficiente de determinación de 0,714). Los estudiantes que

lograron una mejor puntuación en la primera prueba también consiguieron puntuar alto en la segunda.

En la Tabla 7.1, que muestra la correlación de Pearson, el doble asterisco (**) indica que ésta es significativa. Se aprecia que todas las correlaciones estudiadas fueron altas, entre el C-test 3 y las calificaciones previas se observa la mayor. La menor, entre el C-test 1 y las notas del primer trimestre, aunque ya aumenta al compararlo con el segundo.

Tabla 7.1. Correlaciones de Pearson

	1	2	3	4	5	6
1. CTEST1	--					
2. CTEST2	,845**	--				
3. CTEST3	,690**	,722**	--			
4. CTEST4	,740**	,763**	,789**	--		
5. EVAL1	,511**	,597**	,676**	,543**	--	
6. EVAL2	,670**	,577**	,676**	,520**	,814**	--

N= 25 alumnos
**La correlación es significativa al nivel 0,01 (bilateral).

Los datos pusieron de manifiesto la validez concurrente de la prueba y, en consecuencia, las enormes posibilidades del C-test como prueba de evaluación de la competencia en Lengua Extranjera.

En cuanto a la validez aparente, hemos de decir que en este primer estudio no se pasó un cuestionario al alumnado, puesto que esta valoración no era nuestro objetivo primordial. Simplemente, en el contexto del aula, se pidió de manera informal a los alumnos que expresaran libremente su opinión sobre las pruebas. Curiosamente, la mayoría manifestaron que les había gustado el tipo de examen, lo cual contrasta con algunas investigaciones sobre el tema (Bradshaw 1990; Jafarpur 1995) que reflejan lo contrario. A partir de esta apreciación se decidió que en estudios posteriores podría ser de utilidad la aplicación de un cuestionario retrospectivo, puesto que abriría la posibilidad de nuevas vías de investigación.

7.2.6. Conclusión

Desde el primer momento el C-test cumplió nuestras expectativas. Aunque con la debida cautela, ya el primer estudio piloto nos llevó a considerar al C-test como un instrumento válido y fiable para la evaluación del Inglés como Lengua Extranjera. Trabajos posteriores sirvieron para corroborar esta impresión inicial.

La alta correlación con las calificaciones previas en la asignatura de Inglés fue determinante como indicador de validez concurrente y fiabilidad, por tanto cabía incluso plantearlo como posible alternativa a otros tipos de examen.

Sin embargo, el estudio estadístico sobre la recuperación de los distintos tipos de palabra no aportó nada significativo a la investigación. En cuanto a la utilización de textos conocidos o no en la creación de C-tests los resultados tampoco fueron tan concluyentes como cabía esperar.

Por otra parte, el C-test demostró ser una prueba práctica: fácil y cómoda en cuanto al diseño, administración y corrección. Como prueba objetiva no implica más que las mínimas decisiones subjetivas, en concreto, la elección de los textos.

En cuanto a su validez aparente, en general, nos pareció que no incomodaba a los estudiantes. Sólo aquellos con peor nivel en la lengua se sintieron “perdidos” al realizarla puesto que no entendían los textos y no encontraban claves lingüísticas para recuperar el texto original. Pese a todo apreciaron su carácter objetivo. Pero llegamos a esta conclusión exclusivamente a partir de nuestras propias impresiones y del sondeo informal en el aula.

Para estudios posteriores se debía buscar un instrumento más objetivo que permitiera valorar este aspecto. Como explicaremos más adelante, en la Prueba piloto II se trabajó en el diseño de un primer modelo de cuestionario de opinión para el alumnado.

Los resultados iniciales con FP sirvieron para animarnos a profundizar en el C-test como indicador válido y fiable del grado de competencia de los alumnos españoles de COU en Inglés como Lengua Extranjera[61].

[61] El artículo de Esteban, Herrera y Amengual (2001) refleja esta investigación. Fue mencionado en la revisión de Graeme Porte (2001) sobre Evaluación de la lengua en España.

Resumiendo:

1. La alta correlación entre los resultados de los C-tests aplicados con las calificaciones obtenidas en la asignatura de Inglés evidenció la validez concurrente y fiabilidad de la prueba.
2. No se obtuvieron resultados concluyentes ni novedosos en cuanto a la recuperación de los distintos tipos de palabra (términos léxicos y funcionales) en los C-tests estudiados.
3. La variable texto conocido *versus* desconocido tampoco aportó datos significativos a la investigación.
4. Desde la perspectiva de la factibilidad, la prueba demostró su economía de tiempo y esfuerzo en diseño, administración y corrección.
5. La validez aparente del C-test no quedó suficientemente probada. Se decidió profundizar en ella utilizando los instrumentos adecuados.
6. En estudios posteriores debían revisarse cuestiones relativas al diseño de las omisiones y formato de la prueba.
7. Quedó abierta la posibilidad de continuar estudiando las aportaciones del C-test en el contexto de la evaluación de la competencia en Lenguas Extranjeras.

7.3. Prueba piloto II

Antes de plantear el diseño definitivo de la investigación principal de la tesis llevamos a cabo un segundo estudio piloto con alumnos de COU. Como veremos, se introdujeron algunos cambios para mejorar el diseño del primer estudio piloto.

7.3.1. Objetivos del estudio

En esta ocasión decidimos trabajar con alumnos cuya competencia en lengua inglesa es, en principio, superior. Seguimos profundizando en los rasgos del C-test, especialmente en su validez y fiabilidad, en la línea iniciada en el estudio anterior

(apartado 7.2). Esta vez comprobamos las correlaciones del C-test con exámenes del tipo de Selectividad. De este modo, fijamos un referente externo objetivo (Herrera Soler 1999): el examen de Inglés de las PAAU.

Abandonamos el análisis de las otras variables (familiarización con el texto y tipo de palabra), puesto que no aportaron nada significativo al primer estudio. Sin embargo, introdujimos como novedad un cuestionario retrospectivo de opinión dirigido al alumnado.

7.3.2. Sujetos

Las pruebas se aplicaron a un grupo de 21 estudiantes españoles de COU del IES San Isidoro de Sevilla, de Madrid, durante el tercer trimestre del curso 1999/00. A lo largo del curso se venían realizando exámenes tipo PAAU como preparación para la prueba real de Selectividad, a la que se enfrentan los alumnos una vez superado el COU (actualmente 2º de Bachillerato) para acceder a estudios universitarios.

Para este estudio tomamos las calificaciones obtenidas en dos de estos exámenes (realizados a mediados de abril y mayo, respectivamente), desglosados en sus distintas partes, y las comparamos con los resultados del C-test (que se aplicó a primeros de mayo).

La muestra presenta las características lógicas del trabajo con los alumnos de la propia clase; además de ser limitada en cuanto al número de sujetos, se puede dar el efecto Hawthorne y el *halo effect* (Adair 1984; Adair *et al*. 1989; Brown 1988).

El grupo no tenía entrenamiento previo en la técnica del C-test. Se dieron instrucciones claras y un ejemplo del modo de realización de la prueba. También se explicó el sistema de corrección, de nuevo sólo la recuperación de la palabra exacta se consideró correcta. Finalmente, se les pidió completar un cuestionario sobre su opinión acerca de la prueba.

7.3.3. Materiales

Los materiales utilizados en este estudio fueron los siguientes:

1. Un C-test de cien omisiones
2. Dos pruebas tipo PAAU:
   "*Parking*" (Selectividad LOGSE, Madrid. Septiembre de 1994)
   "*Get ready! The euro-day is coming*" (Modelo PAAU LOGSE, Madrid. Curso 1999-2000).
3. Un cuestionario retrospectivo

A continuación, aportamos algunos detalles de interés relativos a los materiales arriba enumerados.

Para la realización del segundo estudio diseñamos un C-test de 100 omisiones. En este caso, seguimos fielmente las instrucciones de Klein-Braley (1985) para la creación de C-tests. Nuestro C-test estaba compuesto por cuatro textos con 25 omisiones cada uno. Los textos, sobre temas de actualidad, procedían del libro de texto *Exam Strategies* (Longman). Se presentaron al alumno en orden de dificultad creciente, según la apreciación del profesor.

Mostramos la primera parte del C-test, que incluye las omisiones 1 a 25, correspondientes al primer texto. También incluimos los títulos de los textos restantes. El C-test completo aparece en el Apéndice de la tesis.

LEARN TO COMMUNICATE

To be fluent in several languages is no longer considered a rare talent, but a necessity to succeed and communicate in the world in which we now live. Many (1)peo--- believe (2)th-- once (3)y-- are (4)pa-- childhood, (5)lear---- a (6)n-- language (7)i- too (8)diff-----. This (9)i- not (10)tr--.

Whether (11)y-- want (12)t- learn English, French, Spanish (13)o- Polish there (14)a-- schools (15)a-- courses (16)gea--- for (17)yo-- needs (18)a-- specifically (19)ai--- at (20)ad--- learning. (21)Ad--- learning (22)i- pro-active; (23)y-- are (24)invo---- with (25)t-- language from the beginning and encouraged to talk, whatever your ability. There are a variety of methods available.

26-50 - The historic voyage of Christopher Columbus
51-75 - Coping with addiction
76-100 - Killing the goose...

Además, se aplicaron en el aula dos exámenes de Inglés de Selectividad: *Parking* (Selectividad LOGSE, Madrid. Septiembre de 1994) y *Get ready! The euro-day is coming* (Modelo PAAU LOGSE, Madrid. Curso 1999-2000).

A partir de los resultados se valoró la validez concurrente, fiabilidad y eficacia del C-test en relación con los exámenes de Selectividad, tomando como base las distintas correlaciones existentes entre ellos.

Después se aplicó el cuestionario retrospectivo, que se basa en el creado por Jafarpur (1995). Fue completado por los alumnos de forma anónima, con el fin de asegurar la veracidad de las opiniones expresadas acerca del C-test.

7.3.4. Procedimiento

Desde la perspectiva de la replicabilidad, uno de los rasgos básicos que define cualquier trabajo empírico, expondremos el procedimiento seguido en esta segunda prueba piloto.

Las pruebas tipo PAAU se aplicaron en dos sesiones normales de clase, a mediados de abril y mayo de 2000, respectivamente. En el periodo de tiempo comprendido entre ambas se aplicó el C-test. Los alumnos dispusieron de 50 minutos para completar cada examen. Tomamos las calificaciones obtenidas en ellos, desglosadas en sus distintas partes, y posteriormente las comparamos con los resultados del C-test. También el C-test se administró en una sesión de clase de Inglés del tercer trimestre (a primeros de mayo). Una vez completado el C-test se entregó el cuestionario retrospectivo para conocer las impresiones que este tipo de examen produce en el alumnado.

Tras la corrección, todos los datos obtenidos se sometieron a análisis estadístico utilizando el programa SPSS 9.01 para *Windows*. Así pudimos comparar los resultados obtenidos y estudiar las correlaciones.

Se subdividió el C-test de 100 ítems en 4 subtests, correspondientes a los cuatro textos en que se basa la prueba y que el alumno debía recuperar. Para su estudio se denominó CTT a los resultados globales del C-test, CT1 a las omisiones 1 a 25, CT2 de la 26 a la 50, y así sucesivamente.

En cuanto a las pruebas de Selectividad, llamamos TT1 a la puntuación total obtenida en el primer examen realizado y TT2 a la del segundo:

TT1: *Parking*

TT2: *Get ready! The euro-day is coming*

También desglosamos la puntuación total de cada examen tipo PAAU en dos partes. Por un lado, tomamos los resultados de las preguntas de tipo objetivo (gramaticales, de comprensión del texto: verdadero o falso y de vocabulario) y, por otro, las subjetivas (preguntas abiertas y redacción sobre uno de los dos temas propuestos). Cada una de las partes tiene un peso de 5 puntos sobre los 10 totales. La puntuación de la parte objetiva del primer examen aparece en las tablas como OBJ1 y la subjetiva se denomina SUB1. Al segundo examen le corresponden OBJ2 y SUB2.

En el siguiente capítulo, sobre la Metodología de la investigación, se analiza detalladamente la prueba de Inglés de las PAAU (capítulo 8, apartado 8.3.4).

7.3.5. Resultados y discusión

La Tabla 7.2 refleja las correlaciones entre el C-test y las pruebas de Selectividad:

Tabla 7.2. Correlaciones de Pearson

	1	2	3	4	5	6	7
1. CTT	--						
2. SUB1	,809**	--					
3. SUB2	,770**	,844**	--				
4. OBJ1	,590**	,651**	,775**	--			
5. OBJ2	,627**	,785**	,772**	,675**	--		
6. TT1	,792**	,944**	,893**	,865**	,808**	--	
7. TT2	,746**	,866**	,950**	,775**	,932**	,906**	--

N= 21 alumnos

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

Se puede observar que las correlaciones son bastante altas en todos los casos, aunque no todas las partes del C-test aportan iguales resultados. El C-test 1 pondera al alta. Corresponde al texto que se consideraba más sencillo y por ello encabezaba el C-test. Sin embargo, las correlaciones del C-test 3 (omisiones 50-75) son algo más bajas que las del resto, esto puede ser debido al grado de dificultad del texto 3. Como en el estudio piloto I, hay que insistir en la importancia de la adecuada selección de los textos.

Se podría hacer también un análisis por ítems. Ya hemos comentado en el capítulo anterior el hecho de que algunos autores (Jafarpur 1999) achacan al C-test la inclusión de ítems no significativos que, aparentemente, no ayudan a discriminar el nivel de los alumnos. En principio, la valoración de Jafarpur parece adecuada, pero siempre cabría trabajar sobre el rango de las puntuaciones, con una media X=0 y una desviación típica 1.

En general, son ítems demasiado fáciles, que cualquier alumno consigue recuperar, independientemente de su grado de competencia lingüística, o demasiado difíciles para el conjunto de los alumnos.

En el caso concreto del C-test 1, vemos, por ejemplo, que los ítems 1, 3 y 9 son muy sencillos, todos los alumnos los recuperan sin problema:

> Many (1)peo--- believe (2)th-- once (3)y-- are (4)pa-- childhood, (5)lear----a (6)n-- language (7)i- too (8)diff-----. This (9)i- not (10)tr--.
> Whether (11)y-- want (12)t- learn English, French, Spanish (13)o- Polish there (14)a-- schools (15)a-- courses (16)gea--- for (17)yo-- needs (18)a-- specifically (19)ai--- at (20)ad--- learning.

Sin embargo, ningún alumno consiguió recuperar los ítems 19 y 20, dos omisiones de carácter léxico difíciles de deducir en el texto. Este análisis nos informa sobre el dominio que tiene el alumno de un tipo de ítem y del otro. También permite observar qué ítems domina toda la clase y cuáles exigen mayor competencia de los alumnos y atención por parte del profesor.

En el capítulo 6 vimos que Jafarpur (1999) se planteó eliminarlos de los C-tests porque, en su opinión, no aportan nada a la prueba. Así, se obtendría un C-test racional, “a la medida”, dirigido y no natural. Sin embargo, concluyó que retirar ese tipo de ítems extremadamente fáciles o difíciles no supone ventaja alguna y no

produce variaciones en los resultados. Además afectaría al diseño de la prueba, pues supondría no seguir siempre la “regla del dos”.

En realidad, comprobamos que estos ítems discriminan lo mismo que cualquier otro test si se transforman las puntuaciones directas en típicas. En ese caso, la distribución de puntuaciones nos permitiría valorar el nivel de conocimientos de los alumnos. De esta manera, no habría lugar para las objeciones de Jafarpur. Podría entenderse que nada sobra y nada falta en los C-tests, en función de los criterios del profesor, ya que hay ocasiones en las que nos interesa motivar al máximo a nuestros alumnos, y otras en las que, por diseño se puede hacer más hincapié en los ítems difíciles.

Los dos exámenes de Selectividad (TT1 y TT2) correlacionaron muy bien entre sí (0,906), dato que indicó la fiabilidad del examen. También comprobamos que el C-test correlacionaba bien con los resultados de los exámenes del tipo de Selectividad (0,792 y 0,746). Lo más sorprendente fue que la correlación era mejor con la parte subjetiva de los mismos: 0,809 con la parte subjetiva del primer examen de selectividad (SUB1) y 0,770 con la del segundo (SUB2) frente a los resultados de la parte objetiva: OBJ1 = 0,590 y OBJ2 = 0,627. Fue éste un dato inesperado y llamativo. Decidimos que debería ser analizado con mayor profundidad en el futuro, puesto que el C-test está clasificado como prueba objetiva, pero sus niveles de correlación nos llevaron a considerarlo próximo a las pruebas subjetivas.

7.3.6. Conclusión

Esta segunda experiencia de aplicación de C-test a alumnos de COU confirmó ya al C-test como una variedad de las técnicas de cierre válida en la enseñanza del Inglés como Lengua Extranjera. Faltaba todavía definir si sólo como complemento o incluso como alternativa a otras pruebas más tradicionales.

Descubrimos en el C-test un tipo de examen que participa de ciertas características de las pruebas objetivas y de las subjetivas. No sólo mide el vocabulario que ha adquirido el alumno sino también su capacidad para inferir a partir del contexto, para reconocer los elementos gramaticales del texto y recuperarlos “re-creando” el texto de origen.

El C-test mostró correlación significativa con los dos exámenes de Selectividad realizados, más con el segundo, que se hizo un mes después. Resultado que puede deberse al efecto del aprendizaje y a la motivación que genera la proximidad de la Selectividad. Se acercaba el fin de curso y la fecha de la convocatoria oficial de las PAAU y, por lo tanto, debió aumentar el estudio. Sin duda, los alumnos habían practicado más y realizaron la prueba con mayor cuidado y atención.

El hecho curioso de que se lograran correlaciones aún más altas con la parte subjetiva de los exámenes de Selectividad (preguntas abiertas y composición) indica que el C-test es un tipo de examen más próximo a las pruebas subjetivas que a las objetivas. Es objetivo en cuanto a la preparación y corrección, pero para realizarlo requiere producción, es creativo pues va más allá del mero reconocimiento. Fue éste el resultado más significativo del estudio realizado y nos llevó a plantearnos, de cara a estudios posteriores, cuáles son los factores de este diseño que le hacen correlacionar mejor con pruebas de tipo subjetivo.

El cuestionario que completaron los alumnos pedía su opinión con respecto al C-test; si les parecía adecuado, completo, y si creían que reflejaría bien sus conocimientos de inglés. Igualmente se les preguntó por las dificultades que encontraban al realizarlo. En general, expresaron que no les gustaría que su acceso a la Universidad dependiera de un C-test, pero sí que formara parte del examen. Una vez más, la carencia de validez aparente que destaca la literatura (Weir 1988; Bradshaw 1990; Jafarpur 1995) no se refleja en este cuestionario. No obstante, fue tomado únicamente como elemento informativo que recogía las impresiones del alumnado.

A pesar de que la muestra utilizada no fue muy grande, consideramos esta segunda aproximación al C-test como el punto de partida para la investigación principal desarrollada en esta tesis[62]. La idea del cuestionario, junto con las conclusiones de estos dos trabajos piloto, sirvieron para determinar qué elementos formarían parte de ella y cuáles serían las principales líneas de investigación. Como veremos en los capítulos siguientes, en el estudio se mantuvieron las PAAU como referencia externa, se analizó el cuestionario retrospectivo, etc.

[62] El artículo “Niveles de correlación entre el C-test y las Pruebas de Inglés de Selectividad”, publicado en el volumen “Estudios y criterios para una Selectividad de Calidad” (2005) está basado en el estudio piloto II.

En resumen:

1. De nuevo aseguramos la validez concurrente del C-test, al comprobar que correlaciona significativamente con las dos pruebas de Selectividad aplicadas.
2. Los resultados obtenidos muestran que el C-test, a pesar de ser prueba objetiva por su formato y corrección, resulta más próximo a las subjetivas que a las objetivas.
3. La validez aparente de la prueba quedó reflejada en un cuestionario retrospectivo de la opinión de los alumnos.
4. Destacamos la versatilidad y factibilidad del C-test como instrumento de evaluación de la lengua.

Después de haber llevado a cabo los estudios piloto previamente comentados, y con la seguridad de haber encontrado en el C-test un instrumento de evaluación de sumo interés, continuamos las investigaciones en torno a esta prueba para averiguar si cumple las expectativas creadas, confirmar los resultados obtenidos anteriormente, comprobar cómo correlaciona con otras pruebas que incluyan aspectos objetivos y subjetivos (PAAU), y estudiar la incidencia de un posible cambio en el formato.

CAPÍTULO 8. DESCRIPCIÓN DEL PROCESO METODOLÓGICO

8.1. Introducción

Como hemos visto en el capítulo anterior, nuestro trabajo empírico tiene su origen en dos estudios piloto que revelaron el potencial del C-test como instrumento de evaluación del Inglés como Lengua Extranjera.

En este capítulo hacemos una descripción de los principales elementos que han formado parte de la investigación que informa la tesis.

Comenzamos con los sujetos participantes en el estudio, a continuación mostramos los distintos materiales utilizados en el mismo y algunas otras características del contexto de la investigación que consideramos pertinentes. Finalmente, explicamos el procedimiento utilizado y comentamos aspectos relativos al tratamiento de los datos.

8.2. Sujetos

Los participantes en este estudio fueron 162 alumnos de 2º curso de Bachillerato pertenecientes a cuatro Institutos de Enseñanza Secundaria de la Comunidad de Madrid, pero de características muy distintas, debido principalmente a su ubicación: IES Ágora de Alcobendas, IES Vicente Aleixandre de Pinto, IES San Isidoro de Madrid, e IES Humanejos de Parla.

Los centros pertenecen a tres Direcciones de Área Territorial diferentes: Madrid-Norte, Madrid-Sur y Madrid-Centro. Las pruebas se aplicaron durante el tercer trimestre del curso 2000/01.

Con respecto a las pruebas piloto, en el estudio definitivo ampliamos el ámbito de la muestra hasta 162 sujetos. Así conseguimos la validez externa que permite que los resultados sean generalizables y relevantes (Mackey y Gass 2005: 119). Hemos conseguido, además, una muestra más variada y representativa.

Todos los sujetos realizaron una prueba modelo de las PAAU, un C-test de cien omisiones (modelo A o B) y por último, de forma anónima, completaron el correspondiente cuestionario retrospectivo.

Para recoger otras informaciones sobre los sujetos, como las calificaciones en Inglés en la 2ª Evaluación del curso escolar y en el examen de Inglés de la PAAU, hemos contado con la inestimable ayuda de las profesoras de Inglés titulares de los grupos que participaron en este estudio empírico. Previamente fueron informadas de los objetivos y características del estudio que se iba a llevar a cabo (véase Apéndice). De nuevo queremos destacar su colaboración eficiente y generosa. Sin ellas no habría sido posible esta tesis.

La variedad de la procedencia de los alumnos proporciona al estudio una visión más amplia y permite generalizar los resultados que se obtengan. La muestra es homogénea en cuanto a la edad y el nivel académico, también en cuanto al tipo de centro de procedencia. Todos son Institutos de Enseñanza Secundaria que forman parte de la red de centros públicos de la Comunidad de Madrid. No obstante, por su ubicación, sus características son muy distintas.

La muestra nos permite estudiar cada grupo por separado, analizar las correlaciones entre las distintas variables correspondientes a las pruebas; pero también hacer un estudio comparativo de otras variables externas (género, ubicación del IES, tipo de población) que puede resultar muy rico, si bien no es el objetivo principal de la tesis y podrá ser objeto de investigaciones posteriores.

En cuanto al tamaño de la muestra, se consiguió un número de alumnos suficiente, desde el punto de vista estadístico, para que los resultados resulten concluyentes. Aunque, según Dörnyei (2003: 73-74), no hay normas estrictas a este respecto: “Unfortunately, there are no hard and fast rules in setting the optimal sample size”, el autor ofrece algunas pistas indicativas:

> From a purely statistical point of view, a basic requirement is that the sample should have a *normal distribution*, and a rule of thumb to achieve this, offered by Hatch and Lazaraton (1991), is that the sample should include 30 or more people. [...] From the perspective of statistical significance, the principal concern is to sample enough learners for the expected results to be able to reach statistical significance. [...] a good rule of thumb is that we need around 50 participants to make sure that these coefficients are significant and we do not lose potentially important results.

Nuestra muestra cumple ampliamente los parámetros indicados por el autor. Además, en este trabajo empírico, de cada sujeto se analizan más de 110 variables (más las correspondientes al cuestionario retrospectivo que suponen otras 20), así completamos un gran número de ítems de información de cada alumno.

Figura 8.1. Distribución de los sujetos de la muestra atendiendo a su procedencia

GRUPO	CENTRO DE PROCEDENCIA	Nº DE ALUMNOS
1	IES San Isidoro de Sevilla (Madrid)	39
2	IES Ágora (Alcobendas)	45
3	IES Vicente Aleixandre (Pinto)	40
4	IES Humanejos (Parla):	35

TOTAL: 162 alumnos

Dado que todos los grupos estudiados están por encima de 30, número mágico en la estadística de las Ciencias Sociales, el tamaño, aunque diferente, no va a incidir en las inferencias que se puedan hacer a partir de los resultados.

La información que se obtiene de cada alumno se centra en los siguientes puntos:

1. Sexo
2. Edad
3. Instituto de Enseñanza Secundaria de procedencia
4. Calificación obtenida en la asignatura de Inglés en la 2ª Evaluación
5. Calificación obtenida en “Cavemen”, una prueba tipo PAAU (Puntuación global, por preguntas y desglosada en parte objetiva y subjetiva)
6. Calificación obtenida en el C-test de100 ítems (Puntuación global y en cada subtest de 25 ítems)

7. Respuestas al cuestionario retrospectivo de opinión
8. Calificación obtenida en la PAAU convocatoria de junio 2001 (de los 81 sujetos presentados)

Los dos primeros reflejan características demográficas de los participantes, "biodata", que no deben faltar en cualquier estudio (Mackey y Gass 2005: 126). El tercero se refiere al IES de procedencia y nos permitirá hacer inferencias sobre el contexto. Y el resto, recopila información sobre la actuación de los sujetos.

Dentro del cuestionario se recogen además otros datos de los sujetos, como las oportunidades de aprendizaje del Inglés externas al entorno escolar. No es necesario recopilar datos como la primera lengua (L1) o el nivel de competencia en la segunda (L2), ya que la muestra es homogénea en estos aspectos.

Aunque cualquier análisis comienza con las variables antropométricas, dada la poca variabilidad que presenta la edad en nuestra muestra, ésta no será analizada. Puesto que todos los participantes en el estudio son alumnos de 2º de Bachillerato, su edad es muy similar y tal análisis no aportaría nada al estudio.

En cuanto al género, la muestra incluye 103 mujeres y 59 varones. A pesar de que *a priori* no consideramos que sea una variable significativa, analizaremos los resultados casi únicamente a título informativo. Nuestro estudio no lo plantea como objetivo y, por tanto, su diseño no es el más adecuado para este propósito.

En lo relativo a la experiencia académica previa, independientemente de que su trayectoria hubiera sido más o menos satisfactoria, todos los participantes en el estudio cursaban el mismo nivel dentro del sistema educativo español (2º de Bachillerato) y al finalizar el curso académico se disponían a enfrentarse a las Pruebas Unificadas de Acceso a la Universidad en la Comunidad de Madrid (si superaban el curso y así lo deseaban, con la intención de acceder a estudios universitarios posteriormente). Por tanto, deberían compartir un nivel de competencia semejante en Inglés, la lengua objeto de estudio.

Se eligió este nivel para realizar la investigación precisamente porque nos permitía unificar criterios, tomar las PAAU como referencia y así manejar los datos de un examen externo a los centros, pero a la vez común a todos ellos.

8.3. Materiales

Los materiales que se utilizaron en el estudio fueron:

1. Un C-test de 100 omisiones formado por cuatro textos distintos. Lo dividimos en cuatro subtests de 25 ítems cada uno. Se diseñaron dos modelos diferentes: C-test A y C-test B.
2. La prueba "*Cavemen?*" propuesta en la convocatoria de septiembre de 1999 de las PAAU para Bachillerato-LOGSE (Inglés) en la Comunidad de Madrid. Fue realizada en clase como preparación para el examen oficial de las PAAU. Se analizó tanto el resultado global en la prueba como los obtenidos en las distintas preguntas y partes que la forman.
3. Las calificaciones de Inglés en la 2ª Evaluación del curso escolar 2000/01.
4. La calificación obtenida en la prueba de Inglés en la convocatoria oficial de junio de 2001 de las PAAU de la CM (este dato se limita a los alumnos de la muestra que se presentaron a ellas).
5. Un cuestionario retrospectivo de opinión acerca del C-test.

En los apartados siguientes aportamos más información acerca de ellos. No obstante, tanto las pruebas definitivas como el cuestionario pueden consultarse en el Apéndice.

8.3.1. C-test: Diseño

Siguiendo los parámetros de Klein-Braley (1985) se elaboró un C-test de cien omisiones, formado por cuatro textos distintos, todos procedentes de exámenes recientes de las Pruebas de Aptitud para el Acceso a la Universidad (Fig. 8.2).

Para crear el C-test se tomó cada texto, respetando la primera oración, y a partir de ese punto, se iniciaron las mutilaciones siguiendo la "regla del dos". Cuando se completaron las 25 omisiones se dejó una última oración intacta y se prescindió del resto del texto original. De este modo, se estructuró el C-test de 100 ítems en

cuatro subtests equilibrados, frente a otros, como el creado por Dörnyei y Katona (1992), cuyo número de ítems en los subtests oscilaba entre los 17 y 24.

Fig. 8.2. Textos seleccionados y procedencia:

Road accidents	Universidades de Madrid, Junio de 1999. Prueba de Acceso de Inglés – Bachillerato LOGSE
Evolution	Universidades de Madrid, Junio de 1998. Prueba de Acceso de Inglés – Bachillerato LOGSE
American imperialism	Universidades de Madrid, Junio de 1997. Prueba de Acceso de Inglés – Bachillerato LOGSE
Women doctors. Are they different?	Universidades de Madrid, Junio de 1996. Prueba de Acceso de Inglés – Bachillerato LOGSE

El siguiente cuadro-resumen muestra la estructura y los modelos del C-test:

Figura 8.3. Estructura y modelos del C-test aplicado

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

Como queda reflejado en la Figura 8.3, se crearon dos diseños de examen, modelos A y B, alternado los mismos textos. En los dos primeros subtests de ambos modelos figuraban los espacios correspondientes a las omisiones de cada ítem, ayuda que debería facilitar la recuperación del texto original, como se comentará más adelante, y cuya eficacia queríamos comprobar.

Para su estudio, el modelo A del C-test se divide en cuatro subtests de 25 omisiones, y del mismo modo se organiza el modelo B. En ambos modelos se mantiene la indicación del número de letras necesario para completar cada omisión en los ítems 1 a 50 (en la Fig. 8.3 aparece indicado mediante una línea discontinua).

En la administración de la prueba, los modelos A y B se distribuyeron al azar entre los sujetos participantes en el estudio; para hacerlo se siguió el sistema de pares e impares según el lugar que ocupaban en el aula.

Antes de comenzar el C-test los sujetos debían completar unos datos sociométricos básicos (nombre, fecha, edad, género, IES) para su identificación y posterior análisis de las variables género e IES de procedencia.

A continuación explicamos con mayor detenimiento el proceso de elaboración del C-test, es decir, las distintas fases y tareas de su diseño. Veremos con qué criterios se llevó a cabo la selección de los textos utilizados y cómo se probó su funcionamiento con nativos antes de ser aplicado a los sujetos participantes en el estudio. Asimismo, comentaremos el criterio elegido para su corrección y las instrucciones que se entregaron a los alumnos justo antes de su administración.

8.3.1.1. Proceso de selección de textos

La literatura insiste en la importancia de una adecuada selección de los textos sobre los cuales se van a crear pruebas de cierre.

Oller (1979) consideraba que, *a priori*, cualquier texto puede servir para este propósito, sin embargo Klein-Braley (1984: 97) constató que, en la práctica, no es tan fácil encontrar textos adecuados para la creación de pruebas de cierre. Los problemas radican en el tema y en el grado de dificultad de los textos. La autora criticó la subjetividad que implica la selección de textos por parte del profesor.

Brown (1993) demuestra que no se puede hacer un *cloze* a partir de cualquier texto, si se quiere que funcione bien: “In short, it appears that it is not a good idea simply to take a book off the shelf, select a passage and develop a cloze test form it”. Por ello, propone el pilotaje de pruebas de cierre creadas a partir de varios textos y la posterior elección de aquella que muestre mejor funcionamiento, es decir, mejores medias y desviación estándar.

Tomando las consideraciones de Oller (1979) como punto de partida, Klein-Braley reflexiona en concreto sobre el C-test. Al ser una prueba de cierre le afectan también el tema y la dificultad de los textos. Propone la utilización de textos cuyo tema pueda ser calificado como "neutral". También es interesante la aportación de Raatz (1983: 125), que insiste en que se cuide la variedad de los textos, pues propicia la representatividad de la lengua: "C-tests use unsystematically selected texts, whose subject matter is as varied as possible".

Reproducimos unos fragmentos de la revisión que hizo Carroll (1987: 102) del libro *C-Tests in der Praxis* de Klein-Braley y Raatz (1985), en ellos se comentan aspectos relativos a la selección de textos:

> The passages are to be selected for their appropriateness for a target population of examinees such that the expected average dificulty is around 50% success in filling the blanks. Adult native speakers of the language, however, would be expected to have at least 95% success.
>
> Usually, passages are selected on the basis of intuitive judgments about difficulty and content, but on the matter of difficulty level, Klein-Braley reports investigations leading to objective estimates based on type-token ratios and sentence lengths.

En cuanto a la dificultad, como se aprecia en la cita anterior, además de la intuición, y con el fin de lograr mayor objetividad, la autora sugiere seguir los criterios de *type/token ratio* (variación léxica) y longitud de las oraciones. En nuestro análisis de los textos aplicados añadiremos el de la densidad léxica, que se utiliza habitualmente en las investigaciones sobre vocabulario.

Diversas investigaciones han utilizado índices de legibilidad (*readability*) para determinar el nivel y la validez de los textos utilizados. Lee (1996) usó en su estudio la fórmula de Dale y Chall (1948)[63]. Pero Klein-Braley (1984: 99) manifiesta claramente que no le interesan tales índices, con frecuencia no demasiado fiables, sino simplemente determinar la dificultad de los textos para cada grupo concreto de alumnos. Para ello, utiliza ecuaciones de regresión.

[63] Dale-Chall Readability Index: (0.0496 * __Average Sentence Length) + (0.1579 * __Percent Difficult Words) + 3.6365 = __Raw Score. Existen otras fórmulas, tales como la de Flesh (1948), Kincaid, Coleman-Liau, Automated Readability Index, Fog Index, etc.

Por otra parte, Babaii y Ansary (2001: 217) recomiendan que los textos utilizados en los C-tests no sean excesivamente fáciles: “to encourage macro-level processing, the text should be challenging to the target test takers”.

Mochizuki (1995), que trabajó con diversos tipos de textos, concluyó que los más adecuados para la creación de C-tests son los narrativos y de cierta longitud. Sin embargo, Ikeguchi (1998), como Klein-Braley (1997), recomienda los C-tests creados a partir de varios textos cortos.

En este estudio empírico se decidió trabajar en la línea propuesta por Klein-Braley (1997). 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. Puesto que nos movemos en el ámbito de las PAAU, tomamos la determinación de ceñir la selección de textos para el diseño de C-tests a los ya utilizados en pruebas o modelos de Selectividad. Esto nos supuso un menor coste de tiempo, lo que contribuyó a aumentar la factibilidad de la prueba, y garantizó la uniformidad de nivel. Así pues, tomamos los textos aparecidos en las pruebas de Selectividad de los últimos años en la CM y los que se proporcionan al profesorado de Inglés de Enseñanza Secundaria como modelo. De forma intuitiva, y con la inevitable subjetividad del profesor, fuimos descartando algunos textos, generalmente menos atractivos por su tema, hasta quedarnos con sólo seis. Klein-Braley (1997) recomienda comenzar con un buen número de textos, siempre más de cuatro, que será el número definitivo.

Como es sabido, los textos en que se basan las pruebas de Inglés de las PAAU son auténticos y tratan temas variados de interés general (divulgativos, periodísticos, etc.). En principio, son textos adecuados, que se adaptan al nivel de madurez y conocimiento que se supone al alumno de 2º de Bachillerato.

Sobre los seis textos seleccionados confeccionamos C-tests siguiendo la “regla del dos” y aplicando las normas de sus creadores. Una vez fijadas las 25 omisiones en cada texto, (hasta llegar a un total de 100 por C-test), Klein-Braley aconseja la administración a hablantes adultos nativos, bilingües, o bien profesores de la lengua meta. Después se debe proceder a elegir los textos definitivos.

Se siguió fielmente este proceso. Probamos el funcionamiento de la prueba con un nativo, una persona bilingüe y varios profesores de Inglés de Enseñanza Secundaria. En todos los casos se alcanzó más del 90% de respuestas correctas

(Klein-Braley 1997: 64). Por tanto, sólo restaba decidir los cuatro textos definitivos y su orden de aparición. Según la autora, el profesor ha de ordenar los textos por orden creciente de dificultad, simplemente de forma intuitiva (ibídem). De este modo se consigue que el alumno entienda bien el mecanismo de la prueba y aumente la motivación, aunque la puntuación media sea la resolución correcta de aproximadamente el 50% de los ítems.

> Such a test can be very frustrating both for teacher and pupils, particularly since in the C-Test the subject is well aware that items have not been solved [...]. For this reason it is suggested that the first text should be very easy and that the difficulty should increase throughout the test so that the final text is very difficult. (Klein-Braley y Raatz 1984: 144)

> C-tests have been shown to have high reliabilities even when they were much too difficult or too easy for the subjects involved and we feel it is important for the icebreaker to be so simple that every test subject understands exactly what the C-principle demands form him or her. (Klein-Braley 1984: 98)

En nuestro caso, debido a su procedencia, el nivel de dificultad de los textos debía ser bastante homogéneo. Así pues, fijamos el orden de los textos de forma intuitiva, aunque en principio, podrían haber sido colocados de forma aleatoria, dada su homogeneidad de nivel: “Road accidents”, “Evolution”, “Women doctors”, “American imperialism”. Más adelante veremos que, a pesar de todo, algunos textos presentan menor dificultad que otros. Las diferencias entre ellos se pusieron de manifiesto al analizar su densidad y variación léxicas.

Para conseguir acentuar el orden creciente de dificultad decidimos incorporar un cambio en el formato. Nos dimos cuenta de que, normalmente, en los C-tests creados por Klein-Braley y Raatz no se señala el número de letras que se omite en cada palabra. En realidad, no es necesario, puesto que el sujeto debe saber que se omite la segunda mitad de la palabra, como se indica en las instrucciones. Pero pensamos que contar con esa ayuda adicional podía facilitar la tarea del alumno e influir en los resultados obtenidos. Finalmente se tomó la decisión de aportarla en los dos primeros subtests del C-test (omisiones 1 a 50). Así logramos seguir también, en cierto modo, la idea de la dificultad creciente propuesta por Klein-Braley (1997).

Como el número total de sujetos de la muestra era lo suficientemente grande, vimos la posibilidad de crear dos modelos de prueba: A y B, cambiando el orden de

los textos pero manteniendo constante el formato (ítems 1-50 con indicación del número de letras omitidas y 51-100 sin pistas). De esta manera se podría analizar el funcionamiento de cada modelo de C-test y buscar el porqué de las diferencias entre ellos, si las hubiera.

8.3.1.2. Elección del criterio de corrección

En cuanto al criterio de corrección, se eligió el de la palabra exacta, que parece el más adecuado para el C-test, ya que las características de la prueba dejan poco margen para que varias palabras distintas sean correctas en un mismo punto y coincidan totalmente en su primera mitad. De este modo, aseguramos la objetividad.

En el cuestionario retrospectivo, algunos alumnos se quejaron porque les parecía que ciertas omisiones del C-test admitían varias posibilidades. Sin embargo, en la posterior revisión de las pruebas no se encontraron datos que fundamentaran esta impresión.

Lo que sí descubrimos en la corrección fueron problemas ortográficos, de *spelling*, en algunas palabras. Indican que el alumno sabía cuál era la palabra correcta, reconocía el término buscado, pero no fue capaz de escribirlo correctamente (falló la producción escrita). En estos casos, quizá el criterio de corrección parezca demasiado estricto. No obstante, en términos generales, no lo consideramos un problema importante, teniendo en cuenta la escasa cantidad de ítems afectados exclusivamente por dudas ortográficas.

8.3.1.3. Instrucciones

Para asegurar la correcta comprensión de la tarea que se pedía a los alumnos y teniendo en cuenta su desconocimiento de la técnica, decidimos entregarles el modelo de C-test resuelto que figura en el Apéndice de la tesis, y acompañarlo de una breve explicación oral.

Hemos comentado en varias ocasiones la importancia que Klein-Braley concede a las instrucciones. Por eso, además de adjuntar el modelo resuelto, el C-test aplicado estaba encabezado por las siguientes instrucciones:

Figura 8.4. Instrucciones para completar el C-test

First of all, read each text carefully trying to understand its meaning. Then, complete the texts filling in the blanks with the appropriate letters.

Remember that the **second half of every second word has been deleted**, beginning with word two in sentence two.

In the first two texts each dash corresponds to a single letter.

Con el modelo resuelto y las instrucciones (orales y escritas) se pretende dar un primer paso hacia la deseable familiarización y valoración de la técnica por parte de los sujetos, que sólo se lograría totalmente mediante la administración repetida de C-tests.

8.3.1.4. Administración a hablantes nativos

Ya en la fase de selección de los textos habíamos aplicado la prueba a varias personas con alto nivel en la lengua (véase apartado 8.3.1.1). En la primera ocasión comprobamos que los C-tests creados a partir de seis textos preseleccionados funcionaban bien, puesto que en todos ellos los sujetos obtuvieron entre el 90% y el 100% de puntuaciones correctas. Posteriormente, los dos textos que parecían menos atractivos fueron descartados.

Una vez trazado el diseño del C-test que pretendíamos aplicar en este estudio, ya decididos los cuatro textos, fijadas las omisiones y el criterio de corrección, se administró la versión definitiva a un grupo de control formado por un hablante adulto nativo, otro de formación bilingüe y dos profesores de Inglés de Enseñanza Secundaria. Klein-Braley (1997) recomienda actuar de este modo para evitar sorpresas en la administración de C-tests. De nuevo, los resultados obtenidos confirmaron la idoneidad de la prueba, por tanto quedó ya lista para su administración a los sujetos de la muestra.

8.3.2. *Cavemen?*

Durante el curso académico se hicieron en la clase de Inglés distintos modelos de pruebas tipo Selectividad como preparación para las PAAU oficiales. Es ésta una práctica común en 2º de Bachillerato y un claro ejemplo de "enseñar para el examen" (véase el capítulo 3, apartado 3.8.5).

Decidimos tomar uno de estos exámenes como punto de referencia para estudiar las correlaciones del C-test. En el estudio se utilizó *Cavemen?*, que corresponde a la prueba propuesta en la convocatoria de septiembre de 1999 de las PAAU para Bachillerato-LOGSE en la Comunidad de Madrid.

Cavemen? comienza con un texto que los alumnos han de leer cuidadosamente. A continuación, propone cinco preguntas relacionadas con el texto. La última consiste en escribir una redacción de 80 a 100 palabras en lengua inglesa sobre uno de los dos temas propuestos.

Los alumnos realizaron la prueba en una sesión de clase, como una más de sus habituales prácticas. Una vez corregidos, tomamos varios datos. En primer lugar, el resultado global en la prueba, es decir la calificación obtenida en escala de 0 a 10 puntos. Además, agrupamos las preguntas de la prueba en dos tipos: las de carácter objetivo y las subjetivas. De este modo, pudimos analizar las correlaciones del C-test con la prueba en general, con cada una de las preguntas en particular, y con las partes objetiva y subjetiva de la misma.

8.3.3. Calificaciones de Inglés en la 2ª Evaluación

También interesaba conocer hasta qué punto los resultados obtenidos en el C-test eran coherentes con la valoración que sus respectivas profesoras hacían del progreso del alumno en la asignatura de Inglés. Sobre este aspecto, Klein-Braley (1984: 136) explica: "The use of teacher ratings is often viewed as problematical because such ratings are themselves not necessarily reliable. Their pragmatic validity in the context of the school system, however, is a fat of life".

Compartimos el punto de vista de Klein-Braley acerca de la validez e importancia de los juicios que el profesor, como profesional de la docencia, emite

sobre los alumnos. Por eso, respetamos la evaluación que las profesoras habían hecho de sus alumnos, con los criterios e instrumentos que consideraran pertinentes para ello. En este caso, nos interesa la consistencia de la medida para cada alumno, como individuo y no como grupo. Con ese fin, las profesoras de Inglés de los grupos que forman parte del estudio nos facilitaron las calificaciones de Inglés en la 2ª Evaluación del curso escolar 2000/01.

Quizá un análisis de los criterios, métodos e instrumentos de evaluación seguidos por cada una de las profesoras pusiera de manifiesto algunas inevitables diferencias entre unas y otras. No obstante, en el contexto académico en que nos encontramos, la evaluación educativa del rendimiento de los alumnos se rige por el mismo currículo (objetivos y contenidos mínimos) y criterios, los fijados por la legislación educativa vigente. Por tanto, la uniformidad está garantizada y podemos incluir este dato en nuestro estudio.

8.3.4. Calificaciones del examen de Inglés de las PAAU oficiales

Otro criterio objetivo de referencia con el que contamos, es la nota obtenida por los sujetos en el examen oficial de Inglés que forma parte de las Pruebas oficiales de Aptitud para el Acceso a las Universidades madrileñas (cada grupo lo realizó en la universidad correspondiente: Complutense, Autónoma y Carlos III).

Este dato aporta aún mayor objetividad a nuestra investigación. Pero, por sus características, también presenta algunas limitaciones destacables. En primer lugar, porque lamentablemente no tenemos dicha información de todos los sujetos. Sólo de aquellos que, una vez superado el Bachillerato, se presentaron a las Pruebas en la convocatoria de junio de 2001, que supone exactamente la mitad de la muestra total. En segundo lugar, porque incluso de los presentados a las PAAU no conocemos más que la calificación final y global de la prueba de Inglés, sin el desglose de los resultados obtenidos en cada pregunta.

Y por último, hemos de tener en cuenta las circunstancias que rodean a la PAAU. Como prueba selectiva presenta un componente importante de ansiedad en los sujetos que la realizan, y esto ha de afectar necesariamente al rendimiento.

A continuación, incluimos algunos detalles acerca de las Pruebas de Aptitud para el Acceso a la Universidad. Por razones de espacio y agilidad de lectura, y puesto que son ampliamente conocidas por todos, no nos detendremos en ellas más que para hacer algunos comentarios generales y mostrar de forma somera la estructura de la prueba de Inglés.

Las PAAU constituyen un referente externo común y obligatorio en el estado español cuyo propósito es unificar u homogeneizar las calificaciones obtenidas por los estudiantes, cualquiera que sea su procedencia, antes de su incorporación a la Universidad[64]. Al ser unas pruebas selectivas de madurez, de los resultados obtenidos en ellas dependerá en gran medida el futuro de los estudiantes españoles. La prueba de Inglés constituye sólo una parte de la PAAU, que incluye además pruebas sobre la mayoría de las asignaturas cursadas en 2º de Bachillerato.

En el capítulo 3 señalamos las implicaciones y el impacto o *washback* que producen en la sociedad (Alderson y Wall 1993, 1996; Messick 1996; Bailey 1996; Shohamy *et al*. 1996; Alderson y Hamp-Lyons 1996; Andrews *et al*. 2002), y en particular en profesores y alumnos, las pruebas estandarizadas de ámbito nacional, como las PAAU en España. Aludimos a cuestiones relativas a su preparación en el aula, el fenómeno conocido como *teaching to the test* (Gipps 1994) o *test-like teaching* (Shohamy 1997). En el contexto educativo español los efectos de las PAAU son evidentes: buena parte de las clases de Bachillerato van dirigidas a la superación de las PAAU, los profesores intentan aumentar la motivación y reducir la ansiedad de los alumnos para que puedan reflejar sus conocimientos en ella de la mejor manera posible. El actual examen de Inglés de las PAAU tiene como objetivo fundamental discriminar entre las actuaciones de los alumnos y hacerlo con el mayor grado de fiabilidad posible.

Por esta razón, aunque también podría ser considerado como *placement test* atendiendo a su función de "filtro selectivo" para el acceso a la Universidad, Herrera (1999: 90) lo categoriza como *proficiency test*:

[64] Como se comentó en capítulos anteriores, las actuales PAAU llevan ya más de treinta años funcionando en España. Actualmente, con la implantación progresiva de la LOE (2006) sigue temporalmente su vigencia hasta que se complete el desarrollo de dicha Ley en el curso 2009/2010, fecha en que está prevista la entrada en funcionamiento de las nuevas Pruebas de Acceso, aún por determinar.

> In the case of the ET, the target is to discriminate as reliably as possible. The University Examination Board looks for an accurate score which enables the academic authorities to rank students according to their proficiency and, which at the same time, allows the students to make their choice of Faculty courses according to the score obtained.

La prueba se basa en el marco teórico de Bachman (Herrera 1999: 91; Amengual Pizarro 2003: 53) y desarrolla uno de los tres componentes del dominio de la lengua: la competencia lingüística. Ésta se compone de competencia organizativa (gramatical y textual) y competencia pragmática. En la prueba se valora el constructo, subdividido en gramática, vocabulario, comprensión y expresión escrita.

A pesar de que se reconoce (García Laborda 2005; Fernández y Sanz 2005) que el examen de Inglés de las PAAU[65] debería valorar también las destrezas orales de la competencia comunicativa, todavía se imponen problemas de carácter económico y de infraestructura. En el momento actual, la prueba de Inglés vigente en el sistema educativo español es una prueba de comprensión y expresión escrita: "it will be observed that reading and writing skills rather than the oral dimension of communicative competence are highlighted" (Herrera 1999: 91).

Las pruebas de producción escrita, por definición, intentan medir la capacidad del individuo para expresarse por escrito en la lengua extranjera. Pueden ser pruebas directas o bien indirectas. Las indirectas buscan evaluar la expresión escrita mediante otras medidas que correlacionen bien con la capacidad de producción escrita. Las directas se basan en la producción real de textos escritos. Hamp-Lyons (1991: 5-14) considera que éstas últimas son las únicas realmente válidas, porque lo importante no es el conocimiento de reglas gramaticales, sino la capacidad real para utilizar la lengua escrita como vehículo de expresión de ideas y emociones.

Aunque las distintas universidades españolas gozan de cierta flexibilidad al plantear la prueba de Inglés, la estructura general presenta sólo leves variaciones. Podemos decir que la prueba de Inglés de las PAAU se estructura en torno a dos partes: una de carácter indirecto, que pretende valorar de forma objetiva el conocimiento de elementos discretos de la lengua, y otra de tipo directo, por tanto con cierto componente subjetivo.

[65] La literatura reclama una revisión de la prueba de Inglés de las PAAU que mejore su validez y fiabilidad, en definitiva, su calidad. El volumen "Estudios y criterios para una Selectividad de calidad en el examen de Inglés" (2005) recoge muchas de estas reivindicaciones, a las que nos sumamos.

La prueba se valora sobre una puntuación total de 10 puntos. En la CM cada una de las dos partes de la prueba tiene asignado un valor máximo de 5 puntos. Se intenta, de este modo, que la prueba sea un instrumento de medida equilibrado.

El ejercicio parte de un texto escrito, que se propone como punto de partida y constituye el eje central del examen, pues proporciona el tema en torno al cual gira toda la prueba. Es, en principio, un texto auténtico o levemente adaptado, y relacionado con temas de carácter divulgativo, periodístico o de interés general. El alumno debe leerlo en primer lugar, para luego responder a diversas cuestiones, unas objetivas (de comprensión del texto, de gramática y de vocabulario) y otras de tipo subjetivo (preguntas abiertas y redacción sobre uno de dos temas propuestos) (Herrera Soler 1999, 2001). El cuadro siguiente (Fig. 8.5) muestra su estructura.

Figura 8.5. Estructura de la prueba de Inglés de las PAAU

PAAU (LOGSE) Comunidad de Madrid	
Parte objetiva (5 puntos)	**Parte subjetiva (5 puntos)**
1. Pregunta de comprensión del texto: Verdadero o falso (2 puntos) 2. Pregunta de vocabulario (1 punto) 3. Pregunta de reflexión gramatical (2 puntos)	1. Pregunta abierta de comprensión del texto. (2 puntos) 2. Redacción sobre uno de los dos temas propuestos (3 puntos)

En nuestro estudio hemos tomado los resultados de una prueba tipo Selectividad, pero realizada en clase como preparación para el examen oficial. Sigue exactamente los parámetros que acabamos de explicar. Para esta tesis, una prueba tipo Selectividad aplicada directamente en el aula presenta ventajas con respecto a la convocatoria oficial de las PAAU, ya que proporciona una información completa: conocemos la puntuación obtenida en cada una de las preguntas y podemos, además, agrupar las puntuaciones de la parte objetiva y subjetiva de la misma. Disponemos de estos datos de todos los sujetos que realizaron el C-test (162), y

además, evitamos el componente de ansiedad propia de las pruebas selectivas externas. También contamos con la calificación obtenida en la convocatoria oficial de las PAAU de junio de 2001, pero sólo de 81 alumnos, el 50% de la muestra total.

En cuanto a la corrección de las PAAU, mencionaremos que la realizan tribunales formados por profesores de Universidad y de Enseñanza Secundaria, dirigidos por un coordinador, que generalmente pertenece al mundo universitario. Los correctores reciben unas instrucciones básicas relativas a las puntuaciones asignadas a cada pregunta, a las que hay que añadir una breve reunión previa a la administración de la prueba, y otra posterior para poner en común los criterios de evaluación y llegar a acuerdos que aseguren una valoración homogénea con independencia del corrector (Amengual 2003). Así pues, parece claro que los correctores de las PAAU no reciben la deseable formación específica (Bachman y Palmer 1996), simplemente algunas instrucciones expresas para la corrección de la prueba y, en concreto, para unificar la valoración de la parte subjetiva del examen, que incluye la redacción. Aunque pocas recomendaciones son necesarias para valorar la parte objetiva, basada en elementos discretos, en cuanto a las tareas de expresión escrita directa sí percibimos un vacío importante.

A pesar de todo, estudios recientes (Amengual 2003) sobre la fiabilidad de las puntuaciones de los ensayos de las PAAU han demostrado la fiabilidad inter-corrector en la valoración de las redacciones. La actuación de los correctores se guía por criterios personales y, en general, valora más los aspectos formales de la lengua, pero es consistente en sus distintas actuaciones.

En el Apéndice incluimos las “instrucciones para el corrector” que se suministraron en las universidades madrileñas en la prueba de Inglés de las PAAU de 2001. Las escasas claves referentes al ensayo simplemente pretenden lograr un equilibrio entre forma y contenido. De los 3 puntos totales, a la expresión (gramática, vocabulario, etc.) le corresponden 1,5 puntos aproximadamente y el mismo valor se da al contenido expresado en el ensayo (ideas, etc.). Todo ello se propone sólo de forma indicativa.

Evidentemente, en este tipo de pruebas se asegura el anonimato de los sujetos presentados como medida para evitar, en lo posible, los sesgos por parte del corrector. Cada corrector recibe un bloque de exámenes asignado al azar, cuyo

número suele oscilar entre los ciento cincuenta y los doscientos, dependiendo de las necesidades del tribunal correspondiente.

Pero, para evitar los sesgos de subjetividad que el ser humano puede aportar al examen, el anonimato no es suficiente, es vital que el corrector tenga la preparación necesaria para juzgar este tipo de prueba (Bachman y Palmer 1996: 221). Y también sería recomendable que cada examen fuera revisado por al menos dos correctores diferentes.

Tanto la falta de formación específica de los correctores en las técnicas de evaluación como la ausencia de doble corrección se deben, principalmente, a motivos de índole económica.

A pesar del elevado coste que tendría la implantación de estas medidas para una prueba a escala nacional, como las PAAU, pensamos que es nuestra obligación recordar a los responsables de la política educativa los beneficios de las mismas. Actualmente se intentan suplir ofreciendo a los examinandos la posibilidad de reclamar o solicitar doble corrección, pero sólo si expresan su disconformidad con la calificación recibida en la primera corrección.

Herrera (1999) cuestionó la validez de la parte objetiva o indirecta de la PAAU de Inglés, puesto que no discrimina entre los sujetos; por otra parte, su trabajo evidenció que la redacción es la parte de la PAAU que mejor muestra las habilidades reales del alumno en lengua inglesa.

En nuestra investigación intentaremos demostrar la validez concurrente del C-test con respecto a las PAAU, tomadas en su conjunto (puntuación global). Pero también analizaremos las correlaciones del C-test con cada una de las partes de la prueba (objetiva y subjetiva).

El C-test y la redacción (dentro de las PAAU) comparten el carácter de pruebas integradoras de producción lingüística (Lee 1996). Nos interesará, por tanto, conocer cómo es la correlación entre ambas.

8.3.5. Cuestionario

Dörnyei (2003: 9) apunta que los cuestionarios se pueden utilizar para averiguar lo que piensa un sujeto, pueden medir "*attitudes*, *opinions*, *beliefs*, *interests* and *values*". Destaca, además, la eficacia del procedimiento y su versatilidad como principales ventajas: "The main attraction of questionnaires is their unprecedented efficiency in terms of (a) researcher time, (b) researcher effort, and (c) financial resources".

En nuestro proceso empírico también creímos interesante conocer la opinión de los sujetos acerca del C-test, y se decidió que la administración de un cuestionario era el procedimiento más adecuado para recopilar la información, ya que el tamaño de la muestra dificultaba el uso de otros métodos aconsejables, como la entrevista personal o los *think-aloud protocols*.

Así pues, para poder analizar algunos aspectos relativos a la validez aparente del C-test, una vez terminada su aplicación, se pasó a los sujetos un cuestionario retrospectivo de opinión basado en el que utilizó Jafarpur (1995: 207). En él se pedía a los alumnos una valoración de la prueba atendiendo a distintos aspectos.

Nuestro cuestionario quedó estructurado en tres partes:

- La primera demanda algunos datos personales, *biodata*, que no debe faltar en un cuestionario (Dörnyei 2003), excepto la identificación de los sujetos, para asegurar el anonimato y la libertad en las repuestas[66].
- La segunda parte del cuestionario solicita información de tipo valorativo sobre diversos aspectos del C-test, las dificultades surgidas en su realización y la impresión producida en los sujetos.
- La tercera parte pide opinión sobre su posible futura utilización en pruebas selectivas, como las PAAU.

En el capítulo 12 se estudia el cuestionario con mayor detalle. La versión definitiva aplicada en esta investigación puede consultarse en el Apéndice.

[66] Respecto al anonimato de los sujetos en este tipo de cuestionarios Mackey (2005: 124) recomienda: "In reporting information about participants, the researcher must balance two concerns. The first is the privacy and anonymity of the participants; the second is the need to report sufficient data about the participants to allow future researchers to both evaluate and replicate the study".

8.4. Contexto: Perfil de los IES en que se realizó el estudio

En este apartado completamos algunos aspectos del contexto en que se centra el presente estudio. En concreto, revisamos el perfil de los centros a los que pertenecen los cuatro grupos de alumnos participantes en el estudio.

En el apartado 8.2 de este capítulo hemos descrito las características del grupo de sujetos que se toma como muestra. Constituye un grupo homogéneo de 162 alumnos que comparten algunas variables, como la edad, el nivel académico (2º de Bachillerato) y su escolarización en IES públicos de la CM.

En cuanto a las características de los cuatro centros educativos, hemos de decir que reflejan realidades sociales bien distintas. Los centros en que se realizó el estudio no fueron elegidos al azar. Se intentó que reflejaran los distintos estratos socioeconómicos presentes en los IES de la Comunidad de Madrid, para reducir en lo posible el error sistemático de la muestra. Nos interesaba contar con una muestra que fuera fiel espejo de la diversidad existente en los IES de la CM, para que los resultados fueran generalizables.

Haremos una breve reseña de las circunstancias de cada IES participante.

Comenzamos con el IES Ágora, perteneciente al Área Territorial Madrid-Norte. El centro se encuentra en Alcobendas, ciudad de la Zona Norte Metropolitana muy próxima a la capital (a sólo 13 kms.). En ella conviven colectivos con niveles de renta muy diferenciados. Destaca la escasa incidencia de los fenómenos de marginalidad urbana. La población es joven y sociológicamente diversa, con un nivel educativo medio-alto.

El IES San Isidoro de Sevilla se encuentra en Madrid capital, en una zona privilegiada, de tipo residencial. Está enclavado en un área abierta y de ambiente eminentemente universitario, rodeado por instalaciones de la Universidad Complutense, el CEU y varios colegios mayores. Es un centro relativamente pequeño, con sólida tradición de prestigio entre los IES madrileños y en la zona. Ha funcionado siempre como centro en el que se impartían enseñanzas de Bachillerato.

Los IES Vicente Aleixandre y Humanejos pertenecen a la Dirección de Área Territorial Madrid-Sur. Se encuentran ubicados en las poblaciones de Pinto y Parla respectivamente. El primero ya comenzó como Instituto de Bachillerato, mientras que el segundo fue fundado como centro de Formación Profesional hace 28 años.

Pinto y Parla se consideran ciudades dormitorio de la periferia sur madrileña. Comparten con Alcobendas el carácter periférico, pero su universo social es muy distinto. Son fundamentalmente ciudades industriales, obreras. Arrastran el lastre de fuertes problemas económicos y sociales, sobre todo Parla. Pero cuentan con una población muy joven, que está cambiando el perfil de la zona, y acogen a un gran volumen de población inmigrante.

El IES Humanejos tiene una arraigada tradición en el campo de la Formación Profesional en Parla, pues fue el primer centro fundado con ese propósito. Sin embargo, encuentra dificultades para suscitar en sus alumnos el espíritu universitario. Al acabar la Enseñanza Secundaria Obligatoria la mayoría se decanta por otras opciones: muchos se incorporan al mundo laboral, otros inician estudios de FP y sólo un pequeño grupo elige continuar su formación en Bachillerato, con intención de acceder a estudios universitarios en el futuro. Como veremos, de los participantes en el estudio, los alumnos de Bachillerato del IES Humanejos son los más reacios a presentarse a las PAAU. En su lugar, abandonan los estudios o se incorporan a los Ciclos Superiores de Formación Profesional.

No se pretende con este trabajo simplemente contrastar los resultados obtenidos entre centros, puesto que dejaríamos múltiples variables fuera de estudio. Más bien al contrario, nuestro objetivo es constatar el funcionamiento del C-test en distintos IES y situaciones, respetando siempre las peculiaridades de los IES y de los grupos de sujetos en que se aplican, independientemente de su origen y características. Por otra parte, la variedad de extracción de los grupos nos asegura un universo realista y plural.

8.5. Procedimiento

En este apartado incluimos los detalles relativos a la selección de los sujetos, la distribución del tiempo y el proceso seguido para completar el estudio, teniendo en cuenta sus objetivos y las variables analizadas.

8.5.1. Selección de los sujetos: muestra

Como paso previo a la recopilación de datos y materiales, contactamos personalmente, ya en el primer trimestre del curso 2000/01, con los distintos IES que iban a formar parte del estudio. El IES Humanejos y el San Isidoro ya habían colaborado desinteresadamente con nosotros en las pruebas piloto.

Ahora la muestra se ampliaba para dar cabida a otros centros de ubicación y características diferentes, con intención de aumentar su representatividad.

Las distintas profesoras de Inglés fueron informadas directamente por la investigadora, tanto oralmente como por escrito, de todos los detalles del trabajo y de los datos que iban a ser necesarios. Era vital una colaboración directa y estrecha. Nos comunicaron qué grupos concretos de 2º de Bachillerato iban a participar en el estudio y se mostraron dispuestas a ayudar activamente en todo el proceso.

8.5.2. Distribución del tiempo

La recopilación del material para este trabajo comenzó en marzo de 2001, cuando se tomaron los datos de las calificaciones de los alumnos en la asignatura de Inglés en la 2ª evaluación del curso académico 2000/01.

Continuó con la aplicación de la prueba *Cavemen*? en una de las primeras sesiones de clase de Inglés del tercer trimestre, en abril de 2001. Siguió con la administración del C-test y el cuestionario en el aula en el mismo mes. Y finalizó en junio de 2001 cuando se recogieron las calificaciones de Inglés de los alumnos presentados a las PAAU (véase la Fig. 8.6).

Siguiendo el mencionado calendario, fuimos recogiendo los distintos materiales de cada grupo de sujetos. A medida que recibíamos los datos, se iban incluyendo en tablas para su posterior tratamiento informático y estadístico, asignando un lugar a cada IES y un número a cada alumno.

Figura 8.6. Distribución del tiempo

Marzo de 2001	Recogida de las calificaciones de Inglés en la 2ª evaluación del curso 2000/01
Abril de 2001	Realización de la prueba *Cavemen*? en una sesión de inglés del tercer trimestre
Abril de 2001	Realización del C-test y del cuestionario retrospectivo en una sesión de clase de Inglés
Junio de 2001	Recogida de las calificaciones de Inglés en la convocatoria oficial de junio de 2001 de las PAAU.

En primer lugar se confeccionaron las tablas con las calificaciones de Inglés de la 2ª Evaluación. Posteriormente se aplicó la prueba *Cavemen?* en una sesión de clase de Inglés del mes de abril. Cada profesora eligió la fecha que mejor se adaptaba a su programación de la asignatura. En ningún caso supuso una ruptura con la rutina de las clases de Inglés, puesto que este tipo de pruebas es habitual como preparación para la Selectividad. Se realizó en 60 minutos. Resulta un periodo de tiempo suficiente, a pesar de que en las PAAU oficiales la asignación de tiempo es mucho mayor. Los alumnos no fueron informados de que sus resultados serían analizados, pretendíamos que la prueba fuera reflejo real de la competencia lingüística del alumno en este punto del curso, cercana ya la Selectividad, y con los mínimos condicionantes externos.

Una vez realizada fue corregida por las respectivas profesoras, siguiendo los criterios habituales que se recomiendan para las PAAU, pero como una más de las aplicadas durante el curso, y posteriormente revisadas por la investigadora. Recibimos una copia de cada uno de ellos y los datos obtenidos se incorporaron a la tabla: puntuación global, puntuación de cada pregunta y puntuación lograda al agrupar las preguntas de tipo objetivo y subjetivo.

Una semana después se aplicó el C-test de 100 omisiones y el cuestionario retrospectivo. Se informó a los alumnos de que la prueba formaba parte de un estudio empírico: se trataba de una técnica nueva para medir su competencia en la

lengua. Después, en algunos casos, y por deseo expreso de las profesoras, los resultados obtenidos fueron tomados en cuenta en la evaluación de la asignatura.

En cuanto al tiempo, Connelly (1997) recomienda que el dedicado a la resolución de C-tests sea generoso para que los sujetos puedan trabajarlos bien. Propone un tiempo mínimo de entre 5 y 7 minutos para cada subtest (20-25 ítems). En este caso, la prueba se administró también durante una sesión de clase de Inglés, de 50 minutos, aunque se hizo uso de los minutos previos para introducir la técnica con las explicaciones pertinentes y el modelo de C-test resuelto. Y cuando finalizó se ocuparon diez minutos más para contestar al cuestionario. Así pues, el tiempo real se incrementó sensiblemente hasta aproximadamente 60 minutos. Una vez recogidos fueron entregados a la investigadora para su corrección y análisis. Los C-tests corregidos se devolvieron a los alumnos pocos días después. Por tanto, los sujetos pudieron comprobar sus resultados, a veces sorprendentes.

Algunas profesoras nos comunicaron que habían "reutilizado" el C-test a posteriori como material de clase para llamar la atención de los alumnos y reflexionar sobre determinados puntos gramaticales, de vocabulario, errores comunes, etc., y destacaron su eficacia, porque ayuda a tomar conciencia de las claves de que disponían para su solución y de las estrategias utilizadas.

Hinofotis (1987) y Buck (1988) sugieren el uso de pruebas de cierre en las clases. En este sentido, Lee (1996: 65) confirma: "the confirmation of the cloze procedure as a valid language proficiency test suggests the use of the procedure beyond a testing format. It can be used as an effective teaching device". En la práctica, constatamos que su apreciación puede aplicarse también al C-test.

Las tablas de resultados se completaron con la puntuación total en el C-test (sobre 100) y la obtenida en cada uno de los subtests de 25 omisiones. Fueron denominados, respectivamente, Ctesttot, Ctest1, Ctest2, Ctest3 y Ctest4. Los cuestionarios de opinión quedaron en nuestras manos para su análisis.

Por último, debimos esperar hasta la entrega de las calificaciones de las PAAU oficiales, en junio de 2001. Con este dato culminó la tarea de recopilación de los materiales necesarios para nuestro estudio.

8.6. Tratamiento de los datos

La investigación actual sobre Evaluación de la Lengua utiliza los medios estadísticos. La ciencia estadística aporta al lingüista herramientas y procedimientos básicos para determinar el funcionamiento y la validez de una prueba. Alderson (en Bachman 2004: ix) explica así los lazos entre ambas ciencias:

> Language tests are intended to measure, and without quantification they cannot measure. Quantification implies numbers and numbers imply statistics. And so, although a firm understanding of the nature of language is essential for the trained language tester, so too is at least a basic familiarity with statistics.

La estadística permite resumir y describir la gran cantidad de información que se extrae de un examen, para después hacer inferencias. Por tanto, podemos distinguir dos tipos de análisis estadístico: el descriptivo y el inferencial (Bachman 2004: 34). Desde esta perspectiva se ha llevado a cabo la investigación sobre el C-test, a partir de los resultados obtenidos en las distintas pruebas aplicadas. Se han realizado ambos tipos de análisis mediante diversas técnicas estadísticas, utilizando el programa *Statistical Package for Social Sciences* (SPSS 12.5).

La distribución de las puntuaciones de las pruebas se refleja en los estadísticos descriptivos: tablas de frecuencias, histogramas, diagramas de cajas y barras. En los histogramas que ilustran nuestra investigación revisamos la simetría, la curtosis (leptocúrtica, mesocúrtica y platicúrtica) y el sesgo de las curvas (positivo o negativo). Para cada prueba se señalan las medidas de tendencia central: moda, mediana y media, así como el rango y la desviación estándar. A veces también se realiza el análisis de la varianza (ANOVA) y las pruebas de significación (T tests).

El estudio intrínseco del C-test se completa desglosando la prueba en subtests y estudiando las correlaciones entre cada subtest y el total de la prueba. Además, según el formato de las omisiones, se divide la prueba en dos partes (con omisiones guiadas y no guiadas) y se comparan las medias obtenidas en cada una de ellas.

Con respecto a la validez concurrente del C-test, el principal referente externo fue la prueba oficial de Inglés de Selectividad (junio de 2001), pero también la prueba *Cavemen?* realizada en el aula (subdividida en parte objetiva y subjetiva) y las calificaciones de Inglés en la 2ª Evaluación. La investigación correlacional intenta determinar la existencia de relación entre dos variables. El coeficiente de correlación

se expresa en valores desde 0 a 1, e indica si la relación entre las variables es lineal y significativa. El análisis de la validez concurrente se hace a través de las correlaciones de Pearson entre el C-test y las distintas pruebas.

Por otra parte, la fiabilidad del C-test se comprueba mediante el método de "análisis por mitades" y el cálculo del Alfa de Cronbach (Klein-Braley 1984), además del análisis de las correlaciones.

Para determinar los factores que condicionan el grado de facilidad/dificultad de recuperación de las omisiones: frecuencia, familiarización, términos léxicos o funcionales, formato, etc., estudiamos los estadísticos de frecuencias de distintos ítems y comparamos su funcionamiento en los modelos A y B. Indican el porcentaje de aciertos y fallos en cada ítem y los sujetos que lo dejan sin hacer.

El análisis de las variables textuales se realiza mediante el cálculo de su variación y densidad léxicas, siguiendo las pautas de Laufer y Nation (1995) y Schmitt (2000: 75).

La técnica de regresión lineal, que tiene valor predictivo, se utiliza en nuestra investigación para el analizar el carácter de la relación entre las distintas partes o subtests que forman el C-test (C-test 1, C-test 2, C-test 3 y C-test 4) y las otras pruebas aplicadas, que consideramos variables dependientes (VDs).

Para explorar la incidencia de factores demográficos, como el género y el IES, en la actuación de los sujetos en el C-test, se analizan los promedios obtenidos en las pruebas, el ANOVA, el modelo lineal general y se hace el análisis de varianza univariante, puesto que la disparidad de promedios no implica necesariamente la existencia de diferencias significativas en la actuación de los grupos.

Por último, la valoración de los datos obtenidos en el cuestionario, encaminado a determinar la validez aparente del C-test, se hace mediante la elaboración de tablas de frecuencias, diagramas de barras y el procedimiento de análisis factorial.

CAPÍTULO 9. ANÁLISIS EMPÍRICO DE LA VALIDEZ DEL C-TEST

9.1. Introducción

En este capítulo comienza el análisis del C-test desde una perspectiva empírica. Se intentará responder a las preguntas de investigación planteadas en la Introducción de la tesis siguiendo el orden de presentación de datos. De este modo, podremos confirmar o rechazar las cuatro primeras hipótesis de trabajo. Tomaremos como referencia las otras pruebas aplicadas a los sujetos en la fase experimental de nuestro trabajo y la Selectividad de junio de 2001. Los resultados corresponden a la sistematización de los datos recogidos a partir de los distintos materiales utilizados: puntuaciones en el C-test y los subtests que lo forman, en la prueba *Cavemen?* y sus subapartados, valoración del profesor de Inglés respectivo (calificaciones de la 2ª Evaluación) y calificación obtenida en la prueba de Inglés de las PAAU oficiales de junio de 2001.

Se llevará a cabo el proceso de validación del C-test como prueba de evaluación. Partiremos del análisis intrínseco de la prueba: estructura en subtests, formato y modelos aplicados. Este proceso se desarrolla en los siguientes pasos:

- Análisis comparativo de los resultados totales del C-test y de cada subtest: promedios e histogramas
- Análisis de los resultados obtenidos en los modelos de C-test A y B
- Análisis de las consecuencias del formato: omisiones guiadas o no
- Análisis de las variables textuales: variación léxica, densidad léxica, tema del texto
- Análisis de los factores que condicionan la recuperación de las omisiones: frecuencia, familiarización, términos léxicos o funcionales

Del análisis de la prueba de Inglés tipo PAAU *Cavemen?*, desglosada en parte objetiva y subjetiva, se pasará a determinar la validez concurrente del C-test y sus correlaciones con las otras pruebas, se analizará su validez criterial y la fiabilidad de la prueba. Se estudiarán los promedios y las correlaciones del C-test con cada una de las variables estudiadas.

9.2. Proceso de validación del C-test como prueba de competencia lingüística

Al acometer el proceso de análisis y validación del C-test retomamos el concepto unitario de validez desarrollado por Messick (1989), basado en que todas las cualidades de las pruebas se interrelacionan.

En el capítulo 3 de la tesis definimos los distintos tipos de validez. Aunque a menudo se solapan, para nuestro estudio es útil validar los diversos aspectos por separado. El C-test pretende ser una prueba de competencia lingüística (*proficiency test*) y, como tal, en su proceso de validación hemos de considerar:

1. Validez de constructo
2. Validez de contenido
3. Validez criterial: concurrente y predictiva
4. Validez aparente
5. Validez consecuencial
6. Fiabilidad

Para demostrar la validez del C-test, en primer lugar es necesario fijar los límites del constructo que pretende medir. En este caso, ya se ha comentado que el C-test fue creado por Klein-Braley y Raatz (1981, 1984) para medir la competencia general en lengua inglesa y, como prueba de cierre, se inspira en los principios de "redundancia reducida" (Spolsky 1973) y "gramática pragmática de expectativas" (Oller 1979).

Por otra parte, también hemos de tener en cuenta su validez de contenido (Hughes 1989; Bachman *et al*. 1996), es decir, la relevancia y representatividad de las estructuras que incluye la prueba. Puesto que el C-test se crea a partir de textos

auténticos y variados, según Klein-Braley y Raatz (1984: 144), su representatividad queda asegurada: “The text is a sample of the language and the mutilations in the C-test sample the text”. Pero será el estudio de la actuación real de los sujetos en la prueba lo que nos dé muestras de su validez de forma definitiva.

Centraremos nuestro análisis en la validez criterial. La literatura contempla dos tipos: concurrente y predictiva. En este capítulo intentaremos mostrar la validez criterial concurrente del C-test con respecto a otras pruebas independientes que miden la misma capacidad, tomadas como referencia (*Cavemen?* y PAAU oficiales). Puesto que las PAAU fueron realizadas con posterioridad al C-test, también encontraremos pistas relativas a su validez predictiva y de constructo.

Como se comentó en el capítulo 3, la validez concurrente viene dada por la correlación entre los resultados de las pruebas. Para que la correlación tenga valor ambas pruebas se deben realizar en un breve intervalo de tiempo (Davies 1983; Hughes 1989). Los datos de nuestra investigación se recopilaron en un periodo aproximado de un mes (véase la temporalización en el capítulo 8, apartado 8.5.2), y las PAAU de referencia dos meses después (junio 2001).

La validez predictiva, por otra parte, se refiere al grado en que los resultados obtenidos en una prueba pueden predecir la actuación del alumno en una situación futura. Si se hiciera un seguimiento, lo cual no es fácil dadas las características del estudio, los resultados de las PAAU deberían ser indicativos del futuro académico de los alumnos ya inmersos en el mundo universitario.

En cuanto a la validez predictiva del C-test, el diseño y temporalización de nuestro trabajo empírico nos permitirán comprobar si los resultados del C-test correlacionan significativamente con los obtenidos por los alumnos en las PAAU de junio de 2001, que se realizaron dos meses después del C-test. La validez consecuencial, es decir, los posteriores efectos beneficiosos de la prueba en los distintos agentes que participan del proceso de enseñanza-aprendizaje, estará garantizada si la prueba demuestra ser válida en los aspectos anteriormente descritos, entendiendo la validez como marco unitario.

El análisis de la validez aparente del C-test aplicado merece un espacio propio (capítulo 12). En él trabajaremos fundamentalmente con los datos del cuestionario retrospectivo.

9.3. Aspectos descriptivos del C-test aplicado: análisis intrínseco

En capítulos anteriores hemos descrito el proceso de creación del C-test aplicado en este trabajo empírico. La fase de diseño es fundamental, puesto que, en realidad, supone el comienzo de la validación de la prueba (Messick 1989, 1996). Siguiendo los consejos de Klein-Braley y Raatz, aunque adaptándolos a nuestras necesidades, como producto final llegamos a dos versiones (modelos A y B) de un C-test de cien omisiones, formado por cuatro subtests de veinticinco omisiones cada uno (véase el capítulo 8).

El análisis de las características del C-test aplicado es el paso previo para determinar, a partir de su funcionamiento en la práctica, la consistencia de la prueba. Analizaremos los resultados obtenidos por los sujetos en cada uno de los modelos y subtests. Para explicar las posibles diferencias entre ellos hemos de atender a las variables:

- Tema de los textos.
- Rasgos de los textos: grado de dificultad.
- Orden de aparición en el C-test.
- Formato aplicado: con o sin pistas acerca del número de letras omitidas.

9.3.1. Promedios del C-test y los subtests que lo forman

Para comenzar nos centraremos en los promedios obtenidos en el C-test (100 omisiones) y en cada una de las partes o subtests en que podemos dividirlo. En este punto del análisis no tendremos en cuenta la división en modelos A y B.

Queremos comprobar la consistencia interna de la prueba en conjunto, como un todo. Después veremos también las correlaciones entre los distintos subtests que lo forman y las de cada uno de ellos con el total del C-test.

En la tabla de los promedios que aparece a continuación (Tabla 9.1) hemos llamado CTESTTOTAL a los resultados totales del C-test (100 omisiones) y a los subtests: CTEST1 (omisiones 1-25), CTEST2 (omisiones 26-50), CTEST3 (omisiones 51-75) y CTEST4 (de la 76 a la 100), para su mejor identificación.

Tabla 9.1. Promedios y desviación típica del C-test y los subtests que lo forman

	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

Nota: N = 162
(Los valores correspondientes a los subtests 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)

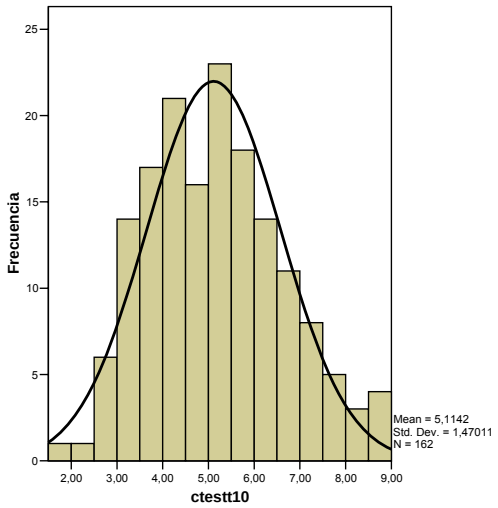
Sin tomar en consideración las posibles diferencias entre los dos modelos aplicados, en conjunto, vemos que la media obtenida por el C-test es de 51,12 puntos en escala de 0 a 100. Esta puntuación está justo en el punto que Klein-Braley y Raatz (1984: 144) y Klein-Braley (1984: 98) consideran adecuado para el C-test como prueba de tipo normativo que garantice la discriminación entre el alumnado:

> The C-test is a norm-oriented test, and as such, aims at medium level of difficulty on average. The mean score should be around 50% in order to ensure maximum differentiation between subjects. [...] If necessary we can afford to let the mean difficulty slide up to 60%. (Klein-Braley 1984: 98)

El C-test que diseñamos y aplicamos cumple el requisito; su grado de dificultad es medio, de este modo aseguramos la diferenciación entre sujetos. También los promedios de los subtests se encuentran en ese punto adecuado de dificultad.

Podemos ver que el histograma del C-test (Ctesttot10) en escala de 0 a 10 (Figura 9.1) presenta una distribución de frecuencias normal. Ningún alumno consiguió alcanzar la puntuación máxima, como tampoco ningún sujeto dejó la prueba en blanco. La desviación estándar es de 1,47.

Figura 9.1. Histograma del C-test (en escala de 0 a 10)



Pasemos a un análisis pormenorizado de los promedios obtenidos en los subtests de la prueba. La Tabla 9.1 muestra que la media obtenida en el C-test 1 (13,26 puntos de un total de 25) es superada en más de 2 puntos por el C-test 2 (15,64). La explicación parece sencilla: los sujetos acababan de conocer la técnica en el C-test 1 y la familiarización por la práctica produjo una sensible mejora de los resultados en el C-test 2. Sin embargo, en los dos subtests restantes observamos el efecto contrario, un decremento notable. La práctica ya no redunda en la mejora de los resultados de los C-tests 3 y 4.

Debemos recordar que en el subtest 3 se introduce una nueva variable; el cambio en el formato de la prueba al introducir las omisiones no guiadas, hecho que, a la luz de los resultados (10,15 puntos de promedio), contribuyó a aumentar el grado de dificultad, tal y como preveía la hipótesis 4.

El análisis de los histogramas de los subtests, sin desglosar en modelos A y B, (Fig. 9.2, 9.3, 9.4 y 9.5) indica que los resultados de los cuatro subtests presentan una distribución bastante normal y equilibrada, especialmente en los C-tests 1 y 4.

El C-test 1 refleja una distribución bimodal manifiesta, es decir, hay dos grupos homogéneos en puntuaciones por encima y por debajo de la media, elemento de información muy importante si se planteara el C-test como elemento de trabajo en el aula más que como prueba de evaluación, ya que nos indicaría que tendríamos que tener en cuenta los dos niveles en nuestra docencia.

El histograma del C-test 2 refleja una distribución normal, con un muy ligero sesgo negativo. En el histograma del C-test 3 la distribución es también normal y no se aprecian grupos heterogéneos, ya que se observa una curva de sesgo positivo, lo que pone de manifiesto que este subtest era más difícil que los demás.

Figura 9.2. Histograma del C-test 1

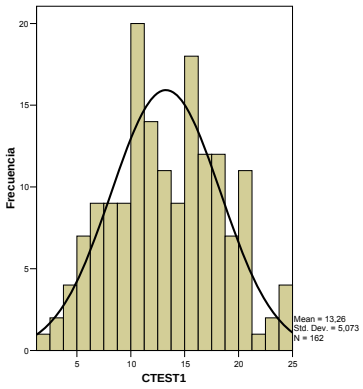


Figura 9.3. Histograma del C-test 2

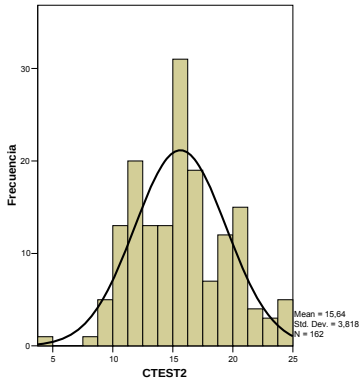


Figura 9.4. Histograma del C-test 3

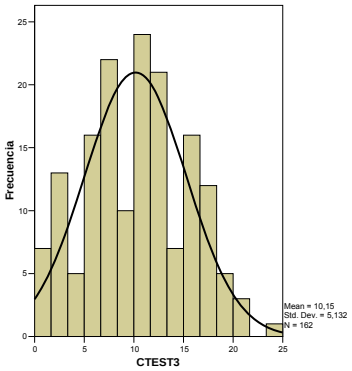
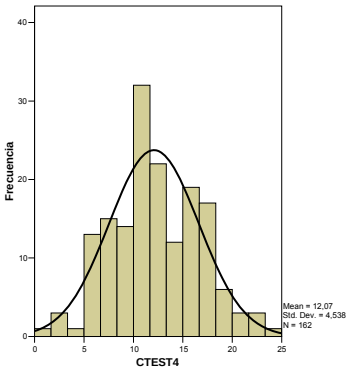


Figura 9.5. Histograma del C-test 4



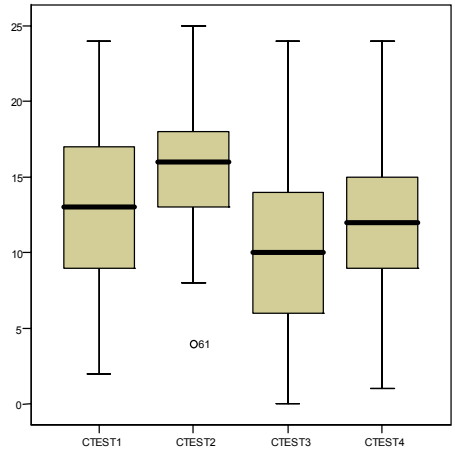
La cuarta pregunta de investigación: "¿Incide el formato utilizado en la recuperación de las omisiones?" nos cuestionaba acerca del formato del C-test en relación directa con la hipótesis 4:

> "Los cambios en el formato influyen directamente en los resultados obtenidos; si se incluye el número de letras que corresponde a cada omisión se facilita la tarea del alumno".

Aunque profundizaremos en ello en el apartado 9.3.3, adelantamos algunos aspectos que llaman nuestra atención al analizar los estadísticos (tablas y gráficos).

En los C-tests 1 y 2 las omisiones eran guiadas (con mención expresa del número de letras omitidas en cada ítem), mientras que esta ayuda desaparece en los C-tests 3 y 4. El diagrama de cajas (Figura 9.6) refleja gráficamente los promedios obtenidos por cada subtest y permite una visión comparativa de conjunto. La linea negra que aparece en cada caja, y que corresponde a la mediana, presenta unos valores semejantes a la media. Se aprecia más dispersión de puntuaciones en los C-tests 1 y 3. En el subtest 2 el diagrama de cajas detecta la presencia de un *outlier* o valor extremo, cuya puntuación es anormalmente baja.

Figura 9.6. Diagrama de cajas de los promedios obtenidos en los subtests



A continuación observamos en las tablas 9.2 y el diagrama de cajas (Fig. 9.7) la comparación de los promedios obtenidos en cada una de las dos mitades del C-test en escala de 0 a 50, para ello agrupamos las omisiones 1-50 correspondientes a los subtests 1 y 2 (guiadas) y las de los subtests 3 y 4 (omisiones 51-100 no guiadas). Queda patente que en las omisiones no guiadas los resultados descienden y hay diferencias significativas, con t= 10,894 y p<0001.

Tabla 9.2a. Media de los ítems guiados (C-tests 1 y 2) y no guiados (C-tests 3 y 4)

Estadísticos de muestras relacionadas

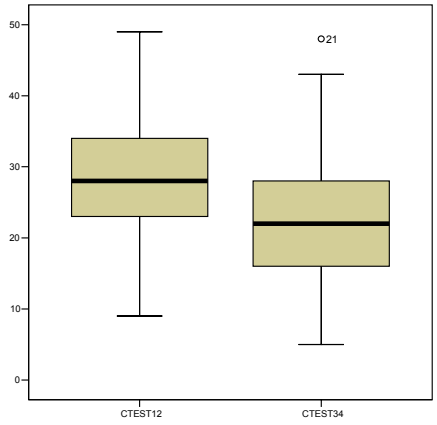
		Media	N	Desviación típ.
Par 1	CTEST12	28,90	162	7,910
	CTEST34	22,23	162	8,693

CTEST12 aglutina los resultados de los subtests 1 y 2 con omisiones guiadas
CTEST34 aglutina los resultados de los subtests 3 y 4 con omisiones no guiadas

Tabla 9.2b. Prueba de muestras relacionadas

		Diferencias relacionadas					t	Gl	Sig. (bilateral)
		Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	CTEST12 - CTEST34	6,667	7,789	,612	5,458	7,875	10,894	161	,000

Figura 9.7 .Diagrama de cajas de los promedios obtenidos en los subtests 1-2 y 3-4



La media de los subtests con omisiones guiadas es de 28,91 puntos en escala de 0 a 50, superior a la de los subtests con omisiones no guiadas (22,25 puntos). También la dispersión de puntuaciones, expresada en la desviación típica (8,675 frente a 7,943 puntos) es mayor cuando no se dan pistas. En los subtests no guiados aparece otra vez un caso extremo. Una vez revisados los datos, comprobamos que corresponde a un sujeto bilingüe, cuya puntuación supera ampliamente a la del resto de la muestra.

El propósito del cambio de formato era lograr que la prueba aumentara su dificultad progresivamente a medida que el alumno se iba familiarizando con la técnica, como recomienda Klein-Braley (1985, 1997), puesto que, por su procedencia, el grado de dificultad de los textos debía ser similar. Además, este nuevo formato plantea un reto al alumno (Babaii y Ansary 2001) y contribuye a que no pierda interés a medida que conoce la técnica.

A pesar de todo, las instrucciones informaban al alumno de que en todo caso las omisiones corresponden siempre a la segunda "mitad" de la palabra.

Nuestro interés se centra en estudiar cómo afecta este mínimo cambio a la actuación de los alumnos (hipótesis 4). A la vista de los resultados, podemos decir que, como se esperaba, supuso realmente una dificultad añadida[67]. Las opiniones de los alumnos, recogidas en el cuestionario retrospectivo, respaldan esta idea (véase el apartado 10.3 del capítulo 10). Pero sería demasiado aventurado achacar exclusivamente a este factor el descenso en la media. En los siguientes apartados veremos que hemos de atender también a la posible incidencia de otras variables, como las características de los textos.

9.3.2. Correlaciones entre el C-test y los subtests que lo forman

Una vez examinados los promedios, pasamos a analizar las correlaciones entre los subtests y los resultados totales del C-test, para comprobar la coherencia interna de la prueba. Mediante el análisis determinaremos la existencia de relación entre las variables, aunque no las causas de ésta (Mackey y Gass 2005).

La Tabla 9.3 muestra las correlaciones de Pearson. Se denominan "*product-moment correlations*".

Observamos cómo los resultados que plasma la tabla son coherentes con los expresados por los promedios y ya comentados. Todos los subtests correlacionan de forma significativa con el total del C-test, las correlaciones son muy altas en los C-tests 2 y 4 (0,841 y 0,877 respectivamente) y descienden algo en los C-tests 1 y 3 (0,727 y 0,741), aún siendo muy buenas. Vemos que el subtest 3 no correlaciona bien con el 1, pues se aprecia la correlación más baja (0,182), y con el 2, aunque mejora notablemente, el coeficiente (0,535) es inferior a los obtenidos por los otros subtests.

[67] El trabajo de Babaii y Moghaddam (2006), recientemente publicado, explora la incidencia del formato en el tipo de procesamiento lingüístico utilizado para resolver C-tests. Se aplicó el mismo C-test con dos versiones (con y sin omisiones guiadas) a dos grupos diferentes de sujetos y se comprobó que cuando no se aporta ayuda extra, los sujetos se esfuerzan más por utilizar "macro-level processing". Este hecho llevó a los autores a recomendar la creación de C-tests a partir de textos de cierta dificultad.

Tabla 9.3. Correlaciones de Pearson entre los resultados globales del C-test y de los distintos subtests

	1	2	3	4	5
1. CTEST1	--				
2. CTEST2	,574(**)	--			
3. CTEST3	**,182**(*)	,533(**)	--		
4. CTEST4	,544(**)	,636(**)	,615(**)	--	
5. TESTTOTAL	,727(**)	,841(**)	,741(**)	,877(**)	--

** La correlación es significativa al nivel 0,01 (bilateral).
* La correlación es significante al nivel 0,05 (bilateral).
N = 162

A pesar de la alta correlación entre el subtest 3 y el total del C-test (0,741), buscaremos evidencias de alguna variable en este subtest (C-test 3) que, junto al cambio de formato, contribuya a bajar los promedios.

El descenso en las puntuaciones no puede achacarse al tema del texto porque no es el mismo en los dos modelos y, por los resultados y correlaciones veremos que un mismo texto (i.e. *Road accidents*, subtest 1 en el modelo A y subtest 3 en el B) funciona de forma totalmente distinta cuando se presenta en otro orden. Concretamente, cuando esto supone la introducción de las omisiones no guiadas (véase el apartado 9.3.2).

En primera instancia, en el contexto del C-test aplicado, cabe pensar que, por el diseño de la prueba, el factor fundamental ha sido el cambio del formato, planteado en la hipótesis 4. Sin embargo, el C-test 4, que mantiene el formato de omisiones no guiadas, presenta la correlación más alta con el total del C-test, y es también el que mejor explica la varianza de la variable dependiente en el análisis de regresión lineal, como se verá en el capítulo siguiente.

Este dato sugiere que, además del aprendizaje y familiarización, existen variables textuales en el C-test 3 que lo diferencian del resto. Pensamos en las características de los textos de partida: tema, variación léxica y densidad. Hacia estos aspectos dirigiremos nuestra atención en el apartado 9.3.4 de este capítulo.

Hasta ahora se ha estudiado la relación entre el C-test en conjunto y los cuatro subtests que lo forman. Hemos visto el funcionamiento de cada subtest a partir de los promedios y que el subtest 3, con el descenso en las puntuaciones, marca un punto de inflexión en la prueba, coincidente con la introducción del cambio de

formato, pero que no puede achacarse exclusivamente a él. En los apartados siguientes concretaremos la incidencia del cambio de formato, revisaremos las características de los textos del C-test y definiremos cómo afecta a los resultados el modelo de C-test aplicado (A y B).

El apartado 9.3.4 analiza las características de los cuatro textos a partir de los cuales se elaboró el C-test en ambos modelos. Si bien *a priori* se había considerado que su grado de dificultad era similar, ahora se impone una revisión más profunda de sus rasgos. Quizá sus peculiaridades contribuyan a explicar los datos estadísticos obtenidos por el C-test 3.

El siguiente paso exige aplicar el análisis de la regresión lineal del C-test, que nos permite explorar y cuantificar la relación entre la variable dependiente, el C-test total, y las independientes, los distintos subtests. Con este tipo de análisis, al que dedicamos la parte final del capítulo, completaremos nuestra investigación.

9.3.3. Resultados obtenidos según el modelo de C-test: A y B

Como hemos visto en la Metodología (capítulo 8), se aplicaron dos modelos diferentes de C-test con idénticos textos y mutilaciones, pero cambiando el orden de presentación de los mismos. En ambos modelos se aportaban pistas sobre el número de letras omitidas en los subtests 1 y 2 (cincuenta primeras omisiones). Fueron repartidos al azar. Exactamente la mitad de los sujetos de la muestra completó el C-test modelo A y la otra mitad el B. Téngase presente en el análisis la Figura 8.3 del capítulo 8, que refleja la estructura de la prueba (textos y formato aplicado).

Basándonos en las puntuaciones observadas en los datos totales del C-test en una escala de puntuaciones del 0 al 100 (Tabla 9.5), que hemos denominado CTESTTOTAL, ya desglosados en ambos modelos, podemos comprobar que se consiguió mejor promedio en el modelo A (media = 51,84), lo cual indica que el modelo B (media = 50,41) resultó más difícil, si bien la diferencia es mínima. Ambos modelos entran en el rango de puntuaciones previsto por Klein-Braley y Raatz (1984) y citado en el apartado anterior.

Tabla 9.5. Estadísticos de grupo: modelos A y B del C-test

Modelo C-test		Media	Desviación típ.	Error típ. De la media
Modelo A	CTESTTOTAL	51,84	14,980	1,664
Modelo B	CTESTTOTAL	50,41	14,438	1,604

N = 81

Hagamos también este tipo de análisis para cada uno de los subtests En la Tabla 9.6 vemos los promedios obtenidos por cada subtest en ambos modelos, esta vez en una escala de puntuaciones de 0 a 25 puntos:

Tabla 9.6. Estadísticos de grupo: modelos A y B de los distintos subtests

Modelo C-test		Mínimo	Máximo	Media	Desv. Típ.
Modelo A	CTEST1	5	24	16,15	4,126
	CTEST2	9	25	16,12	3,858
	CTEST3	0	20	7,47	4,799
	CTEST4	2	22	12,10	4,437
Modelo B	CTEST1	2	21	10,37	4,226
	CTEST2	4	22	15,15	3,739
	CTEST3	5	24	12,84	3,923
	CTEST4	1	24	12,05	4,663

N = 81

En todos los casos, excepto en el C-test 3, se consiguió mejor promedio en el modelo A. En el caso del C-test 1 se aprecia la mayor diferencia de promedios entre el modelo A y el B (de 16,15 a 10,37 puntos), y en el C-test 4 la menor (de 12,10 a 12,05 puntos).

El C-test 3 de nuevo llama nuestra atención, puesto que presenta un comportamiento totalmente distinto. En este caso es el modelo B el que mejor funciona, con una diferencia de más de 5 puntos (promedio de 12,84 puntos en el modelo B frente a los 7,47 del modelo A).

Esta diferencia de promedios obtenidos por el C-test 3 en uno y otro modelo muestra que las diferencias se deben más a los textos que al cambio de formato.

Los promedios indican que las mayores dificultades aparecieron en el texto *American imperialism* cuando no tiene omisiones guiadas (modelo A). Sin embargo,

cabe destacar que cuando las tiene, aunque el promedio mejora sensiblemente (media con omisiones guiadas = 10,37; media sin pistas en las omisiones = 7,47), sigue siendo el más bajo de los obtenidos en los subtests con omisiones guiadas.

Además, hemos de notar que la diferencia de promedios mencionada para el C-test 1 se produce con el mismo texto en el modelo B: *American imperialism*. Sin embargo, la mayor puntuación en subtests guiados corresponde al C-test 1 del modelo A (16,15) y en los no guiados al C-test 3 del B (12,84), ambos subtests creados a partir del texto *Road accidents*. La incidencia de los textos es ya evidente.

Partiendo de estos datos profundizaremos en el análisis textual para buscar las causas de las particularidades detectadas en las puntuaciones. Nos planteamos varias cuestiones que iremos retomando a lo largo de nuestro análisis:

- ¿qué características de los textos determinan las diferencias en los promedios?,
- el tema de los textos, su variación léxica o su densidad, ¿determinan el grado de dificultad del C-test?, ¿son elementos que discriminan la competencia de los sujetos? (véase la pregunta de investigación 7)

El C-test 3 del modelo A se basa en el texto *American imperialism* y el del modelo B en *Road accidents*. Si nos fijamos en los promedios obtenidos por ambos textos cuando se indica el número de letras omitidas en cada mutilación (Tabla 9.7), veremos que, efectivamente, se obtienen puntuaciones más elevadas. Pero también se observa que uno de los textos, *Road accidents,* resultó más fácil que el otro, independientemente del formato aplicado.

Tabla 9.7. Diferencia de promedios según el formato de las omisiones

Texto base	Formato	Media	Texto base	Formato	Media
Road accidents	- - - - -	16,15	American imperialism	- - - - -	10,37
Road accidents	_____	12,84	American imperialism	_____	7,47

Por otra parte, en el modelo B se produce un aumento de la media en el C-test 2 con respecto al C-test 1 (Tabla 9.6). En el texto *Women doctors* se alcanza un

promedio de 15,15 puntos frente a 10,37 en *American imperialism*. En el caso del modelo A, los promedios de los subtests 1 y 2 son muy semejantes (16,15 y 16,12 respectivamente).

Sin embargo, sí es importante el decremento de la puntuación en el C-test 3, circunstancia común a ambos modelos pero especialmente significativa en el A. Ya hemos comentado que un ligero decremento podría explicarse por el propio diseño de la prueba, que guía las omisiones 1-50 (C-test 1 y 2), lo que contribuye a facilitar la tarea del alumno. Cuando se deja de aportar la guía la dificultad aumenta, a pesar de que la práctica y el conocimiento progresivo de la prueba deberían también traducirse en una mayor destreza. En el modelo A, la media baja en el C-test 3 de forma mucho más llamativa. Por tanto, como hemos apuntado, cabe buscar los motivos del descenso de promedios en los rasgos del texto sobre el que se ha creado el C-test 3 del modelo A: *American imperialism*. En el apartado 9.4 analizaremos la incidencia de las variables tema, variación léxica y densidad.

9.3.4. Incidencia del cambio de formato

En el diseño definitivo de la prueba se introdujo un cambio en el formato para que se produjera un progresivo aumento del grado de dificultad de los subtests a medida que el alumno se iba familiarizando con la técnica (omisiones 50 a 100). Klein-Braley propone hacerlo simplemente de forma intuitiva, atendiendo a la dificultad de los textos. Pero en nuestro diseño, al pasar de omisiones guiadas a no guiadas, la dificultad se guía por parámetros más objetivos.

El diseño original del C-test que proponen Klein-Braley y Raatz no indica el número de letras correspondientes a las omisiones, aunque las instrucciones informan al alumno de que se omite “la segunda mitad” de cada palabra.

En apartados anteriores hemos adelantado que, según los datos estadísticos de nuestro estudio, el cambio de formato influye en los promedios de todos los subtests, cualquiera que sea el texto sobre el que se crearon. Los promedios descienden en ambos modelos según se trate o no de omisiones guiadas. Así contestamos a la pregunta de investigación 4, sobre la incidencia del formato y confirmamos la hipótesis 4.

Hemos comprobado que en el C-test 3 se produce siempre un punto de inflexión y las medias descienden en los dos modelos con respecto al C-test 2. La tabla siguiente muestra las medias obtenidas para cada subtest con o sin omisiones guiadas.

Tabla 9.8. Comparación de las 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

Además de los promedios, es interesante ver gráficamente en los histogramas cómo varía la recuperación de un mismo texto, según se planteen omisiones guiadas o no en el C-test.

A continuación aportamos ambos histogramas de *American imperialism,* en los que no sólo se reflejan las diferencias en los promedios, sino que también se aprecia claramente cómo cambia la distribución de puntuaciones.

Cuando no se aporta el número de letras de cada omisión (AmImp.NG) obtenemos una curva sesgada positivamente, es decir, que las frecuencias más altas corresponden a los valores más bajos de la tabla (Fig. 9.8a: modelo A, C-test 3). La mayor parte de los sujetos de la muestra obtuvo una puntuación de menos de 7 puntos en escala de 0 a 25, y la puntuación máxima fue de 20 puntos.

Por el contrario, en la figura 9.8b (AmImp.G), vemos que cuando se guían las omisiones (modelo B, C-test 1) se obtiene una distribución de frecuencias mucho más normal, reflejada en la curva. También es menor la desviación estándar: 4,22 frente a 4,79. En este caso, el C-test discrimina mejor entre los sujetos.

Fig. 9.8a. y 9.8b. *American Imperialism* con omisiones no guiadas y guiadas: Histogramas

Fig. 9.8a. Omisiones no guiadas

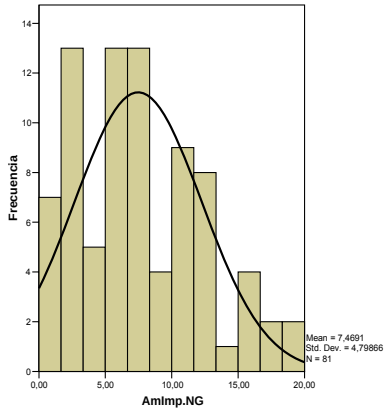
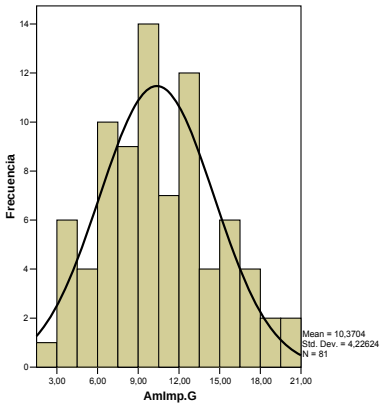


Fig. 9.8b. Omisiones guiadas



9.3.4.1. El cambio de formato y la recuperación de algunos ítems

Veamos a continuación cómo incide el formato en la recuperación de los ítems con el fin de confirmar definitivamente la hipótesis 4.

Para determinar las pautas de la incidencia del cambio de formato nos basamos en algunos ejemplos indicativos. Tomamos simplemente los estadísticos de frecuencias de tres ítems de carácter léxico y uno funcional de los C-tests modelos A y B creados a partir del texto *American Imperialism*. Indican el porcentaje de aciertos y fallos en cada ítem y los valores perdidos. Para cada modelo el total de sujetos de la muestra es de 81.

Comenzamos con los ítems 1 y 2 del modelo B (omisiones guiadas), que se corresponden con los ítems 51 y 52 del modelo A (en este caso sin indicación expresa del número de letras omitidas).

Tabla 9.9. Ítem 1 mod. B: hun_ _ _ (*hunger*)

		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 9.10. Ítem 51 mod. A: hun_____ (*hunger*)

		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		

Los estadísticos de frecuencias indican que la recuperación del ítem *hunger* (término léxico) es más fácil si se aporta el número de letras omitidas; un 29,6% de los sujetos lo resuelven correctamente frente al 18,5% de aciertos cuando no se da la información. Además aumenta notablemente el porcentaje de sujetos que dejan el ítem sin hacer; un 42%. Este dato llama poderosamente nuestra atención porque muestra que prescindir del número de letras restantes produce un efecto psicológico de desánimo en los sujetos, muchos de los cuales ni siquiera intentan resolverlo.

En cuanto a la recuperación del ítem 2 del modelo B, poverty, también un término de tipo léxico, vemos que de nuevo el porcentaje de aciertos es mayor cuando se dispone del número de letras omitidas; 49,4% frente al 42%, pero en este caso la diferencia es menor. Al igual que ocurría con los ítems 1(B) y 51(A), analizados anteriormente, aumenta el porcentaje de sujetos que abandonan la tarea al tener menos información, aunque ese efecto que hemos denominado de "desánimo" no es tan acusado en esta ocasión.

Tabla 9.11. Ítem 2 mod. B: pov _ _ _ _ (*poverty*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	8	9,9	16,7	16,7
	Correcto	40	49,4	83,3	100,0
	Total	48	59,3	100,0	
Perdidos	sin hacer	33	40,7		
Total		81	100,0		

Tabla 9.12. Ítem 52 mod. A: pov _____ (*poverty*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	8	9,9	19,0	19,0
	Correcto	34	42,0	81,0	100,0
	Total	42	51,9	100,0	
Perdidos	sin hacer	39	48,1		
Total		81	100,0		

Las mismas pautas se siguen en los ítems 21(B) y 71(A): mejor recuperación de las omisiones y mayor proporción de intentos cuando se trata de omisiones guiadas y se dispone de más información sobre los ítems.

En este caso se trata de un término muy frecuente en la lengua, factor que contribuye a su recuperación correcta.

Tabla 9.13. Ítem 21 mod. B: pe _ _ _ (*peace*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	2	2,5	4,3	4,3
	Correcto	45	55,6	95,7	100,0
	Total	47	58,0	100,0	
Perdidos	sin hacer	34	42,0		
Total		81	100,0		

Tabla 9.14. Ítem 71 mod. A: pe____ *(peace*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	10	12,3	26,3	26,3
	Correcto	28	34,6	73,7	100,0
	Total	38	46,9	100,0	
Perdidos	sin hacer	43	53,1		
Total		81	100,0		

Por último, vemos la recuperación de un término funcional: *and*, con omisiones guiadas o no, para corroborar que las frecuencias siguen el mismo patrón de comportamiento. Sin duda, este tipo de términos funcionales cortos, frecuentes y muy repetidos en la lengua, animan al sujeto en la realización de la prueba. Son factores de motivación más que de discriminación, por eso, al contrario que Babaii y Moghaddam (2006), consideramos que deben mantenerse en el C-test.

Tabla 9.15. Ítem 15 mod. B: a _ _ (*and*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	18	22,2	24,7	24,7
	Correcto	55	67,9	75,3	100,0
	Total	73	90,1	100,0	
Perdidos	sin hacer	8	9,9		
Total		81	100,0		

Tabla 9.16. Ítem 65 mod. A: a___ (*and*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	23	28,4	39,7	39,7
	Correcto	35	43,2	60,3	100,0
	Total	58	71,6	100,0	
Perdidos	sin hacer	23	28,4		
Total		81	100,0		

Todos los ejemplos analizados están tomados del texto *American Imperialism* (C-test 1 en el modelo B y C-test 3 en el A), que es el que presenta las menores diferencias en los promedios con/sin omisiones guiadas, como vimos en la Tabla 9.8 que inicia este apartado. Pero también es el texto que obtiene los menores promedios y, por tanto, merece un análisis más profundo.

Podríamos seguir con el análisis detallado de cada ítem del C-test para constatar estos hechos y reafirmar así nuestra hipótesis de partida: el diseño que incluye los espacios de las omisiones constituye una ayuda eficaz, ya que suministra un elemento más de información que permite al alumno potenciar y desarrollar el

grado de inferencia. Circunstancia que supone un incremento o decremento de varios puntos en los promedios, según se trate de omisiones guiadas o no.

A esto hemos de añadir que la ayuda extra de conocer de forma expresa el número de letras de cada omisión supone también un elemento clave de motivación para el sujeto que realiza la prueba, anima a intentar la recuperación de la palabra, como prueban los estadísticos de frecuencias de los ítems y manifiestan los propios alumnos en el cuestionario retrospectivo de opinión (véase el capítulo 12).

9.4. Análisis de los textos a partir de los cuales se creó el C-test aplicado

Aunque todos los textos proceden de exámenes de Inglés de las PAAU propuestas por la Consejería de Educación de la CM para el acceso a las Universidades madrileñas, y por tanto su nivel se presuponía semejante, los resultados de los subtests indican claramente que no es así.

Procede, así pues, analizar las características de los textos a partir de los cuales se diseñó el C-test. En la longitud de las oraciones de cada texto, rasgo que podría considerarse, no parece existir diferencias notables, y por tanto, queda fuera de este análisis. Tampoco el orden de presentación de los textos tiene gran incidencia en los resultados, excepto porque supone además el cambio en el formato de las omisiones.

En este apartado estudiaremos cómo las variables tema, variación y densidad léxicas del texto influyen en los resultados del C-test.

9.4.1. La variable temática

En cuanto a la variable temática, hemos de decir que todos los textos que aparecen en el C-test son de carácter divulgativo o periodístico, tratan temas de actualidad e interés general y, en este sentido, pueden considerarse semejantes.

Sasaki (2000) mostró en su investigación cómo la familiarización con el vocabulario de los textos activa los esquemas que se utilizan en la resolución de pruebas de cierre. Los sujetos de su estudio encontraron en la familiarización

cultural un factor de motivación añadido. De ello, podemos deducir que el tema (culturalmente familiar o no) incide tanto en la comprensión de los textos como en la motivación que aporta a los sujetos para intentar recuperar el texto original.

En nuestro caso, quizá los alumnos pudieran estar más familiarizados con unos temas que con otros, dependiendo de su conocimiento del mundo y sus preferencias o circunstancias personales, pero en primera instancia no parece que la variable temática pueda tener gran incidencia en la actuación en la prueba. Aún así, el tema que plantea *American imperialism* es de un registro menos común, político-histórico, presenta una realidad algo más lejana al alumno que los otros textos:

- *Road accidents* plantea el problema de los accidentes de tráfico en Francia y las posibles soluciones,
- *Evolution* trata de forma pseudocientífica la relación entre los primates y el ser humano, en el contexto de la evolución de la especie humana,
- *Women doctors* introduce el tema de la mujer en el mundo del trabajo, el cambio progresivo de los roles de hombres y mujeres, en concreto, en la Medicina.

Los tres temas anteriormente citados pueden considerarse familiares y relativamente cercanos a la realidad que vivimos. *American Imperialism* plantea un tema relacionado con la política internacional y las desigualdades entre países ricos y pobres, desde la perspectiva de los Estados Unidos. Es también un tema frecuente, incluso recurrente en los medios de comunicación en la actualidad. Pero pertenece a la esfera del pensamiento político-social y tiene mayor grado de abstracción que los otros. Es probable, además, que los asuntos relacionados con la política resulten distantes y no capten el interés de los adolescentes.

Quizá esta diferencia temática planteara una primera dificultad a los sujetos, no obstante, veremos en el epígrafe siguiente que la densidad y variación léxicas del texto también deben ser tenidas en cuenta como factores determinantes del grado de dificultad de los subtests.

9.4.2. Variación y densidad léxicas de los textos

En los C-tests creados a partir del texto *Road Accidents* se consiguen las puntuaciones más altas, tanto con omisiones guiadas como no guiadas. Por el contrario, *American Imperialism* es, según los resultados, el texto más difícil de recuperar de los cuatro que forman el C-test, como vemos en los promedios obtenidos por el C-test 1 modelo B y el C-test 3 modelo A (Tablas 9.6, 9.7 y 9.8). Tratando de explicar mejor este factor de dificultad recurrimos a un análisis de la densidad y variación léxicas[68] de cada texto, como aconsejan Laufer y Nation (1995) y Schmitt (2000: 75). Éstos son los resultados:

Tabla 9.17. Variación y densidad léxica de los textos

TEXTOS	VARIACIÓN LÉXICA	DENSIDAD LÉXICA
Road Acccidents	69,47	58,94
Evolution	63,46	45,19
American Imperialism	**70,37**	**60,18**
Women Doctors	69,14	58,51

La lectura de estos datos refleja que los valores obtenidos son muy semejantes, tanto desde el punto de vista de la variación como de la densidad léxica. Pero, efectivamente, en un continuo de facilidad/dificultad *Evolution* y *American imperialism* aparecerían en uno y otro extremo respectivamente. Las diferencias de los porcentajes en ambos textos podrían explicarse en función de la mayor o menor prototipicidad del léxico.

El texto *Road accidents* obtiene los promedios más altos tanto en omisiones guiadas como no guiadas, aunque muy cercanos a los de *Evolution* y *Women doctors*, a pesar de no ser el más fácil, si atendemos a los datos de variación y densidad léxicas.

[68] La variación léxica viene dada por la proporción entre *types* y *tokens* (*type-token ratio*). La proporción entre palabras funcionales y de contenido léxico indica la densidad léxica de un texto (véase el capítulo 4, apartados 4.2.3.1 y 4.2.3.2).

Los datos obtenidos al analizar la variación y densidad léxicas de los textos no parecen aportar suficiente información como para justificar las diferencias en los promedios. Así pues, en el apartado 9.5 daremos un paso más en el análisis, dirigiremos nuestra atención al grado de dificultad de los ítems más que al de los textos. Estudiaremos las características de los términos afectados por la mutilación en cada texto. Veremos cómo el tipo de término, léxico o funcional, afecta a su recuperación correcta en el C-test.

9.5. Factores que determinan la facilidad o dificultad de los ítems

La mera observación de un C-test pone de manifiesto que no todas las omisiones son iguales y, por tanto, no se recuperan de la misma forma. El análisis pormenorizado prueba que no todos los ítems del C-test presentan el mismo grado de dificultad, tal y como predecíamos en la hipótesis 3, en torno a la cual girará ahora nuestro trabajo.

En la Introducción de la tesis esbozamos la pregunta de investigación:

- ¿Incide el tipo de término, léxico o funcional, afectado por la mutilación en la recuperación de las omisiones?

Este apartado aborda el grado de dificultad de las omisiones, sus causas y posibles consecuencias para la prueba.

9.5.1. Términos léxicos y funcionales

Laufer (1997) apuntó un buen número de factores que inciden en el grado de dificultad de las palabras. En este caso nos ocupamos de la categoría gramatical. Pero la autora menciona también la pronunciación, la correspondencia pronunciación-ortografía, la longitud de la palabra, su morfología (irregularidades en la flexión, derivación, etc.), la semejanza entre formas léxicas en L1 y L2, aspectos semánticos como el grado de abstracción, el registro lingüístico (frecuencia), la idiomaticidad y la multiplicidad de significados.

Aunque en principio se puede considerar que los nombres son la categoría gramatical más sencilla, frente a la dificultad de los adverbios (Laufer 1997), con respecto a su recuperación en las pruebas de cierre, la literatura indica que, en general, los términos funcionales resultan más fáciles que los léxicos, debido a su frecuencia de uso y su cantidad limitada en la lengua (Klein-Braley 1985: 91, Dörnyei y Katona 1992: 197, Farhady y Keramati 1996:196).

Partimos de esa base, aunque veremos que a veces se entrecruzan varios factores. Analizaremos algunos casos concretos y veremos los resultados estadísticos alcanzados por cada tipo de término en ambos modelos de C-test.

Por lógica, hemos de pensar que los ítems excesivamente fáciles o difíciles no discriminan entre los sujetos. Babaii y Moghaddam (2006: 596) proponen, por ejemplo, que los términos repetidos queden intactos. Jafarpur (1999), en la línea de Grotjahn (1987) y Kamimoto (1993), planteó suprimirlos para obtener C-tests racionales, sin seguir siempre la *regla del dos*. Pero comprobó que su eliminación no cambia los resultados (véase el apartado 6.10.2 del capítulo 6).

Nuestro análisis de estos valores extremos por su facilidad o dificultad confirma la tesis de Jafarpur. Si se toman los ítems que prácticamente recuperan con éxito todos los alumnos, por ejemplo, 7, 8, 12 y 32 del modelo A, se observa que la mayoría corresponden a preposiciones (*in, to*), términos de uso muy frecuente en la lengua, funcionales y de una exigencia mínima del nivel de inferencia ya que el alumno tan sólo tiene que recuperar una letra.

En el caso del ítem *animals* se trata de un término léxico que aparece previamente en el texto (redundancia) y, por tratarse de un hiperónimo del inglés básico de gran frecuencia en la lengua, su recuperación no ofrece problemas. Por otra parte, términos como *harm* (ítem 58 del modelo A) resultan excesivamente difíciles.

Ni los ítems que resuelven la mayoría de los alumnos ni los que nadie consigue recuperar discriminan entre los sujetos. Ahora bien, los fáciles (incluso los repetidos) aportan a la prueba un factor extra de motivación que es también interesante para la validez aparente de la prueba, por ello, al contrario de lo propuesto inicialmente por Jafarpur (1999), no consideramos procedente su eliminación de la prueba.

El estudio de Dörnyei y Katona (1992: 198) con alumnos de Universidad y de Enseñanza Secundaria plantea también que a mayor competencia en la lengua

mejores predictores del comportamiento del sujeto son las omisiones de términos léxicos. Por el contrario, en niveles más bajos los términos funcionales son los que mejor discriminan. Al limitar nuestra investigación a un solo nivel (2º de Bachillerato) no se indagó en este aspecto, que podría retomarse en futuras investigaciones.

Veamos ahora las tablas de frecuencia de algunos ítems de carácter léxico y funcional, a modo de ejemplo. Comenzamos con el ítem *to*, término funcional que exige la recuperación de sólo una letra y que, dado su escaso grado de dificultad, fue recuperado con éxito por el total de los sujetos de la muestra.

Muy cerca de este resultado, los obtenidos por los ítems de función *in* (8 y 23 del modelo A), *the* y el frecuente, aunque término léxico, *animals*, que tampoco aportan información para discriminar a los sujetos.

Tabla 9.18. Ítem 7 modelo A: t_ (*to*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	correcto	81	100,0	100,0	100,0

Tabla 9.19. Ítem 8 modelo A: i_ (*in*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	incorrecto	1	1,2	1,2	1,2
	correcto	80	98,8	98,8	100,0
	Total	81	100,0	100,0	

Tabla 9.20. Ítem 12 modelo A: i_ (*in*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	incorrecto	2	2,5	2,5	2,5
	correcto	79	97,5	97,5	100,0
	Total	81	100,0	100,0	

Tabla 9.21. Ítem 32 modelo A: ani_ _ _ _ (*animals*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	incorrecto	1	1,2	1,2	1,2
	correcto	80	98,8	98,8	100,0
	Total	81	100,0	100,0	

Tabla 9.22. Ítem 82 modelo A: t___ (*the*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	incorrecto	5	6,2	6,3	6,3
	correcto	74	91,4	93,7	100,0
	Total	79	97,5	100,0	
Perdidos	sin hacer	2	2,5		
Total		81	100,0		

En el otro extremo del continuo facilidad-dificultad se encuentran términos como los que presentamos a continuación. Los términos *harm* y *wave* resultaron difíciles, pocos alumnos los recuperan y muchos ni siquiera lo intentan.

Tabla 9.23. Ítem 58 modelo A: ha___ (*harm*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	incorrecto	42	51,9	95,5	95,5
	correcto	2	2,5	4,5	100,0
	Total	44	54,3	100,0	
Perdidos	sin hacer	37	45,7		
Total		81	100,0		

Tabla 9.24. Ítem 88 modelo A: Wa___ (*wave*)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	incorrecto	44	54,3	88,0	88,0
	correcto	6	7,4	12,0	100,0
	Total	50	61,7	100,0	
Perdidos	sin hacer	31	38,3		
Total		81	100,0		

Según estos datos, es evidente que no siempre los términos de función se recuperan mejor. Hay otros factores que determinan la facilidad o dificultad de un ítem, como la redundancia del texto en que se encuentra.

A pesar de todo, como veremos en los histogramas (Fig. 9.9, 9.10, 9.11 y 9.12) que reflejan la recuperación de ambos tipos de términos en los modelos A y B, en general, los términos funcionales se resuelven mejor. En el modelo A, el promedio para los funcionales es de 27 puntos frente a 24 en los léxicos. Además el histograma de los léxicos (Figura 9.10) muestra un sesgo positivo que indica que la mayor parte de las puntuaciones son bajas.

9.5.2. Incidencia del tipo de término omitido en la recuperación del texto. Análisis por modelos.

Para determinar cómo incide el tipo de término omitido en la recuperación de los textos, en esta investigación, al igual que en la primera prueba piloto (véase el capítulo 7, apartado 7.2), se realizó un análisis estadístico de la recuperación de las palabras con significado léxico y gramatical para cada modelo de C-test (A y B).

En la identificación de términos léxicos y funcionales de los textos seguimos de nuevo la clasificación de Quirk y Greenbaum (1973), Aarts y Aarts (1986), y Huddleston (1988), que coinciden con lo que Alarcos (1994) denomina términos dependientes e independientes. Tenemos presente que cabe un leve margen de error, ya que algunas palabras pueden adscribirse a una u otra clase (*multiple membership*).

9.5.2.1. Recuperación de omisiones de términos léxicos y funcionales

Puesto que la densidad léxica del texto indica la proporción de términos léxicos y funcionales (apartado 9.3.5), en primer lugar, confeccionamos un listado de las palabras de cada tipo afectadas por la mutilación. En la tabla siguiente mostramos los porcentajes correspondientes a los cuatro textos utilizados en el C-test:

Tabla 9.25. Porcentaje de omisiones en palabras léxicas y funcionales para cada texto

TEXTO	Nº DE PALABRAS AFECTADAS POR LA MUTILACIÓN DE CONTENIDO LÉXICO		FUNCIONALES	
Road Accidents	13	52%	12	48%
Evolution	10	40%	15	60%
American Imperialism	**19**	**76%**	**6**	**24%**
Women doctors	14	56%	11	44%

Llaman la atención los datos correspondientes al texto *American imperialism*. Cuando analizamos la densidad del texto, ya vimos que resultaba ligeramente superior a la de los demás (60,18), si bien la diferencia no era tan acusada como al tener en cuenta las palabras afectadas por la mutilación. En este caso, las palabras de contenido léxico mutiladas suponen el 76% del texto frente al 24% de los términos de función (debido al azar al seguir la regla del 2).

Esta circunstancia es clave para explicar la actuación de los alumnos. Sin duda, afecta a la resolución del C-test. No es casual que los C-tests creados a partir de este texto (con o sin omisiones guiadas) presenten los promedios más bajos: 10,37 y 7,47 respectivamente.

Así mismo, destaca el bajo porcentaje de palabras de contenido léxico que resultan mutiladas en el texto *Evolution,* sólo el 40%, frente al 52% en *Road accidents*. Este dato aporta luz a la cuestión planteada en el apartado 9.3.5 relativa al grado de facilidad/dificultad de los textos; no sólo inciden la densidad y variación léxicas de los textos, sino también la proporción de palabras léxicas y funcionales que resultan afectadas por la mutilación al aplicar la “regla del dos”.

Para analizar las diferencias en la recuperación de las palabras pertenecientes a clases cerradas (términos de función) y las que tienen carga semántica (términos léxicos), se hizo un análisis estadístico diferenciando entre los dos modelos de C-test aplicado (A y B).

A continuación, vemos los datos globales de la recuperación de términos gramaticales y léxicos en ambos modelos de C-test. En la tabla 9.26 quedan reflejados los estadísticos del modelo A, cuyo promedio (51,84 puntos de un total de

100) es superior al del modelo B (50,41). En este modelo el promedio de recuperación de términos funcionales (27,10) es mayor que el de léxicos (24,74).

Las diferencias entre los promedios de términos léxicos y de función son menores de lo que cabría esperar. Sin embargo, a partir del valor de la varianza se observa mayor dispersión de puntuaciones en los términos léxicos que en los funcionales, es decir, que los términos léxicos se recuperan muy bien o muy mal, según el dominio de la lengua que tienen los alumnos. La distribución de la asimetría no muestra diferencias notorias.

Tabla 9.26. Estadísticos modelo A

		CTESTTOTAL modelo A	T. Función	T. Léxicos
N	Válidos	81	81	81
	Perdidos	0	0	0
Media		51,84	27,10	24,74
Desv. típ.		14,98	6,741	8,819
Varianza		224,386	45,440	77,769
Asimetría		,404	,116	,506
Error típ. De asimetría		,267	,267	,267
Curtosis		-,497	-,548	-,312
Error típ. De curtosis		,529	,529	,529
Mínimo		27,00	13	10
Máximo		86,00	42	48

Tabla 9.27. Promedios de recuperación de cada tipo de término en el Modelo A

	Media	Desviación típ.	Error típ. de la media
CTESTTOTAL	51,84	14,980	1,664
T.Funcionales	27,10	6,741	,749
T.Léxicos	24,74	8,819	,980

N = 81

El histograma de la recuperación de términos funcionales del modelo A presenta una distribución de frecuencias normal (Fig. 9.9). Sin embargo el de los términos léxicos (Fig. 9.10) tiene un sesgo positivo (*positively skewed*), lo que indica que los ítems son más difíciles y en consecuencia, las puntuaciones más bajas.

Figura 9.9. Modelo A: términos de función

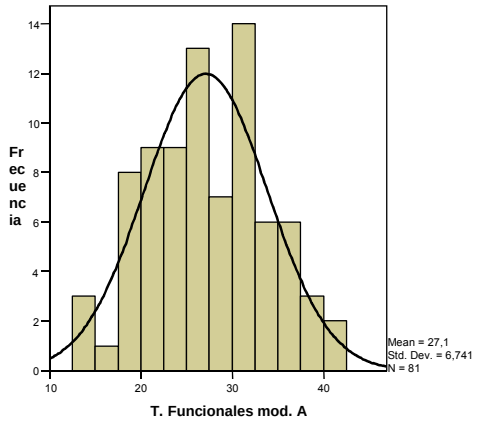
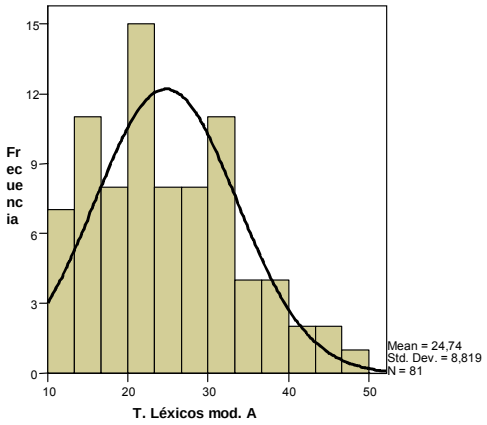


Figura 9.10. Modelo A: términos léxicos



El análisis de las correlaciones también aporta datos muy interesantes, puesto que muestra correlación "casi perfecta" entre los resultados obtenidos en el C-test modelo A y los de cada tipo de término. La correlación de los términos léxicos con el total del C-test modelo A es muy significativa (0,972). La de los términos funcionales también es muy alta (0,951), y la existente entre términos léxicos y funcionales del modelo A asciende a 0,851.

Estos resultados indican que las variables término léxico y funcional están asociadas significativamente. En la Tabla 9.28 vemos el test estadístico de muestras relacionadas, en el que hay diferencias significativas con p<0,001.

Tabla 9.28. Test de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
Modelo A				Inferior	Superior			
Par 1 T.Funcionales - T.Léxicos	2,358	4,694	,522	1,320	3,396	4,521	80	,000

Por el contrario, en el modelo B del C-test no hemos encontrado diferencias significativas entre la recuperación de términos léxicos y funcionales, con T=0,816 y p=0,417, como veremos en la Tabla 9.29.

En cuanto a los estadísticos, por agilidad, nos limitaremos a exponer los promedios totales y para cada tipo de término, que se reflejan en los histogramas correspondientes (Figuras 9.11 y 9.12).

Hemos visto en el apartado 9.3.2 que modelo B baja ligeramente el promedio global del C-test (50,41). Igual que en el A, se mantiene un mejor promedio en la recuperación de los términos gramaticales, pero la diferencia en este modelo es mínima (25,47 puntos de promedio frente a 24,94). Las varianzas indican también una mayor dispersión de puntuaciones en los términos léxicos (8,758 frente a 6,684).

De nuevo, las correlaciones entre la recuperación de términos léxicos y funcionales, y con el C-test modelo B en conjunto son significativas. Es especialmente alta en el caso de los términos léxicos (0,951). Podemos afirmar que las variables léxico y función están muy asociadas en el C-test modelo B. Los histogramas correspondientes lo reflejan claramente (Fig. 9.11 y 9.12).

Tabla 9.29. Test T: Estadísticos de muestras relacionadas

	Diferencias relacionadas					T	Gl	Sig. (bilateral)
Modelo B	Media	Desviación típ.	Error típ. De la media	95% Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 T.Funcionales - T.Léxicos	,531	5,857	,651	-,764	1,826	,816	80	,417

Figura 9.11. Modelo B: términos de función

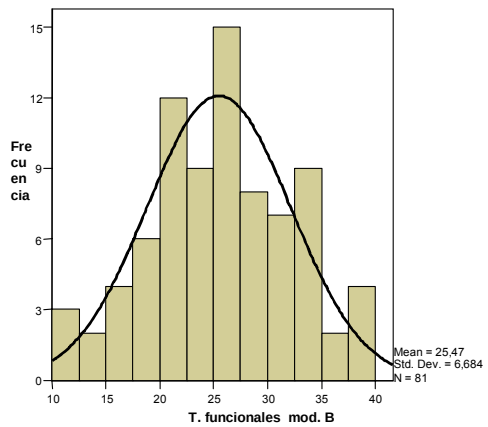
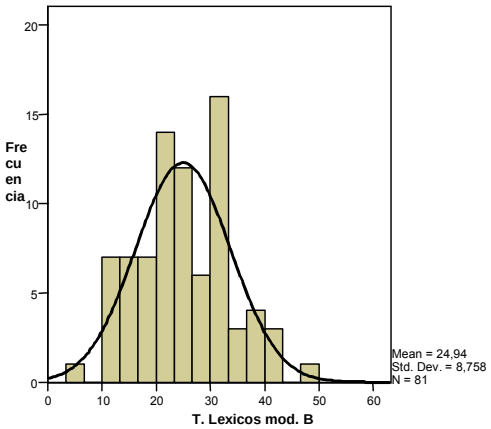


Figura 9.12. Modelo B: términos léxicos



A la luz de los estadísticos podemos concluir que los términos gramaticales se recuperan con mayor facilidad, pero la diferencia es mínima en ambos modelos de C-test.

La redundancia de los textos es otro aspecto que hay que tener en cuenta. Al ser relativamente sencillos, en ellos abundaba la repetición de palabras relevantes desde el punto de vista léxico. Pensamos que la redundancia facilitó su recuperación y acortó diferencias. A pesar de todo, los términos gramaticales se recuperaron mejor, en consonancia con las investigaciones de Klein-Braley (1985: 91) y Dörnyei y Katona (1992:197).

Así pues, en términos generales, se confirma la hipótesis 3:

> "En este tipo de prueba el alumno recuperarán mejor los términos funcionales que los de contenido léxico".

Ahora bien, hemos de matizar que, aunque en general, efectivamente, los términos funcionales se recuperan mejor, los de contenido léxico también resultan fáciles si aparecen previamente en el texto o son muy frecuentes en la lengua. Por otra parte, el hecho de que la recuperación de términos funcionales sea más sencilla no afecta de forma significativa en los C-tests aplicados.

9.6. Casuística en la recuperación de las omisiones: Análisis de los errores

En apartados anteriores hemos analizado las diferencias de recuperación entre unos ítems y otros, basadas principalmente en su grado de dificultad. Entre los factores que determinan el grado de dificultad de un ítem hemos señalado:

- su frecuencia en la lengua
- su adscripción a clases abiertas o cerradas (términos léxicos y funcionales)
- la redundancia del texto

Se ha realizado un análisis cuantitativo, estudiando los estadísticos correspondientes. Sin embargo, también el análisis cualitativo de la recuperación de algunos ítems puede resultar interesante y aportar claves para futuros estudios. No

pretendemos hacer aquí un análisis de los errores exhaustivo, que escaparía a los objetivos de nuestra investigación, tan sólo mostrar algunas pautas de error e inferir sus causas.

Para la corrección del C-test se adoptó el criterio de la "palabra exacta". Aunque pareció ser el más adecuado y objetivo, hay que señalar que tiene sus limitaciones, ya que hemos comprobado que algunos términos no se recuperan correctamente, y en consecuencia no computan, debido a errores meramente ortográficos, que se podrían considerar "menores". En un modelo comunicativo, este tipo de errores carece de importancia. Es obvio que, en estos casos, el sujeto reconoce el ítem omitido, pero comete fallos en la producción escrita. Se producen errores de transcripción, mucho más frecuentes aún en omisiones no guiadas:

Los ítems 25 del modelo A y 75 del B, correspondientes a la omisión del término *government*, muestran los siguientes estadísticos:

Tabla 9.30a. Ítem 25 modelo A: *government*

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	34	42,0	47,9	47,9
	Correcto	37	45,7	52,1	100,0
	Total	71	87,7	100,0	
Perdidos	Sin hacer	10	12,3		
Total		81	100,0		

Tabla 9.30b. Ítem 75 modelo B: *government*

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	59	72,8	76,6	76,6
	Correcto	18	22,2	23,4	100,0
	Total	77	95,1	100,0	
Perdidos	Sin hacer	4	4,9		
Total		81	100,0		

Cuando se aportan las pistas del número de letras omitidas (modelo A) 37 sujetos recuperan correctamente el término. Sin omisiones guiadas el número desciende hasta sólo los 18 sujetos.

Sin embargo, en ambos casos es muy elevado el número de sujetos que intenta la recuperación, curiosamente mayor en las omisiones no guiadas. Esto se explica por la frecuencia de uso del término en la lengua. A pesar de todo, muchos de los sujetos de la muestra (números 39, 127, 129, 138) que no lo recuperaron consignaron *governament[69] o *goverment en su lugar. Es decir, que muchos más sujetos de los que refleja la tabla de frecuencias habían identificado la palabra aunque no la recuperaron con éxito.

Caso semejante encontramos en la recuperación de los términos *vehicle* y *profession* que aparecen con frecuencia transcritos como *vehicule, *vehiculs y *profesions respectivamente. La palabra *awful* se recupera como *awfull, sobre todo cuando la omisión no es guiada.

Otras veces el término mutilado no se recupera correctamente pero se deduce que hay cierto grado de comprensión del texto, por ejemplo cuando se recupera *hungry por *hunger* o *patron por *pattern*. En el primer caso el sujeto ha comprendido el campo semántico y la familia a que pertenece el término buscado, aunque gramaticalmente realiza una incorrección grave. En el segundo, de nuevo el sujeto comprende el texto, pero no encuentra la palabra deseada y busca otra que se adapte al contexto (en este caso, contaminada por la L1).

En ocasiones se cometen errores en la concordancia, algunos pueden deberse simplemente a la falta de atención, como cuando se recupera *animal por *animals* en el ítem 91 del modelo B (sujeto número 6).

Otros errores evidencian el total desconocimiento del término por parte del sujeto, que actúa de forma aleatoria, esperando que el azar le ayude. Cuando el sujeto (al que se asignó el número 137 en el estudio) consigna *hundred por *hunger*, demuestra que no comprendió el texto ni buscó en el contexto inmediato de la omisión, simplemente intentó encontrar en el vocabulario que maneja otro término que comenzara del mismo modo y tuviera un número de letras similar.

Varios sujetos (números 2, 93) consignaron el auxiliar *had en lugar de *has* en el ítem 10 del modelo A. Esto parece indicar que se fiaron del procesamiento automático al ver el inicio de la palabra requerida, pero no comprobaron mediante la adyacencia léxica las claves sintácticas que aparecían en el texto inmediato. Faltó

[69] Los términos precedidos por el asterisco (*) son incorrectos en lengua inglesa, corresponden a la trascripción literal de cómo algunos sujetos de la muestra recuperan las omisiones del C-test.

atención y reflexión gramatical para identificar que el auxiliar formaba parte de la forma verbal “has been much reduced” en el contexto de “the last two decades”.

Por otra parte, a partir de la observación minuciosa de las pruebas (lectura detallada, búsqueda de pautas de comportamiento al realizar la prueba, errores comunes, etc.) descubrimos algunos datos cualitativos que investigaciones posteriores deberán valorar de forma cuantitativa. Así, a los factores citados al comenzar este epígrafe (frecuencia, clase abierta o cerrada y redundancia del texto), que determinan la facilidad/dificultad de recuperación de los ítems, hemos de añadir dos más, que ya hemos apuntado a lo largo de este capítulo:

- la longitud de la palabra
- su grado de abstracción

Hemos observado que la longitud del término mutilado afecta a su recuperación. Efectivamente, las palabras de mayor número de letras (y/o sílabas) coinciden con las que se recuperan peor a pesar de ser frecuentes (ver datos de la recuperación del término *government*), aunque reconocemos que esta variable se solapa en muchos casos con la de términos léxicos y funcionales, ya que los términos de función son generalmente cortos, frente a mayor variabilidad de tamaño de los términos léxicos.

También el grado de abstracción de los textos, y por tanto, de su léxico parece dificultar la tarea de recuperación de las omisiones, en sintonía con los resultados que apuntan Babaii y Moghaddam (2006).

Si nos fijamos en el texto *American Imperialism*, cuyo promedio desciende con respecto a los otros textos tanto en omisiones guiadas como no guiadas, veremos que contiene muchos términos “largos” (*leadership, benefit, worldwide, weapons, peaceful, resolution, conflicts, proliferation, development, education, engagement, cooperation, prosperity, generous, peacemaking,* etc.), algunos afectados por la mutilación, lo que pudo influir en la comprensión del texto y posterior recuperación correcta de los ítems. Son términos menos frecuentes en la lengua y muchos se pueden considerar “técnicos” o “académicos” (véase el capítulo 4, apartado 4.2.3.4).

Laufer (1997) identificó la longitud de la palabra, la parte del discurso y su grado de abstracción como factores intraléxicos cuyo efecto en el aprendizaje del vocabulario no está claro "with no clear effect". En la resolución de C-tests, sin embargo, es evidente su incidencia en el grado de dificultad de los ítems.

9.7. Análisis empírico de los resultados obtenidos en *Cavemen?*

Por las razones que esbozamos en el capítulo anterior, se decidió administrar a los alumnos una prueba tipo PAAU: *Cavemen?*. Junto a los resultados de las PAAU oficiales, *Cavemen?* será tomada como principal referencia para el estudio de la validez concurrente del C-test mediante el análisis de su correlación con las otras pruebas aplicadas.

Esta prueba aporta en sí misma gran cantidad de información a nuestra investigación. Disponemos de la puntuación global en la prueba. Además, contamos con la puntuación obtenida en cada una de las cinco preguntas de que consta. Por último, si las agrupamos atendiendo a su carácter objetivo/subjetivo, daremos un paso más, muy ilustrativo para nuestro trabajo. Y, al analizar sus correlaciones con el C-test, responderemos a la hipótesis de trabajo 2.

9.7.1. Descripción de *Cavemen?* Estructura e interrelaciones

La tabla que aparece a continuación muestra de forma esquemática cómo se estructura la prueba, el rango de puntuaciones y el tipo de preguntas que contiene:

Tabla 9.31. Estructura de la prueba de Inglés de las PAAU (Herrera 1999: 95)

Item	Score	Competence	Type of the item	Technique
1.	0-2	Communicative	Subjective	Open answer
2.	0-2	Comprehension	Objective	True/False
3.	0-1	Lexis	Objective	Matching
4.	0-2	Syntax	Objective	Cloze
5.	0-3	Communicative	Subjective	Non-directed essay

La Tabla 9.32 refleja los estadísticos descriptivos obtenidos (N, medias y desviación típica) para cada pregunta de la prueba *Cavemen?* en escala de 0 a 10. **T/F10** corresponde a la primera pregunta, de "verdadero o falso". La segunda pregunta aparece como **OPQ10** y corresponde a las preguntas abiertas sobre el texto. Ambas se valoran sobre 2 puntos en la prueba. La pregunta de vocabulario se identifica como **LEX10** y supone 1 punto del total de la nota. La gramática aparece como **SYNT10** y de nuevo tiene un valor de 2 puntos. La redacción final supone el 30% de la nota y aparece con la denominación de **COMP10**.

Tabla 9.32. Estadísticos descriptivos

	Media	Desv. Típ.
T/F10	8,1096	2,70595
OPQ10	4,5448	2,90078
LEX10	6,0988	2,36554
SYNT10	4,3657	2,48429
COMP10	4,8097	2,67357
Ctestt10	5,1130	1,46754

N = 162

Llama la atención el elevado promedio conseguido en la pregunta de "verdadero o falso" (8,1 puntos), indicador de su escasa dificultad y pobreza discriminatoria. También la pregunta de vocabulario obtiene un buen promedio (6,09), frente a los de la gramática (4,36) y la parte subjetiva de la prueba (4,54 y 4,8). El C-test queda en un punto medio y su desviación típica es la menor.

A continuación, observaremos los resultados obtenidos en la prueba agrupados según el tipo de pregunta. Llamamos **CavemenTotal** a la calificación global en *Cavemen?.* **CavemenObj** indica el resultado obtenido en la parte objetiva de la prueba (*true/false* + *lexis* + *syntax*) en escala de 0 a 10. **CavemenSubj** engloba las puntuaciones de la parte subjetiva (*open questions* + *composition*) en la misma escala. Puesto que es una prueba equilibrada, a cada parte le corresponden 5 puntos de los diez totales.

En la Tabla 9.33 figura el promedio global de la prueba (5,44 puntos), ligeramente superior al del C-test (5,1) pero bastante inferior al conseguido en el examen oficial de las PAAU (6,3). No obstante, la diferencia con las PAAU oficiales

se explica por el número y las características de los sujetos presentados (recordemos que la muestra se reduce a los 81 alumnos con el Bachillerato superado en junio).

En cuanto al tiempo, las PAAU oficiales se realizaron aproximadamente dos meses después. Se presupone la motivación y el estudio de los alumnos presentados, debido a la trascendencia de su actuación para su futuro académico.

Dentro de la prueba *Cavemen?,* la parte objetiva fue la mejor resuelta por los sujetos, ya que la media alcanza los 6,21 puntos. Mientras que en la parte subjetiva, sólo se obtiene un promedio de 4,69.

Es evidente que la parte objetiva resultó más asequible a los alumnos, tanto, que se impone reflexionar acerca de la validez de algunos ítems que condicionan claramente los promedios, como la pregunta de verdadero o falso, que consigue una media de 1,6 en una escala de 0 a 2 puntos (véase la Tabla 9.32). Evidentemente, su grado de dificultad es mínimo y cabe valorar incluso el azar.

En su estudio sobre el examen de Inglés de las PAAU, basándose en los datos de una muestra de 450 exámenes, Herrera (1999) cuestionó la validez de las preguntas de tipo objetivo de las PAAU, porque no discriminan entre el alumnado, cuando en realidad es esa la finalidad de la prueba. Observó el sesgo de las curvas correspondientes a los ítems objetivos y concluyó que el índice de facilidad/dificultad de los mismos es el causante de las altas puntuaciones. De este modo, se pierde el pretendido equilibrio de la prueba y la discriminación de los alumnos depende casi exclusivamente de la parte subjetiva del examen.

Tabla 9.33. Estadísticos de muestras relacionadas

	Media		Desv. Típ.
	Estadístico	Error típ.	Estadístico
CavemenTotal	5,4485	0,15800	2,01101
CavemenObj	6,2105	0,14829	1,88741
CavemenSubj	4,6920	0,20134	2,56266

N = 162

Los datos que hemos obtenido a partir de *Cavemen?* corroboran los resultados de Herrera (1999). Coincidimos en que algunas preguntas objetivas, especialmente la de “verdadero o falso”, no discriminan entre los alumnos. Este hecho nos llevará a

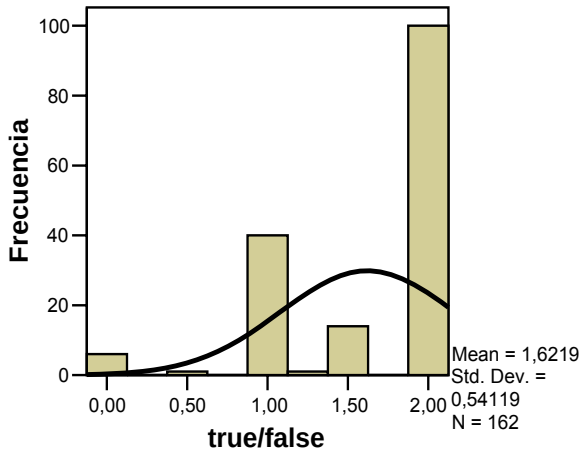
proponer la revisión de la prueba de Inglés de las PAAU, uniéndonos a otras voces (Herrera 2005; García Laborda 2005; Watts y García Carbonell 2005; Fernández y Sanz 2005).

Los histogramas que vemos a continuación muestran de forma clara que las preguntas de la prueba no están bien planteadas.

En el histograma de la pregunta de verdadero/falso (Fig. 9.13) vemos que la curva es mesocúrtica y la distribución de las puntuaciones muy irregular. El sesgo negativo indica que la mayoría de las puntuaciones son altas, resultó muy fácil. Se observa que las frecuencias se concentran en los números enteros: 1 y 2, sobre todo en el 2. Así debe ser si los correctores siguen fielmente las instrucciones de corrección y califican cada apartado del ejercicio con notas enteras: 0 ó 1.

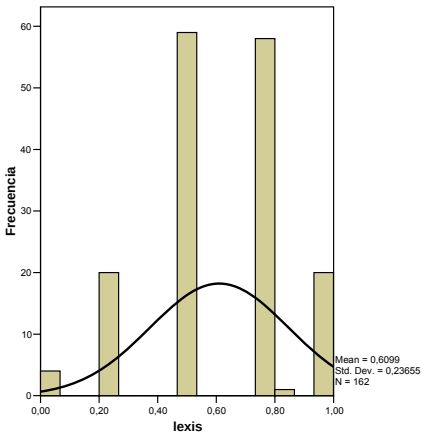
Esta pregunta supone buena parte de la puntuación total de la prueba (dos puntos de los diez totales), puede resultar determinante en los resultados globales y, sin embargo, como vemos, su potencial para discriminar a los sujetos queda gravemente cuestionado.

Figura 9.13. Histograma de la pregunta de “verdadero o falso”



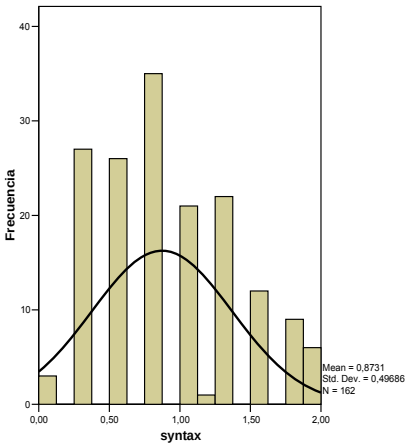
De nuevo en la pregunta de vocabulario, que hemos denominado “lexis”, se aprecia un leve sesgo negativo y una distribución irregular de frecuencias. El sesgo indica que esta pregunta también resultó fácil (media de 6,09 puntos en escala de 0 a 10). Las calificaciones obtenidas se concentran en las puntuaciones múltiplos de 0,25, puntuación asignada a cada apartado de la pregunta.

Figura 9.14. Histograma de la pregunta de vocabulario



No obstante, en la pregunta de gramática, la tercera que incluimos en la parte objetiva de las PAAU (Herrera 1999), la distribución tiende a la normalidad, como se aprecia en el histograma siguiente, en el que aparece con la denominación “syntax”.

Figura 9.15. Histograma de la pregunta de gramática



Veamos ahora los resultados de la parte subjetiva de la prueba, formada por las preguntas abiertas y la redacción.

El histograma de las preguntas abiertas presenta una media de 0,9 en una escala de 0 a 2 y una distribución mesocúrtica (Fig. 9.16). El correspondiente a la redacción presenta una curva bimodal, con sus picos en el 1 y en el 2. La media es de 1,44 en escala de 0 a 3 puntos (Fig. 9.17) y una distribución de puntuaciones muy similar a la del anterior.

Se aprecia claramente en él la dispersión de frecuencias y la tendencia del corrector a puntuaciones cerradas en los ensayos, es decir, a calificar con números enteros cuando no son valores extremos (Amengual 2003).

Figura 9.16. Preguntas abiertas: histograma

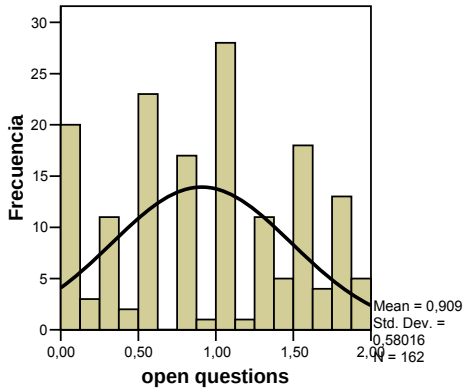
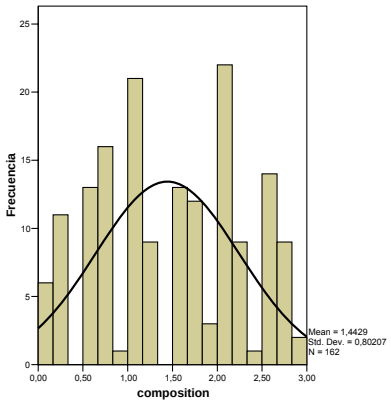


Figura 9.17. Redacción: histograma



Los histogramas que vemos a continuación corresponden a los resultados de la parte objetiva y subjetiva de la prueba *Cavemen?* respectivamente (Fig. 9.18 y 9.19).

Figura 9.18. *Cavemen?*: parte objetiva

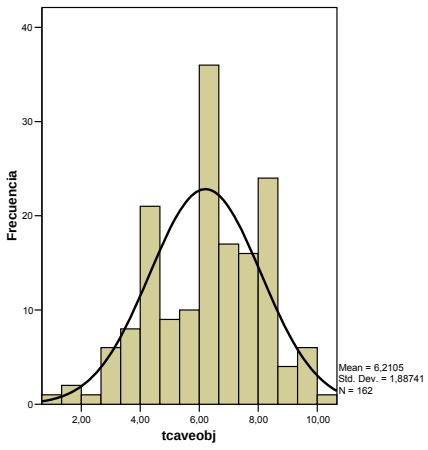
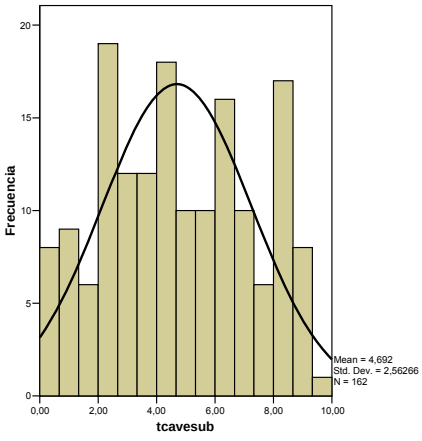


Figura 9.19 *Cavemen?*: parte subjetiva



En la parte objetiva se obtuvo un promedio de 6,2 puntos en una escala de 0 a 10 (aunque, como hemos visto, supone la mitad del total del examen), lo cual indica que esta parte de la prueba resultó demasiado fácil y con poca potencia discriminatoria.

Si lo comparamos con el histograma de la parte subjetiva de *Cavemen?* (preguntas abiertas y redacción), vemos que la media baja considerablemente, hasta 4,69 puntos sobre 10. La parte subjetiva resultó más difícil, pero a la vez es la de mayor poder discriminatorio. La distribución de frecuencias es normal en ambos casos.

9.7.2. Correlaciones entre *Cavemen?* y las otras pruebas: PAAU 2001, 2ª Evaluación y C-test

La Tabla 9.34 refleja la correlación de Pearson entre las distintas pruebas analizadas.

En primer lugar, nos fijamos en la correlación entre la prueba tipo PAAU realizada en clase, *Cavemen?*, y la calificación obtenida por los sujetos presentados en la prueba de Inglés de las PAAU de junio de 2001. Son semejantes en su formato y estructura, pero se diferencian, sobre todo, en las circunstancias de su aplicación y la trascendencia para los sujetos.

Tabla 9.34. Correlaciones de Pearson entre las pruebas aplicadas en el estudio

	1	2	3	4
1. *Cavemen?*	--			
2. 2ª Evaluación	,805(**)	--		
3. Selectividad 2001	**,654(**)**	,575(**)	--	
4. Ctestt10	,750(**)	,723(**)	,722(**)	--

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

Se aprecia que la correlación de Pearson entre ambas pruebas es significativa: 0,654. Al compararlas estamos contrastando dos pruebas de rasgos comunes, pero con diferencias que hemos de valorar:

- **Variable textual**: a pesar de compartir estructura (tipología) y grado de dificultad, cada prueba es diferente ya desde el texto elegido en torno al cual gira todo el examen (tema, densidad, registro, vocabulario, etc.)

- **Variable temporal**: la prueba *Cavemen?* se aplicó tres meses antes, durante los cuales se llevó a cabo una labor de aprendizaje, preparación y práctica en este tipo de prueba (*teaching to the test*).

- **Lugar y condiciones de aplicación**: mientras que una de ellas se aplicó en el aula, con el profesor de la asignatura, las PAAU de junio se realizaron en las instalaciones de la Universidad correspondiente[70].

- **Corrección**: en la prueba *Cavemen*? la variable "corrector" queda controlada, puesto que fue corregida por las profesoras y la investigadora. En las PAAU, sin embargo, puede aparecer algún tipo de sesgo (dependiendo del tribunal).

- **Grado de significación**: la primera era parte de la práctica habitual en 2º de Bachillerato, mientras que en la segunda entran en juego el nerviosismo y la ansiedad propias de las pruebas externas "a gran escala" en las que la actuación del sujeto tiene repercusiones concretas en su futuro (en este caso, supone el acceso a la Universidad y a una carrera determinada).

Todas ellas determinan que la correlación entre las dos pruebas, a pesar de ser significativa, no sea perfecta. La Tabla 9.34 muestra que *Cavemen?* correlaciona incluso mejor con las calificaciones de la 2ª Evaluación (0,805) que la PAAU (0,575) y el C-test (0,722). La correlación PAAU-C-test es 0,750.

Revela, además, que el C-test correlaciona de forma significativa en todos los casos. Queda patente, por tanto, la fiabilidad de la prueba y su validez de constructo.

Resulta de especial interés para nuestro estudio constatar su alta correlación (0,722) con la PAAU de junio de 2001. Indica la validez predictiva del C-test (véase el apartado 9.7), basándonos en estos datos podemos recomendar el C-test como instrumento adecuado para la preparación de las actuales pruebas de Selectividad.

[70] Entre las propuestas para la Selectividad en el futuro, García Laborda (2005:36) incluye que las pruebas de Acceso a la Universidad se realicen en los propios Institutos de Enseñanza Secundaria para reducir la ansiedad de los alumnos. Naturalmente, este cambio implicaría posiblemente también otros, como la realización de la prueba a través de Internet.

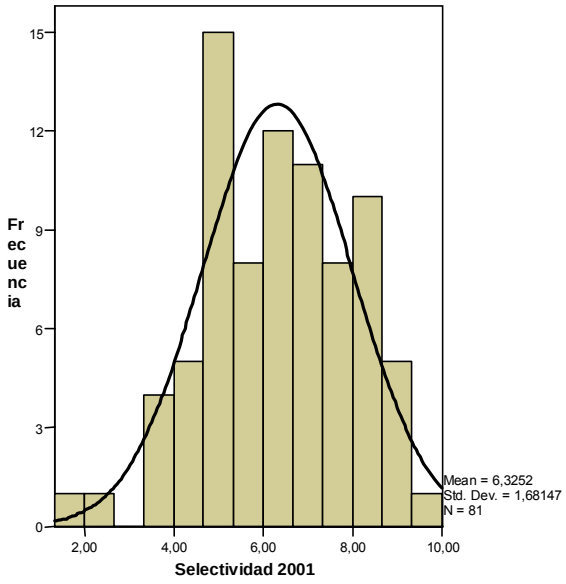
Por otro lado, como veremos, consideramos que podría formar parte de la prueba de Selectividad sustituyendo a preguntas de la actual que no discriminan[71].

El histograma de la PAAU de Inglés de Junio de 2001 para los sujetos de la muestra presentados (Fig. 9.20) presenta una distribución normal, con un ligero sesgo negativo que refleja los buenos resultados obtenidos en la prueba.

Se explica por los rasgos de la muestra; alumnos motivados y con el 2º curso de Bachillerato superado en todas las áreas. La mayoría de los sujetos consiguen más de cuatro puntos en la prueba (en escala de 0 a 10).

La actuación de los sujetos se ve incentivada en el caso de la Selectividad por la relevancia ya comentada de la prueba externa para el futuro académico y personal del alumno.

Figura 9.20. Histograma de la prueba de Selectividad (PAAU Junio 2001)



[71] Esta propuesta aparece ya en el artículo "Niveles de correlación entre el C-test y la prueba de Inglés de Selectividad" del libro *Estudios y criterios para un Selectividad de calidad en el examen de Inglés* (2005).

9.8. Análisis de la validez concurrente del C-test: correlaciones

Aunque en el apartado anterior anticipamos algunos datos, continuamos con el estudio de las correlaciones entre las distintas pruebas que forman este estudio:

- C-test
- Calificaciones en Inglés en la 2ª evaluación del curso académico 2000/01
- *Cavemen?*
- PAAU de junio de 2001

En este apartado respondemos a las preguntas de investigación que se corresponden con la primera hipótesis de trabajo. Son fundamentales, puesto que determinan la capacidad del C-test para medir lo que realmente pretende: la competencia global del alumno en lengua inglesa:

1. ¿Existe correlación significativa entre las puntuaciones obtenidas por un sujeto en un C-test y en la prueba de Inglés de las PAAU? ¿Y con respecto a la valoración que hace el profesor acerca de su progreso en la asignatura?
2. ¿Hay diferencias entre la correlación del C-test con las puntuaciones obtenidas en preguntas objetivas y subjetivas? Si las hay, ¿a qué se deben y cómo se explican?
3. ¿Discrimina el C-test de forma adecuada entre los sujetos, atendiendo a su competencia lingüística?

Para comenzar, haremos un análisis comparativo de los promedios obtenidos en cada una de ellas (Tabla 9.35). Después, en la Tabla 9.36 presentaremos las correlaciones de Pearson entre las pruebas aplicadas y en la 9.37 la prueba de muestras relacionadas. En las tablas, la denominación **CtestTotal10** corresponde a la nota obtenida en el C-test en una escala de 0 a 10 puntos. La prueba *Cavemen*? aparece también desglosada en parte objetiva: **CavemenObj,** y subjetiva: **CavemenSubj**.

En la Tabla 9.35 se observa que el rango de variación de los promedios de las pruebas no es muy amplio, lo que es signo de consistencia y fiabilidad en la evaluación. No obstante, destaca la puntuación de la Selectividad 2001 en un extremo (6,32) y la de la parte subjetiva de Cavemen (4,69) en el otro.

Tabla 9. 35. Estadísticos descriptivos: promedios

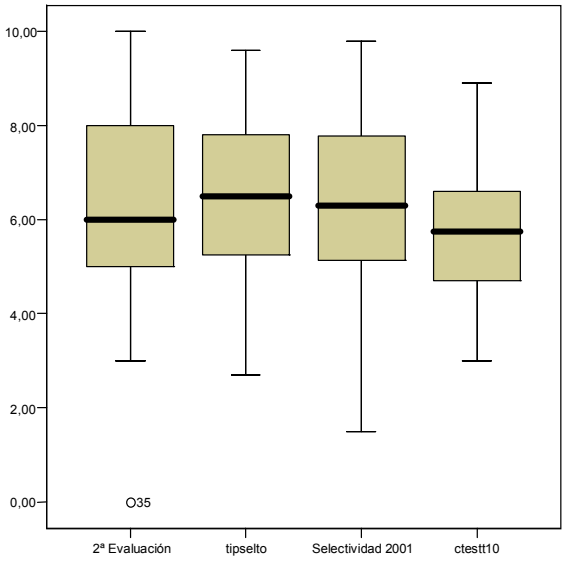
	N	Media		Desv. típica
	Estadístico	Estadístico	Error típico	Estadístico
2ª Evaluación	161	5,50	,163	2,068
Selectividad 2001	81	6,3252	,186683	1,68147
CtestTotal10	162	5,1142	,11550	1,47011
CavemenTotal	162	5,4485	,15800	2,01101
CavemenObj	162	6,2105	,14829	1,88741
CavemenSubj	162	4,6920	,20134	2,56266
N válido (según lista)	80			

El promedio más bajo (5,1) se lee en el C-test, si exceptuamos la parte subjetiva de *Cavemen?* cuya media es de sólo 4,69 puntos. Los promedios más altos en las pruebas que se han tomado como referencia: *Cavemen?* (5,44), las calificaciones de Inglés en la 2ª Evaluación (5,5) y el examen oficial de Inglés de las PAAU de Junio de 2001 (6,32). Podría justificarse, bien por la novedad del diseño del C-test, bien por la motivación que tienen cuando realizan un examen en función de una puntuación.

En el caso de las PAAU de junio de 2001 existe otro componente más: el reducido número de alumnos que las realizan y su nivel de competencia, que se supone superior, ya que han superado el curso en todas las áreas, incluida la de Inglés.

El diagrama de cajas, Figura 9.22, nos da la visión de conjunto. En el C-test se aprecia la menor dispersión de puntuaciones, tanto por encima como por debajo de la mediana, y en la 2ª Evaluación la mayor dispersión por encima de la mediana.

Figura 9.22. Diagrama de cajas de los promedios en las distintas pruebas



Así pues, al valorar los resultados obtenidos en el C-test, no podemos olvidar la novedad de la técnica. Para nuestros alumnos era una prueba totalmente desconocida, lo cual no implica necesariamente su rechazo. Bachman (1990: 114) alerta de que la experiencia previa con una prueba concreta afecta a la actuación del sujeto, pero también al profesor y, en consecuencia, ha ser tenido en cuenta.

> Utilizing familiar testing techniques does, to some extent, simplify the test developer's task, since this effectively narrows the range of techniques that must be considered. A potentially adverse effect of this, however, is the tendency for certain testing techniques to become institutionalized to the extend that they are perceived by test takers and test users alike as the only appropriate methods for testing language abilities. (Bachman 1990: 47-48)

Quizá el razonamiento de Bachman explique la resistencia del profesorado y las instituciones a innovar en materia de evaluación.

Una de las preguntas de investigación (3), en relación directa con la hipótesis 2, se refiere al tipo de prueba (objetiva o subjetiva) con que correlaciona mejor el C-test:

> "Si subdividimos la prueba de Inglés de las PAAU en las distintas preguntas que la forman, ¿hay diferencias entre la correlación del C-test con las puntuaciones obtenidas en preguntas de tipo objetivo y subjetivo? Si las hay, ¿a qué se deben y cómo se explican?"

Para contestarla manejamos los datos obtenidos al desglosar el examen *Cavemen?* en parte objetiva y subjetiva, cuyos resultados muestran un notable nivel de semejanza con los de la segunda prueba piloto estudiada en el capítulo 7.

Comprobamos que el subtest objetivo es el que mejor resuelven los alumnos (media 6,2), sin embargo, el que presenta mayor correlación con el C-test es el subjetivo, si bien en todos los casos las correlaciones son significativas según se aprecia en la Tabla 9.36, que amplía la 9.34, al incluir el desglose de *Cavemen?* en parte objetiva y subjetiva:

Tabla 9.36. Correlaciones de muestras relacionadas

		1	2	3	4	5	6
1. Ctestt10		**--**					
	N	**162**					
2. 2ª Evaluación		**,723(**)**	--				
	N	**161**	161				
3. PAAU 2001		**,722(**)**	,575(**)	--			
	N	**81**	80	81			
4. *Cavemen?*		**,750(**)**	,805(**)	,654(**)	--		
	N	**162**	161	81	162		
5. Caveobj		**,659(**)**	,672(**)	,571(**)	,867(**)	--	
	N	**162**	161	81	162	162	
6. Cavesub		**,692(**)**	,772(**)	,558(**)	,929(**)	,622(**)	--
	N	**162**	161	81	162	162	162

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

Para precisar el carácter de la relación entre el C-test y cada una de las partes de la PAAU *Cavemen?*, en la siguiente Tabla (9.37) tomaremos las correlaciones desglosadas, lo que nos ayudará a rechazar o confirmar la hipótesis 2:

> "De ello se sigue que, por sus características, al ser una prueba objetiva de elementos discretos, para un mismo sujeto, el C-test correlacionará mejor con las pruebas de tipo objetivo que con las de tipo subjetivo y holístico". (Hipótesis 2)

Tabla 9. 37. Correlaciones C-test-*Cavemen?*

		ctestt10	T/F10	OPQ10	LEX10	SYNT10	COMP10
CTESTT10	Correlación de Pearson	--	,385(**)	,616(**)	,309(**)	,672(**)	,665(**)

N = 162
** La correlación es significativa al nivel 0,01 (bilateral).

Se aprecia que los valores más altos corresponden a la gramática (0,672), en consonancia con otros estudios (Cohen *et al*. 1984; Connelly 1997; Eckes y Grotjahn 2006). Sin embargo, curiosamente, la pregunta de vocabulario de la PAAU no correlaciona con el C-test, a pesar de que se reconoce a éste como prueba que mide el vocabulario (este aspecto contrasta con el estudio de Eckes y Grotjahn 2006). Posiblemente se deba al formato de la pregunta de lexis, recordemos que se trata simplemente de localizar cuatro sinónimos en el texto. Por ello, es una muestra insuficiente para formular conclusiones al respecto. Así pues, la parte objetiva de *Cavemen?* aglutina ítems muy distintos y resultados dispares.

En cuanto a la parte subjetiva, tanto la redacción como las preguntas abiertas obtienen una correlación aceptable y uniforme (0,665 y 0,616). A la vista de los datos empíricos hemos de rechazar la hipótesis 2, puesto que, a pesar de ser una prueba objetiva de elementos discretos, el C-test correlaciona mejor con pruebas de tipo holístico, como las preguntas abiertas y la redacción.

Estos resultados coinciden con los de Lee (1996), que estudió la validez concurrente de las pruebas de cierre de ratio fija y los ensayos. Considera que su alta correlación prueba “the common integrative nature between the essay and cloze tests” (op. cit.: 69). Lee (1996) se decanta por las pruebas de cierre como alternativa a otras pruebas: “This result confirms the finding of two previous studies (Fotos 1991, Hanania and Shikhani 1986) that cloze tests can be an *alternative* to essay tests” (la cursiva es mía). Además, sugiere su uso en la práctica docente: “as a teaching device in classroom situations” (op. cit.: 62). Al final de esta tesis haremos una propuesta similar para el C-test, apoyándonos en los datos estadísticos de nuestra investigación.

Resulta llamativo que una prueba objetiva de elementos discretos, como el C-test, correlacione mejor con pruebas subjetivas (exceptuando su consabida conexión

con los aspectos gramaticales). Parece indicar que el C-test es un tipo de examen más próximo a las pruebas subjetivas que a las objetivas; a pesar de que es objetivo sobre todo en la corrección, en su realización requiere producción y presenta rasgos creativos, puesto que va más allá del mero reconocimiento.

Resumiendo, destacamos la correlación del C-test con los resultados globales de la prueba modelo de las PAAU realizada en clase, *Cavemen?* (0,750), aunque si tenemos en cuenta las dos partes en que se subdivide, la correlación baja sobre todo en la parte objetiva (0,659), aún siendo siempre significativa.

La correlación entre las PAAU oficiales y el C-test es de 0,722, ligeramente inferior a la observada entre *Cavemen?* y el C-test (0,750). La tabla que muestra los promedios refleja que en las PAAU se alcanza la media más alta: 6,32. Se justifica por las características de las pruebas a gran escala (*high-stake tests*), a las que sólo se accede previa superación del Bachillerato, y en las que la motivación, competencia y rendimiento de los alumnos son superiores a la media, a pesar de la incidencia de otros factores (nervios, ansiedad, etc.).

No obstante, es una correlación significativa en ambos casos, lo que indica que el C-test, dado su carácter económico en diseño y administración bien podría tenerse en cuenta como componente significativo en la configuración de la prueba de Inglés de las PAAU o de cualquier otra prueba selectiva que pretenda discriminar entre el alumnado. Confirmamos así parcialmente la primera hipótesis de trabajo:

> "Partiendo de las características de la prueba podemos predecir que el C-test correlacionará bien con otras pruebas estandarizadas que midan la competencia global en lengua extranjera, como las PAAU, y también con las calificaciones obtenidas por los alumnos en la asignatura de Inglés".

Efectivamente, el C-test correlaciona de forma significativa con otras pruebas estandarizadas que miden el mismo constructo: la competencia global en Inglés como Lengua Extranjera. En nuestro caso concreto, con las PAAU (tanto si se aplica este formato en el aula como si se hace de forma oficial).

En cuanto a la apreciación de los respectivos profesores acerca de la competencia de los alumnos en lengua inglesa, expresada en la calificación de la asignatura en la 2ª Evaluación del curso, la correlación con el C-test es de 0,723.

Es importante corroborar que la valoración de los profesores está en sintonía con los resultados del C-test. Refleja un aspecto básico para la práctica docente y confirma ya en su totalidad la hipótesis que acabamos de citar; por sus características el C-test puede considerarse un instrumento de evaluación válido en el contexto del aula.

La Tabla 9.38 presenta la prueba de muestras relacionadas. En ella, el T-test muestra que al comparar al C-test con la 2ª Evaluación se obtiene t=3,484 con p<0,001, lo que indica que existe diferencia significativa, es decir que el C-test resultó más difícil, circunstancia común a las otras pruebas (véase Fig. 9.22). Pero, a pesar de todo, las correlaciones reflejadas en la Tabla 9.36 indican la armonía entre las pruebas.

Tabla 9.38. Prueba de muestras relacionadas: T-test.

		Diferencias relacionadas					t	Gl	Sig. (bilateral)
		Media	Desv. típ.	Error típ. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	2ª Evaluación – ctestt10	,39255	1,42953	,11266	,17005	,61505	3,484	160	,001
Par 2	Selectividad 2001 – ctestt10	,57210	1,17885	,13098	,31143	,83276	4,368	80	,000
Par 3	Ctestt10 – tipselto	-,33426	1,33127	,10459	-,54081	-,12771	-3,196	161	,002
Par 4	Ctestt10 – tcaveobj	-1,09630	1,43790	,11297	-1,31939	-,87320	-9,704	161	,000
Par 5	Ctestt10 – tcavesub	,42222	1,87461	,14728	,13137	,71308	2,867	161	,005

Por otra parte, en la información que aportan los distintos modelos aplicados, los datos indican que hay pocas variaciones en el comportamiento de los modelos A y B del C-test, con correlaciones significativas en ambos modelos y promedios también similares: 5,18 en el A y 5,04 en el B.

9.9. Validez predictiva

Puesto que no es fácil realizar un seguimiento a largo plazo que pruebe la validez predictiva del C-test, veremos únicamente si sus resultados predicen bien los obtenidos en las PAAU, realizadas con posterioridad.

Si tenemos en cuenta a los sujetos que se presentaron a la Selectividad, la muestra queda reducida a la mitad. Veremos los promedios de este grupo en las PAAU oficiales y en el C-test. Los alumnos que posteriormente se presentaron a las pruebas de Selectividad alcanzaron una media de 5,75 puntos en el C-test realizado dos meses antes. En las PAAU, la media se incrementó sensiblemente hasta llegar a 6,32 puntos. La mejora en los resultados se explica por la mayor competencia de los sujetos, el tiempo de estudio y práctica (*teaching to the test*) transcurrido entre ambas pruebas, y por factores derivados de la trascendencia que tiene la actuación en las PAAU para el futuro de los alumnos. Mientras que la media del total de la muestra (162 sujetos) en el C-test fue de 5,1 puntos, el promedio de los 81 sujetos que realizaron las PAAU asciende a 5,75.

La correlación de Pearson entre los resultados del C-test y las PAAU de junio de 2001 es de 0,722, y expresa la validez predictiva del C-test (Véase Tabla 9.33). Este valor es incluso más alto que los correspondientes a las correlaciones entre las PAAU y las calificaciones de la 2ª Evaluación (0,575), y entre las PAAU y la prueba modelo de Selectividad, *Cavemen?* (0,654). Estos datos tan satisfactorios indican que, partiendo de la actuación de los sujetos en el C-test (abril 2001) podíamos predecir de forma empírica su actuación en un futuro cercano en las PAAU (junio 2001), a pesar de que otros muchos factores entran en juego (ansiedad, etc.). Y que podíamos hacerlo de manera más fiable que con otras pruebas.

En concreto, llama la atención que el C-test obtenga mayor correlación con las PAAU oficiales que la prueba modelo de PAAU aplicada en clase, *Cavemen?*. Las implicaciones pedagógicas que se infieren de este dato son claves. El C-test se manifiesta como instrumento de evaluación de la lengua cuya validez predictiva queda demostrada. Entre otras virtudes (objetividad, facilidad de creación y corrección), este diseño ofrece al profesorado la posibilidad de predecir actuaciones futuras del alumno. Su uso en la preparación de la prueba de Inglés de la Selectividad resulta muy recomendable.

9.10. Fiabilidad

Entendemos la fiabilidad como consistencia entre distintas actuaciones del mismo sujeto (véase el capítulo 3, apartado 3.3.1). La fiabilidad, como la validez de una prueba, es un indicador de su calidad. Siguiendo a Messick (1996), al asegurar que el C-test es una prueba fiable expresamos que refleja una consistencia en la actuación del sujeto en distintas tareas y ocasiones, y con correctores diferentes.

La literatura ha señalado los posibles métodos para cuantificar la fiabilidad de una prueba (métodos *test-retest* y *split-half*, aplicar versiones paralelas de la prueba, cálculo del error estándar, Alfa de Cronbach, etc.) y los obstáculos que se encuentran en la práctica.

Si hubiéramos administrado a nuestros alumnos un segundo C-test, el mismo o una versión paralela, se habría producido un aprendizaje (familiarización con la técnica) y posible desmotivación, que falsearían los resultados. Así pues, el método *test-retest* fue inmediatamente rechazado. Se decidió afrontar el análisis cuantitativo de la fiabilidad del C-test mediante el método de "análisis por mitades" y calculando el Alfa de Cronbach, además del análisis de las correlaciones con otras pruebas.

9.10.1. Análisis por mitades

Planteamos aplicar en nuestro estudio de la fiabilidad del C-test el método *split half.* El número de ítems de la prueba permite subdividir fácilmente los resultados obtenidos en dos mitades equivalentes (de 50 ítems cada una) y así asignar dos puntuaciones a cada alumno.

Debido al cambio de formato introducido en los ítems 51 al 100 del C-test, se decidió no hacer la división del modo más sencillo (ítems 1-50 y 51-100), sino tomando para cada una de las mitades 25 ítems guiados y 25 no guiados. El C-test de cada alumno quedó subdividido en:

1. Resultado obtenido en los subtests 1 y 3 (ítems 1-25 y 51-75).
2. Resultado obtenido en los subtests 2 y 4 (ítems 26-50 y 76-100).

Los estadísticos descriptivos (Tabla 9.39) muestran que la media obtenida en la primera mitad es menor que la de la segunda, probablemente debido al aprendizaje. No obstante, se aprecian valores muy similares en las puntuaciones máximas y mínimas de ambas mitades, el error típico y la desviación, signo evidente de fiabilidad.

Tabla 9.39. Estadísticos descriptivos de ambas mitades

	Mínimo	Máximo	Media		Desv. típica
	Estadístico	Estadístico	Estadístico	Error típico	Estadístico
Split half 1	5,00	45,00	23,4136	0,61638	7,84528
Split half 2	5,00	46,00	27,7099	0,59435	7,56478

N = 162

Tabla 9.40. Análisis del C-test por mitades: Correlaciones

SplitHalf 1= Ctest1+Ctest3
SplitHalf 2= Ctest2+Ctest4

Correlaciones de Pearson

	1	2
1. SplitHalf1	--	
2. SplitHalf2	,816(**)	--

N = 162
** La correlación es significativa al nivel 0,01 (bilateral).

La correlación entre los resultados obtenidos en ambas mitades es otro indicador del grado de fiabilidad de la prueba. En este caso se logra una correlación muy alta y significativa: 0,816, como queda reflejado en la Tabla 9.40.

9.10.2. Alfa de Cronbach

La potencia de los tests estadísticos de que disponemos hace posible también el análisis de la fiabilidad del C-test mediante el cálculo del Alfa de Cronbach. En el documento electrónico de Klein-Braley y Raatz (1998) “Introduction to language

testing and C-tests" se recomienda el uso de esta fórmula en lugar de la de Kuder-Richardson para medir la consistencia del C-test.

El C-test consigue un 0,794 (teniendo en cuenta los cuatro subtests que forman la prueba), como revelan la Tablas 9.41a y 9.41b. Podemos diferenciar además entre los dos modelos aplicados, A y B. Los datos estadísticos de fiabilidad, tanto del modelo A como del B muestran un Alfa de Cronbach muy elevada: 0,890 y 0,892 respectivamente (Tabla 9.42).

Tabla 9.41a. Alfa de Cronbach del C-test

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
,794	,809	4

Tabla 9.41b Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación múltiple al cuadrado	Alfa de Cronbach si se elimina el elemento
CTEST1	37,86	133,087	0,456	0,807
CTEST2	35,49	135,829	0,541	0,698
CTEST3	40,97	130,266	0,481	0,799
CTEST4	39,05	119,265	0,590	0,662

Tabla 9.42. Alfa de Cronbach del C-test: modelos A y B

Modelo C-test	Alfa de Cronbach	N de elementos
Modelo A	0,890	4
Modelo B	0,892	4

Klein Braley y Raatz (1984: 140) señalaron que incluso en los C-tests que resultaron demasiado fáciles o difíciles para los sujetos se encontraban coeficientes de validez y fiabilidad aceptables. El estudio de Dörnyei y Katona (1992: 193) respalda este hecho. Encontraron coeficientes de fiabilidad de 0,75 y 0,77 en

estudiantes universitarios y de secundaria, respectivamente. Las tablas anteriores reflejan que los resultados de nuestra investigación están en la misma línea.

9.10.3. Validez y fiabilidad

En el apartado 3.4 del capítulo 3 apuntamos que también podríamos basar el análisis de la fiabilidad en el de la validez, ya que estos dos rasgos de las pruebas están tan íntimamente relacionados que a veces se consideran conceptos superpuestos (Hughes 1989; Weir 1988, 1993; Bachman y Palmer 1990*)*.

Las correlaciones reflejan la consistencia entre las actuaciones de los alumnos en pruebas que miden el mismo constructo, como ocurre con el C-test y las PAAU. Se ha demostrado la validez concurrente del C-test a partir del análisis de sus correlaciones con las otras pruebas aplicadas en este estudio empírico, y por tanto, implícitamente, su validez de constructo y fiabilidad (apartado 9.8).

Así pues, distintas vías nos llevan a concluir que el C-test es una prueba válida y fiable como instrumento de evaluación de la competencia general en lengua inglesa, con lo que la hipótesis 1 queda confirmada:

> "Partiendo de las características de la prueba podemos predecir que el C-test correlacionará bien con otras pruebas estandarizadas que midan la competencia global en lengua extranjera, como las PAAU, y también con las calificaciones obtenidas por los alumnos en la asignatura de Inglés".

Consideramos de especial importancia para nuestra investigación la confirmación de esta hipótesis, que engloba y da sentido a las restantes y al trabajo experimental aquí desarrollado. Estos resultados también contribuyen a demostrar lo que planteábamos en la Introducción como objetivo fundamental de la tesis:

> "El C-test es una prueba válida y fiable para medir la competencia global de los alumnos españoles de Enseñanzas Medias en Inglés como Lengua Extranjera".

9.10.4. Fiabilidad del corrector

No podemos olvidar en nuestro análisis la fiabilidad del corrector. Para que una prueba sea fiable ha de serlo también su corrección (Hughes 1989, 1994), y las pruebas de tipo objetivo facilitan la tarea.

Amengual (2003) estudió las distintas variables (intra e inter corrector) y sesgos que intervienen en la corrección de pruebas subjetivas, como son los ensayos de las PAAU. En el caso del C-test, por las propias características de la prueba, cuyo diseño deja poco margen a la subjetividad, un alto grado de fiabilidad está asegurado. A pesar de todo, incluso en las pruebas más objetivas hay que reconocer cierta subjetividad, aunque sólo sea en el diseño y creación de la prueba. En la fase de creación del C-test, las únicas decisiones del profesor se reducen a la selección de los textos. Una vez elegidos, las normas de Klein-Braley y Raatz (1981, 1984, 1997) son estrictas. El criterio de corrección también garantiza la objetividad, puesto que en nuestro estudio empírico sólo la palabra exacta se considera válida. A pesar de que impone otras limitaciones (véase 9.3.8) fue el elegido para nuestra investigación.

Después de este análisis estamos en condiciones de afirmar la validez y fiabilidad del C-test, objetivo fundamental de la tesis. A lo largo del capítulo han quedado confirmadas las hipótesis de trabajo 1, 3 y 4, que citamos de nuevo a continuación, aunque en algunos casos ha sido necesario matizar ciertos aspectos.

Hipótesis 1

"Partiendo de las características de la prueba podemos predecir que el C-test deberá correlacionar bien con otras pruebas estandarizadas que midan la competencia global en lengua extranjera, como las PAAU, y también con las calificaciones obtenidas por los alumnos en la asignatura de Inglés".

Hipótesis 3

"En este tipo de prueba el alumno recuperará mejor los términos funcionales que los de contenido léxico".

Hipótesis 4

Los cambios en el formato influyen directamente en los resultados obtenidos; si se incluye el número de letras que corresponde a cada omisión se facilita la tarea del alumno.

Por el contrario, la hipótesis 2 ha sido rechazada:

Hipótesis 2

“De ello se sigue que, por sus características, al ser una prueba objetiva de elementos discretos, para un mismo sujeto, el C-test correlacionará mejor con las pruebas de tipo objetivo que con las de tipo subjetivo y holístico”.

CAPÍTULO 10. ANÁLISIS DE REGRESIÓN LINEAL

10.1. Introducción

El procedimiento de regresión lineal es una técnica estadística que se utiliza para el análisis de la relación entre variables cuantitativas. En esta investigación lo utilizaremos para explorar y cuantificar las relaciones entre las distintas partes o subtests que forman el C-test (C-test 1, C-test 2, C-test 3 y C-test 4) y las otras pruebas aplicadas: *Cavemen?*, la Selectividad de junio de 2001 y la calificación en la 2ª Evaluación, que consideraremos como variables dependientes (VDs).

10.2. Análisis de regresión lineal de la 2ª Evaluación

Mediante este procedimiento comprobaremos el carácter de la relación existente entre los cuatro subtests del C-test[72] y las calificaciones obtenidas por los sujetos de la muestra en Inglés en la 2ª Evaluación del curso. La 2ª Evaluación es la variable dependiente o *criterio* (VD) y los subtests, las variables independientes o *predictoras* (VIs). Se trata de regresión múltiple porque interviene más de una variable independiente.

Determinaremos cómo cada subtest contribuye a explicar una parte de la varianza.

En los diagramas de dispersión, como veremos, la relación quedará expresada por el grado en que la nube de puntos se ajuste a una línea recta.

[72] En las tablas y gráficos los subtests se identifican como CTEST1, CTEST2, y así sucesivamente. La variable dependiente aparece como 2ª Evaluación.

En el apartado 9.8 del capítulo 9 comentamos la alta correlación entre las dos variables: C-test y calificaciones en Inglés en la 2ª Evaluación: 0,723. Mencionamos la importancia de este dato porque supone un alto grado de acuerdo entre la valoración de los profesores y los resultados conseguidos en el C-test. Si tenemos en cuenta los dos modelos de C-test se mantienen las cifras, especialmente en el modelo B (0,678 con el C-test modelo A y 0,788 con el B).

Los promedios obtenidos, en escala del 1 al 10, fueron 5,5 puntos en la 2ª Evaluación y 5,1 en el C-test, aunque en el C-test se aprecia la menor dispersión de puntuaciones y en la 2ª Evaluación la mayor (véase la Tabla 9.34 del capítulo 9).

La Tabla 10.1 resume los modelos aplicados para explicar la varianza de la variable dependiente (VD) 2ª Evaluación. Esta tabla recoge el coeficiente de correlación múltiple para cada paso, que va de 0,646 en el modelo 1, a 0,726 en el modelo 3. Cabe observar cómo, de manera automática, el sistema ha partido del C-test 4 como el mejor predictor de la VD, la 2ª Evaluación. el C-test 3 sólo entra en el último modelo y el C-test 1 ni siquiera aparece para explicar el 0,527 de la varianza, valor de R cuadrado.

Tabla 10.1. Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,646(a)	,417	,414	1,583	,417	113,912	1	159	,000
2	,707(b)	,499	,493	1,472	,082	25,908	1	158	,000
3	,726(c)	,527	,518	1,436	,027	9,113	1	157	**,003**

a Variables predictoras: (Constante), CTEST4
b Variables predictoras: (Constante), CTEST4, CTEST2
c Variables predictoras: (Constante), CTEST4, CTEST2, CTEST3

El valor R^2, o coeficiente de determinación, expresa la proporción de varianza de la VD que está explicada por la variable independiente (VI), que en este caso oscila entre 0,417 en el modelo 1 y 0,527 en el modelo 3. Conviene subrayar que estos valores del análisis de regresión no permiten afirmar que las relaciones detectadas entre la VD y las VIs sean de tipo causal sino que tan sólo muestran el grado de relación.

Los valores de la R cuadrado corregida, cuarta columna, son una corrección a la baja de R^2, ya que en su cálculo se tiene en cuenta el número de casos y de VIs.

En la siguiente columna, el error típico de la estimación es la raíz cuadrada de la media cuadrática residual, es decir, la desviación típica de las distancias existentes entre las puntuaciones en la 2ª Evaluación y los pronósticos efectuados con la recta de regresión. La disminución del error típico en cada modelo indica la mejora en el ajuste.

Los estadísticos de cambio nos permiten contrastar la hipótesis de que el cambio en R^2 vale cero en la población. Con la primera variable predictora, C-test 4, Modelo 1, el valor de R^2 es 0,417. Al contrastar la hipótesis de que R^2 es cero se obtiene un estadístico F de 113,912, que con 1 y 160 grados de libertad, tiene una probabilidad asociada de 0,000. Puesto que este valor es <0,05, puede afirmarse que la varianza explicada por el C-test 4, Modelo 1, es significativamente distinta de cero.

En el segundo paso, Modelo 2, el R^2 aumenta hasta 0,082 y el valor del estadístico F es 25,908, que con 1 y 159 grados de libertad, tiene una probabilidad asociada de 0,000.

En el tercer paso, Modelo 3, el valor del cambio en R^2 es de 0,027 y, aunque el estadístico de cambio en F sigue siendo significativo (9,113), sólo contribuye a explicar de forma significativa (0,003) el comportamiento de la VD.

Estos valores nos llevan a concluir que las tres VIs: C-test2, C-test3, y C-test 4, seleccionadas en el modelo final, consiguen explicar el 52,7 % de la variabilidad observada en la 2ª Evaluación (VD).

La tabla resumen del ANOVA de la variable dependiente 2ª Evaluación (Tabla 10.2) refleja el valor del estadístico F obtenido al contrastar la hipótesis de que el valor poblacional de R^2 en cada paso es cero. Ahora no se valora el cambio que se va produciendo en el valor de R^2 de un paso a otro, sino el valor de R^2 en cada paso.

Aunque sólo mostramos el tercer paso, Modelo 3, la relación es significativa en cada uno de los pasos según se observa en los valores de los estadísticos F y la probabilidad asociada. El valor del nivel crítico (*Sig*. = 0,000), al ser menor que 0,05, indica que, además de existir una relación lineal significativa, el hiperplano definido por la ecuación de regresión tiene un buen ajuste a la nube de puntos.

Tabla 10.2. Tabla resumen del ANOVA de la variable dependiente 2ª Evaluación

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
3	Regresión	360,550	3	120,183	58,291	,000c
	Residual	323,698	157	2,062		
	Total		160			

c. Variables predictoras: (Constante), CTEST4, CTEST2, CTEST3

La Tabla 10.3, que aparece a continuación, muestra los coeficientes de la recta de regresión parcial en el Modelo 3.

Tabla 10.3. Coeficientes: Variable dependiente: 2ª Evaluación

		Coeficientes no estandarizados		Coeficientes estandarizados		
Modelo		B	Error típ.	Beta	t	Sig.
3	(Constante)	,195	,479		,407	,685
	CTEST4	,138	,036	,304	3,840	,000
	CTEST2	,176	,040	,325	4,440	,000
	CTEST3	,088	,029	,217	3,019	,003

Los coeficientes de regresión parcial correspondientes a cada una de las variables incluidas en el modelo de regresión sirven para construir la ecuación de regresión en cada paso. Las primeras columnas recogen el valor de los coeficientes de regresión parcial y su error típico.

A continuación aparecen los coeficientes de regresión parcial estandarizados (*Beta*), los cuales informan acerca de la importancia relativa de cada variable dentro de la ecuación. Las dos últimas columnas muestran los estadísticos t y los niveles críticos de significación obtenidos al contrastar la hipótesis de que los coeficientes de regresión parcial valen 0 en la población.

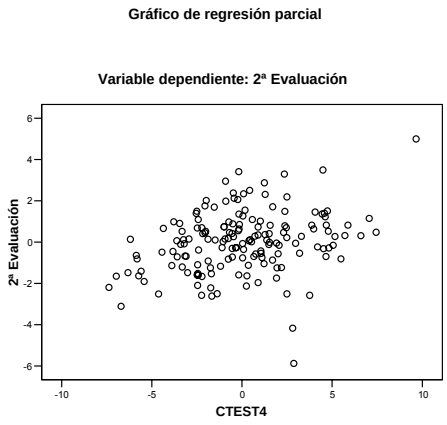
En consecuencia, los coeficientes *Beta* dan una pista sobre la importancia relativa de cada variable independiente, a mayor coeficiente estandarizado, mayor peso. En este caso, vemos que el C-test 4 es el de mayor coeficiente (de 0,646 a 0,304 cuando tenemos en cuenta las tres variables independientes), mientras que el

C-test 3 presenta el menor coeficiente *Beta* (0,217). Es el subtest que introduce el cambio en el formato de las omisiones y la última variable para explicar la varianza. Debido a los efectos de interacción entre los factores, el coeficiente del C-test 2 en el Modelo 3 es el más alto de los tres (0,325); supera incluso al del C-test 4. El nivel de significación <0,05, en las VIs (0,000) indica que las tres variables utilizadas poseen coeficientes significativamente distintos de cero y que todas contribuyen a explicar lo que ocurre con la variable dependiente, de hecho, entre las tres la explican en un 94,6%, como vemos al sumar los coeficientes *Beta* de los subtests en el modelo 3.

El comportamiento de cada subtest en relación con la 2ª Evaluación queda también reflejado en los diagramas de regresión parcial. A través de esta representación plástica de cada subtest nos formamos una idea rápida del tipo de relación con la VD. Los diagramas no se basan en las puntuaciones originales de las dos variables representadas, sino en los residuos obtenidos al realizar el análisis de regresión con las variables independientes. Muestran la relación *neta* entre las variables representadas, porque se controla el efecto de todas ellas.

Se obtienen tantos gráficos como VIs; en nuestro estudio son C-test 2, C-test 3 y C-test 4. En los tres gráficos de regresión parcial de los subtests (VIs) se observa que la relación entre la VD y las VIs es positiva. Reflejan unas líneas de regresión aceptables que indican que los subtests discriminan entre los sujetos. A continuación se muestra el del C-test 4 a modo de ejemplo.

Figura 10.1. G. de dispersión: C-test 4



Los residuos del modelo estadístico son las diferencias entre los valores observados y los pronosticados. Informan sobre el grado de exactitud de los

pronósticos; a menor error típico de los residuos, mejores pronósticos y por tanto mejor ajuste de la recta de regresión a los puntos del diagrama de dispersión.

En nuestro modelo, el valor del rango del residuo tipificado está entre –4,370 y 2,548, lo que indica un buen ajuste, que queda reflejado en los gráficos siguientes. En primer lugar, observamos el histograma de los residuos tipificados de la 2ª Evaluación (VD), su gráfico de dispersión y el de probabilidad acumulada. Cuando los residuos se distribuyen normalmente, como en este caso, la nube de puntos se encuentra alineada sobre la diagonal del gráfico.

Fig. 10.2. Histograma de los residuos de la VD

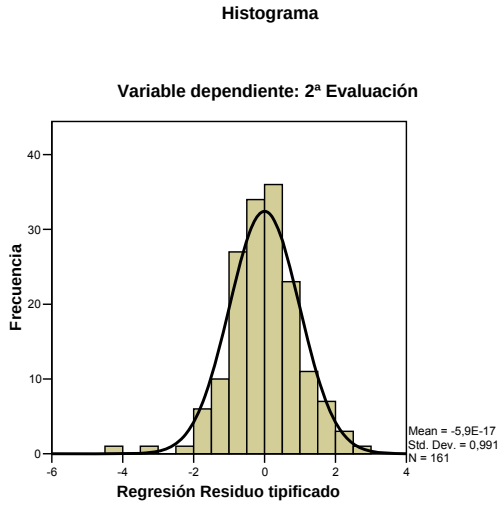


Fig. 10.3. Gráfico de dispersión de la VD

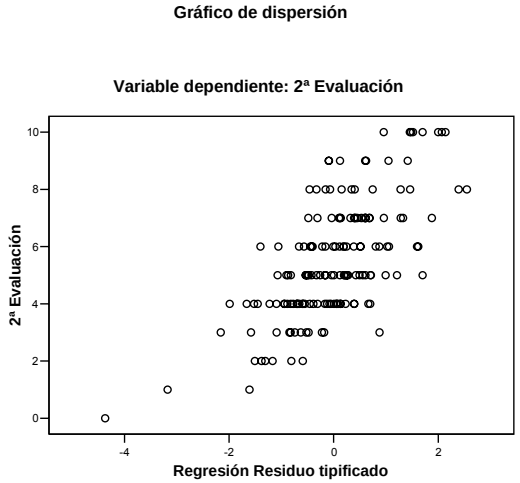
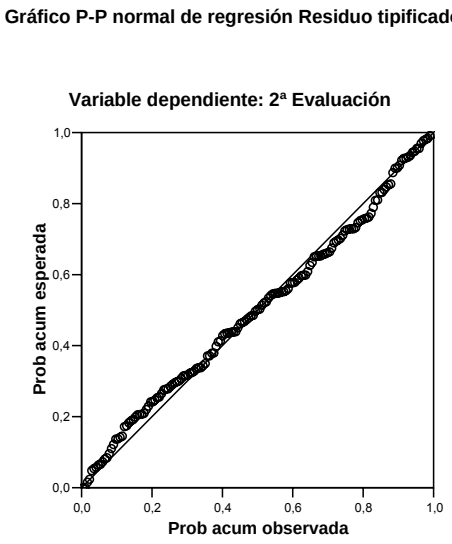


Figura 10.4. Gráfico de probabilidad acumulada del C-test



Después del análisis realizado, podemos afirmar que la prueba de regresión lineal corrobora el buen funcionamiento del C-test como predictor de los resultados obtenidos en Inglés en la 2ª Evaluación. Nos sirve, además, para determinar que, en concreto, el subtest C-test 4 es el que mejor predice los resultados de los sujetos, y que, por el contrario, el C-test 1 no contribuye a explicar la varianza, lo cual no implica que se pudiera haber prescindido de él como predictor de la VD 2ª Evaluación en el diseño de la prueba. Podemos pensar que se debe a que supuso el primer contacto de los alumnos con la técnica del C-test, mientras que, tanto en el C-test 2 como en el C-test 4, los mejores predictores en el análisis de regresión, ya había un aprendizaje previo motivado por la práctica. Por otra parte, probablemente la explicación de los valores del C-test 3 hay que buscarla en el cambio de formato y las características textuales. La introducción de omisiones no guiadas, junto a la densidad y dificultad temática de los textos pueden haber sido la causa de que su contribución a la explicación de la varianza sea la más baja de los tres subtests.

Así pues, los resultados del primer análisis de regresión nos hacen volver a algunas cuestiones ya discutidas a lo largo de este capítulo y el precedente. A pesar de la incidencia de la práctica previa, resulta llamativo que el subtest que mejor explica la varianza de las calificaciones en la 2ª Evaluación sea uno de los que tiene omisiones no guiadas (C-test 4). Si revisamos los textos sobre los que se diseñó el C-test 4 en los modelos A y B, *Women doctors* y *Evolution* respectivamente, veremos que sus características pueden influir en los resultados (véase capítulo 9).

10.3. Análisis de regresión lineal de *Cavemen?*

En este apartado continuamos con el análisis de regresión lineal. Ahora tomamos la prueba *Cavemen?*, modelo PAAU realizada en el aula por todos los sujetos de la muestra, como variable dependiente (VD) y vemos cómo se relaciona con los distintos subtests del C-test aplicado (VIs). Como en el apartado anterior, partimos de las correlaciones de Pearson entre los distintos elementos. La correlación observada entre el C-test total y la prueba *Cavemen?* es de 0,750, la mayor obtenida en el estudio. Pero además, las correlaciones son significativas para todos los subtests, y especialmente altas en el C-test 4 (0,680).

La tabla resumen del modelo incluye en esta ocasión a todos los subtests de la prueba (Tabla 10.4). El C-test 4 es, de nuevo, el que mejor explica o predice la VD, *Cavemen?.* Por eso el sistema parte de él, sin embargo, el C-test 1 sólo aparece en el último modelo. En el modelo 1 el coeficiente de correlación múltiple ya es de 0,680 y alcanza el 0,753 en el modelo 4 con la entrada de la variable independiente C-test 1. El valor R cuadrado expresa la proporción de varianza explicada por el modelo. A partir de los valores obtenidos podemos concluir que las cuatro VIs seleccionadas en el modelo 4 explican el 56,8 de la VD. El cuadro resumen del ANOVA, con los valores de los estadísticos F y la probabilidad asociada, muestra que la relación es significativa en todos los pasos (Tabla 10.5).

Tabla 10.4. Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,680(a)	,463	,459	1,47893
2	,724(b)	,525	,519	1,39538
3	,743(c)	,552	,544	1,35864
4	,753(d)	,568	,557	1,33922

a Variables predictoras: (Constante), CTEST4
b Variables predictoras: (Constante), CTEST4, CTEST2
c Variables predictoras: (Constante), CTEST4, CTEST2, CTEST3
d Variables predictoras: (Constante), CTEST4, CTEST2, CTEST3, CTEST1
e Variable dependiente: Cavemen

Tabla 10.5. Tabla resumen del ANOVA

Modelo		Suma de cuadrados	Gl	Media cuadrática	F	Sig.
1	Regresión	301,148	1	301,148	137,683	,000a
	Residual	349,960	160	2,187		
	Total	651,107	161			
2	Regresión	341,520	2	170,760	87,700	,000b
	Residual	309,587	159	1,947		
	Total	651,107	161			
3	Regresión	359,453	3	119,818	64,910	,000c
	Residual	291,654	158	1,846		
	Total	651,107	161			
4	Regresión	369,527	4	92,382	51,509	,000d
	Residual	281,580	157	1,794		
	Total	651,107	161			

a Variables predictoras: (Constante), CTEST4
b Variables predictoras: (Constante), CTEST4, CTEST2
c Variables predictoras: (Constante), CTEST4, CTEST2, CTEST3
d Variables predictoras: (Constante), CTEST4, CTEST2, CTEST3, CTEST1
e Variable dependiente: Cavemen

Por otra parte, los coeficientes de la recta de regresión parcial aparecen a continuación, en la Tabla 10.6.

Tabla 10.6. Coeficientes

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error tip.	Beta	t	Sig.
1	(Constante)	1,809	,331		5,463	,000
	CTEST4	,301	,026	,680	11,734	,000
2	(Constante)	,250	,464		,540	,590
	CTEST4	,210	,031	,475	6,695	,000
	CTEST2	,170	,037	,323	4,554	,000
3	(Constante)	,350	,452		,773	,441
	CTEST4	,166	,034	,374	4,916	,000
	CTEST2	,143	,037	,271	3,823	,000
	CTEST3	,085	,027	,216	3,117	,002
4	(Constante)	,294	,447		,658	,511
	CTEST4	,131	,036	,296	3,613	,000
	CTEST2	,101	,041	,193	2,486	,014
	CTEST3	,108	,029	,276	3,784	,000
	CTEST1	,067	,028	,169	2,370	,019

a. Variable dependiente: Cavemen

Finalmente, podemos observar el histograma de la regresión del residuo tipificado de la variable dependiente, la prueba tipo PAAU *Cavemen?*, y su gráfico de dispersión. En él, la nube de puntos tiende a alinearse sobre la diagonal.

Fig. 10.5. Histograma de la VD

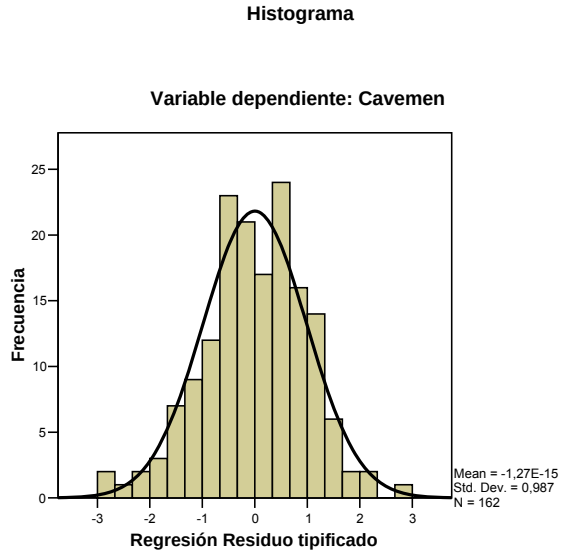
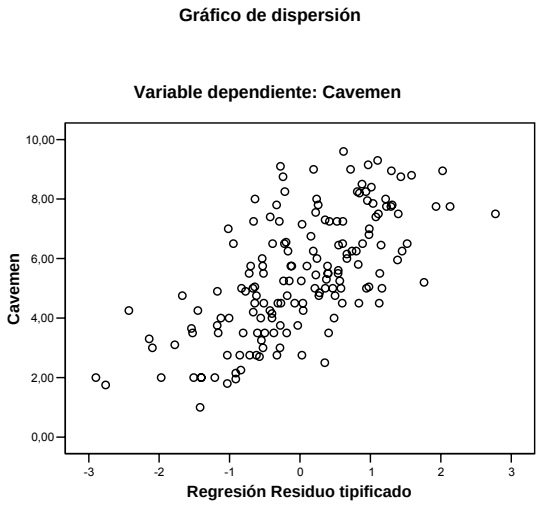


Fig. 10.6. Gráfico de dispersión de la VD



10.4. Análisis de regresión lineal de la Selectividad de junio de 2001

Con el presente apartado culminamos los análisis de regresión lineal. Hasta ahora hemos tomado la 2ª Evaluación y la prueba *Cavemen?* como variables dependientes y, como variables independientes, los subtests del C-test. Por último, este tipo de análisis nos dará pistas sobre la relación existente entre los subtests del C-test y la prueba de Inglés de Selectividad (PAAU) de junio de 2001, que hemos tomado como referencia externa en el proceso de validación del C-test.

Hemos visto anteriormente que la correlación entre el C-test y la Selectividad es también muy alta (0,722). Como en los casos anteriores, las correlaciones son altas para todos los subtests; la mayor, de nuevo, para el C-test 4 (0,621). La matriz de correlaciones bivariadas muestra el número de casos sobre el que se calcula cada coeficiente (N), en este caso, los 81 sujetos presentados a las PAAU.

En el apartado 9.9 del capítulo 9 se comentó que los promedios conseguidos por los 81 sujetos de la muestra son más elevados, se constata su mayor competencia (6,32 en la Selectividad y 5,75 en el C-test).

En cuanto al análisis de regresión, en la Tabla 10.7, que resume el modelo, podemos observar cómo se explica la varianza de la prueba de Selectividad de 2001 (VD). El coeficiente de correlación múltiple para cada paso va de 0,621 en el modelo 1 a 0,730 en el 3.

De nuevo, vemos que el sistema parte del C-test 4 como la variable que mejor predice la VD. En esta ocasión el C-test 2 no entra en el modelo. El valor R^2 muestra la proporción de varianza explicada en cada modelo; de 0,378 (37,8%) a 0,515 (51,5%).

Tabla 10.7. Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,621(a)	,385	,378	1,32648
2	,688(b)	,473	,459	1,23658
3	,730(c)	,533	,515	1,17076

a Variables predictoras: (Constante), CTEST4
b Variables predictoras: (Constante), CTEST4, CTEST1
c Variables predictoras: (Constante), CTEST4, CTEST1, CTEST3
d Variable dependiente: Selectividad 2001

La tabla 10.8, resumen del ANOVA, indica el valor del estadístico F en el tercer paso, Modelo 3.

Tabla 10.8. ANOVA: Variable dependiente: Selectividad 2001

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
3	Regresión	120,645	3	40,215	29,340	,000c
	Residual	105,542	77	1,371		
	Total	226,187	80			

c. Variables predictoras: (Constante), CTEST4, CTEST1, CTEST3
d. Variable dependiente: Selectividad 2001

Los coeficientes de regresión parcial conforman la ecuación de regresión en cada paso. Vemos que el C-test 4 tiene el coeficiente más alto (0,621) y el C-test 3 el más bajo (0,291). Por el nivel de significación constatamos que las tres variables independientes explican algo de la variable dependiente.

Tabla 10.9 Coeficientes de regresión parcial: Variable dependiente: Selectividad 2001

		Coeficientes no estandarizados		Coeficientes estandarizados		
Modelo		B	Error típ.	Beta	t	Sig.
3	(Constante)	1,880	,492		3,819	,000
	CTEST4	,103	,042	,262	2,454	,016
	CTEST1	,124	,030	,385	4,150	,000
	CTEST3	,101	,032	,291	3,165	,002

a. Variable dependiente: Selectividad 2001

Los coeficientes de regresión parcial de las variables excluidas de la ecuación en cada modelo reflejan que el C-test 2 queda fuera incluso en el modelo 3.

Los residuos de un modelo estadístico son muy importantes en el análisis de regresión, porque informan sobre el grado de exactitud de los pronósticos. A menores residuos mejor ajuste de de la recta de regresión a los puntos del diagrama de dispersión. A continuación mostramos el histograma y el gráfico de dispersión de los residuos tipificados de la VD.

Figura 10.7. Histograma

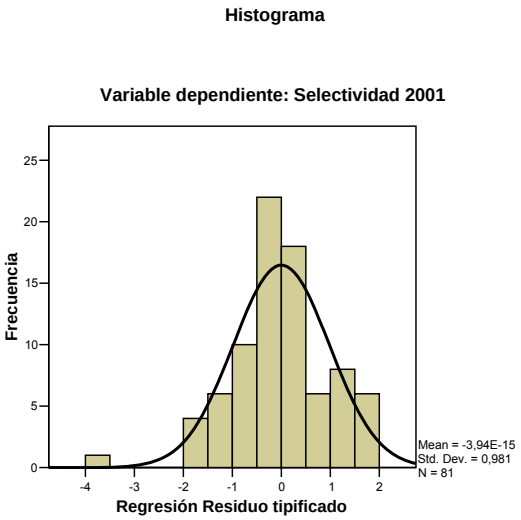
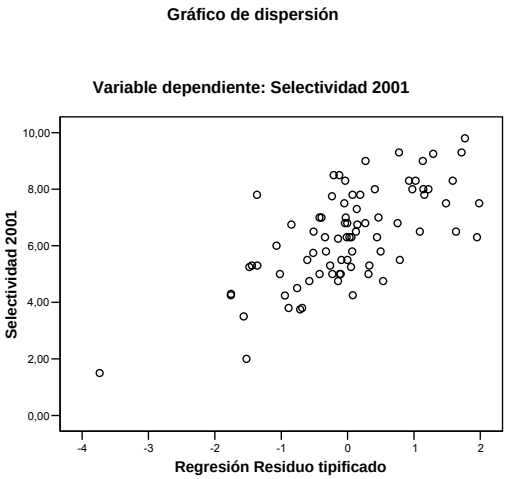


Figura 10.8. G. de dispersión



10.5. Conclusión

A modo de conclusión, podemos decir que el análisis de regresión lineal aplicado, considerando como variables dependientes (VD) a las distintas pruebas que han participado en esta investigación (2ª Evaluación, *Cavemen?* y Selectividad 2001) y como variables independientes (VIs) a los cuatro subtests del C-test, ha puesto de manifiesto que:

- Cuando tomamos las calificaciones de los sujetos en Inglés en la 2ª Evaluación como variable dependiente comprobamos que el C-test 4 es la variable independiente que mejor la predice. El C-test 1 queda excluido del modelo.
- La situación persiste, tanto si la variable dependiente es la prueba tipo Selectividad realizada en las aulas, *Cavemen?*, como la Selectividad oficial de junio de 2001. El C-test 4 de nuevo es el que presenta el mejor funcionamiento. Estos datos coinciden con los del apartado 9.3.1 (Tabla 9.3) con los que abríamos este capítulo. De los cuatro subtests, es el C-test 4 el que consigue la mejor correlación con el C-test: 0,877 (aunque no los promedios más altos).

En primer lugar, hemos de buscar las razones del buen funcionamiento del C-test 4 como predictor de todas las variables dependientes en el diseño global de la prueba. Al comenzar a realizar el C-test, el alumno se enfrenta a un formato totalmente novedoso y desconocido. Cuenta tan sólo con el modelo y las explicaciones del investigador, así pues, será la propia práctica la que le descubra lo que realmente se espera de él. A medida que se van completando los ítems y subtests se va produciendo un aprendizaje. En el subtest 3 la tarea de recuperar los textos se complica al ser omisiones no guiadas, pero en el subtest 4 deberían haberse superado ya todos los escollos propios de la técnica, para quedar únicamente las dificultades derivadas de los propios textos. Como hemos visto en el capítulo 9, en el modelo A el subtest 4 parte del texto *Women doctors* y en el modelo B de *Evolution*.

En segundo término, tendremos en cuenta los rasgos textuales. Curiosamente se obtiene una media muy semejante en ambos textos, en torno a los 12 puntos en escala de 0 a 25, al comparar los promedios por modelos. Cuando se trata de omisiones guiadas los promedios mejoran sensiblemente.

En el capítulo anterior hemos visto que ambos textos son los que presentan menor densidad y variación léxicas (Véase el apartado 9.4.2 del capítulo 9), lo que podría contribuir a facilitar su recuperación.

De este modo, el análisis de regresión lineal del C-test ha contribuido a respaldar los resultados obtenidos en los análisis anteriores.

CAPÍTULO 11. ACTUACIÓN EN EL C-TEST EN FUNCIÓN DEL ESTATUS DEMOGRÁFICO DE LOS SUJETOS

11. 1. Introducción

En el capítulo 8, dedicado a la Metodología de nuestra investigación, hemos expuesto las principales características de la muestra en la que apoyamos el trabajo empírico de la tesis, constituida por 162 sujetos, todos ellos alumnos de 2º curso de Bachillerato en distintos IES de la Comunidad de Madrid.

Atendiendo a las características de la misma distinguimos la existencia de factores demográficos que pueden llegar a determinar diferencias en la actuación de los sujetos en el C-test. Dada la homogeneidad de la muestra en variables como la edad y el nivel académico de los sujetos, en este capítulo analizaremos únicamente la incidencia de dos variables externas al C-test: el género y el IES de procedencia de los sujetos.

11.2. Incidencia de la variable de género

A pesar de que la incidencia de la variable de género no se plantea como objetivo primordial de esta investigación, intentaremos determinar si el género influye en los resultados obtenidos en el C-test. Pretendemos confirmar o rechazar, con la ayuda de las herramientas estadísticas pertinentes, la hipótesis 6:

"No habrá diferencias significativas al aplicar la variable de género".

Lo haremos mediante el análisis de los promedios obtenidos en las pruebas, el ANOVA y el modelo lineal general.

11.2.1. Características de género de la muestra y promedios obtenidos en las pruebas

Un primer acercamiento pone de manifiesto que la muestra no es equilibrada en cuanto a género, ya que participaron en el estudio 103 mujeres (64%) frente a sólo 59 varones (36%). Sin embargo, la estadística refleja la tendencia social actual; el porcentaje de mujeres que acceden a estudios superiores en España es mayor que el de los hombres, muchos de los cuales deciden incorporarse antes al mundo laboral.

Para valorar la posible incidencia de la variable género observamos los promedios obtenidos por hombres y mujeres en cada una de las pruebas aplicadas en el estudio. La tabla que aparece a continuación muestra que los resultados son siempre superiores en el caso de las mujeres, cualquiera que sea la prueba analizada (Tabla 11.1).

Tabla 11.1. Estadísticos de grupo: promedios por género

Género		N	Media	Desviación típ.	Error típ. de la media
Varón	**CTESTTOTAL**	**59**	**4,863**	**1,2979**	**,1690**
	2ª Evaluación	58	4,84	1,795	,236
	Selectividad 2001	25	5,8896	1,23310	,24662
	Cavemen?	59	4,9347	1,95845	,25497
Mujer	**CTESTTOTAL**	**103**	**5,255**	**1,5455**	**,1523**
	2ª Evaluación	103	5,86	2,128	,210
	Selectividad 2001	56	6,5196	1,82325	,24364
	Cavemen?	103	5,7427	1,99056	,19614

(Los datos están calculados en escala de puntuaciones de 0 a 10)
(Los datos correspondientes al C-test aparecen destacados en negrita)

Comprobamos que las diferencias son bastante amplias en algunos casos. La más acusada se aprecia en la calificación de Inglés de la 2ª Evaluación, en torno a un punto en escala de 0 a 10. La menor, en el C-test (sólo 0,39 puntos en la misma escala).

11.2.2. Repercusiones de la variable género en el C-test: modelos y subtests

El C-test es la prueba que presenta las menores diferencias de género en los promedios, aunque sigue la tendencia general: media de 4,86 puntos obtenida por los varones frente a los 5,25 puntos de las mujeres (en escala de 0 a 10). En el caso del C-test esto supone que, como media, las mujeres del estudio recuperan de forma correcta aproximadamente cuatro omisiones más que los varones. Hay que notar, no obstante, que la desviación típica es ligeramente superior en las mujeres, lo que indica mayor dispersión de puntuaciones.

Si analizamos lo ocurrido en cada modelo de C-test: A y B, veremos que se mantiene la pauta comentada previamente. Las tablas siguientes muestran los datos estadísticos para cada subtest de los modelos A (Tablas 11.2 y 11.3) y B (Tablas 11.4 y 11.5).

Tabla 11.2. Estadísticos descriptivos por género. Modelo A: C-test y subtests.

Descriptivos

		N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
						Límite inferior	Límite superior		
CTEST1	varón	31	15,81	3,754	,674	14,43	17,18	9	24
	mujer	50	16,36	4,365	,617	15,12	17,60	5	24
	Total	81	16,15	4,126	,458	15,24	17,06	5	24
CTEST2	varón	31	15,55	3,434	,617	14,29	16,81	9	23
	mujer	50	16,48	4,092	,579	15,32	17,64	10	25
	Total	81	16,12	3,858	,429	15,27	16,98	9	25
CTEST3	varón	31	6,84	4,670	,839	5,13	8,55	1	17
	mujer	50	7,86	4,882	,690	6,47	9,25	0	20
	Total	81	7,47	4,799	,533	6,41	8,53	0	20
CTEST4	varón	31	11,00	4,626	,831	9,30	12,70	2	20
	mujer	50	12,78	4,220	,597	11,58	13,98	3	22
	Total	81	12,10	4,437	,493	11,12	13,08	2	22
CTESTTOTAL	varón	31	49,19	14,605	2,623	43,84	54,55	28	81
	mujer	50	53,48	15,119	2,138	49,18	57,78	27	86
	Total	81	51,84	14,980	1,664	48,53	55,15	27	86

(Los promedios para cada subtest se calculan en escala de 0 a 25)
(Los promedios para el total del C-test, en escala de 0 a100)

En los estadísticos descriptivos (Tabla 11.2) observamos que los promedios de las mujeres superan aproximadamente en 1 punto por subtest a los de los varones en el modelo A. En el C-test 1 se aprecia la menor diferencia, que ronda el medio punto, y en el C-test 4 la mayor. En este último, la media de las mujeres asciende

casi 2 puntos (1,78), en una escala de 0 a 25, con respecto a la obtenida por los varones. La tabla revela que aunque, en general, las mujeres obtienen mejores promedios, la dispersión de las puntuaciones es también mayor en todos los casos, excepto en el subtest 4.

Tabla 11.3. ANOVA de un factor: Género. Subtests del modelo A.

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
CTEST1	Inter-grupos	5,864	1	5,864	,342	,561
	Intra-grupos	1356,359	79	17,169		
	Total	1362,222	80			
CTEST2	Inter-grupos	16,608	1	16,608	1,117	,294
	Intra-grupos	1174,157	79	14,863		
	Total	1190,765	80			
CTEST3	Inter-grupos	19,959	1	19,959	,865	,355
	Intra-grupos	1822,214	79	23,066		
	Total	1842,173	80			
CTEST4	Inter-grupos	60,630	1	60,630	3,162	,079
	Intra-grupos	1514,580	79	19,172		
	Total	1575,210	80			
CTESTTOTAL	Inter-grupos	351,595	1	351,595	1,578	,213
	Intra-grupos	17599,319	79	222,776		
	Total	17950,914	80			

El análisis del ANOVA[73] (Tabla 11.3), que nos sirve para comparar ambos grupos (mujeres y varones), indica que, a pesar de las diferencias en los promedios, no hay diferencias significativas en la actuación en el C-test. En el subtest 4 del modelo A se aprecian un poco más (Sig. 0,79), pero tampoco son notorias; se encuentran en el umbral de la significación.

En el modelo B del C-test, que obtuvo un promedio ligeramente inferior al modelo A, las diferencias son menores (Tabla 11.4), especialmente en los subtests 1 y 2, en que los resultados obtenidos por varones y mujeres no llegan a un punto de diferencia en una escala de 0 a 25. Igual que en el modelo A, la desviación típica es levemente superior en la actuación de las mujeres.

[73] Rietveld y van Hout (2005: 1) se refieren al Analysis Of Variance (ANOVA) de este modo: “This technique is the main instrument for social scientists and their linguistic colleagues to analyze the outcomes of research designs with more than two treatments or groups. Moreover, analysis of variance enables the researcher to assess the effects of more than one independent variable at the same time”.

Tabla 11.4. Estadísticos descriptivos por género. Modelo B: C-test y subtests.

Descriptivos

		N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
						Límite inferior	Límite superior		
CTEST1	varón	28	10,00	3,590	,678	8,61	11,39	2	17
	mujer	53	10,57	4,547	,625	9,31	11,82	3	21
	Total	81	10,37	4,226	,470	9,44	11,30	2	21
CTEST2	varón	28	14,75	3,075	,581	13,56	15,94	9	20
	mujer	53	15,36	4,058	,557	14,24	16,48	4	22
	Total	81	15,15	3,739	,415	14,32	15,97	4	22
CTEST3	varón	28	12,14	3,461	,654	10,80	13,48	6	19
	mujer	53	13,21	4,129	,567	12,07	14,35	5	24
	Total	81	12,84	3,923	,436	11,97	13,71	5	24
CTEST4	varón	28	11,11	3,414	,645	9,78	12,43	5	19
	mujer	53	12,55	5,165	,709	11,12	13,97	1	24
	Total	81	12,05	4,663	,518	11,02	13,08	1	24
CTESTTOTAL	varón	28	48,00	11,139	2,105	43,68	52,32	25	69
	mujer	53	51,68	15,860	2,178	47,31	56,05	18	89
	Total	81	50,41	14,438	1,604	47,21	53,60	18	89

El análisis del ANOVA muestra de nuevo que, a pesar de las diferencias en los promedios, no hay diferencias significativas en la actuación de varones y mujeres en el C-test modelo B (Tabla 11.5). Con estos datos, en principio, podríamos ya confirmar la hipótesis 6, pero antes profundizaremos en ello mediante el modelo lineal general.

Tabla 11.5. ANOVA: subtests del modelo B

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
CTEST1	Inter-grupos	5,870	1	5,870	,326	,570
	Intra-grupos	1423,019	79	18,013		
	Total	1428,889	80			
CTEST2	Inter-grupos	6,784	1	6,784	,482	,498
	Intra-grupos	1111,439	79	14,069		
	Total	1118,222	80			
CTEST3	Inter-grupos	20,768	1	20,768	1,356	,248
	Intra-grupos	1210,146	79	15,318		
	Total	1230,914	80			
CTEST4	Inter-grupos	37,992	1	37,992	1,764	,188
	Intra-grupos	1701,811	79	21,542		
	Total	1575,210	80			
CTESTTOTAL	Inter-grupos	248,008	1	248,008	1,193	,278
	Intra-grupos	16429,547	79	207,969		
	Total	16677,556	80			

En el capítulo 9 hemos señalado que el C-test es la prueba que obtiene los promedios más bajos de las aplicadas en este estudio empírico, con independencia de la variable genérica. En este capítulo vemos que, además, es la prueba que más acerca los promedios obtenidos por varones y mujeres (véase la Tabla 11.1).

11.2.3. Análisis de promedios mediante el modelo lineal general

Para completar el análisis de los promedios y el ANOVA, veremos también la incidencia de la variable genérica mediante el modelo lineal general. En este caso tomamos como muestra sólo los 81 alumnos que realizaron la PAAU oficial, para poder tomar en consideración los datos de las cuatro pruebas (C-test, 2ª Evaluación, *Cavemen?* y PAAU de junio de 2001) en todos los sujetos. La Tabla 11.6 refleja las características de la muestra analizada.

Tabla 11.6. Muestra de los sujetos presentados a las PAAU: género

		Etiqueta del valor	N
Género	1	Varón	24
	2	Mujer	56

Aunque el análisis multivariante aplicado al género indica que, en general, en este grupo no hay diferencias significativas en la actuación de los sujetos, la Tabla 11.7 refleja datos más concretos acerca de cada prueba.

En este grupo no se aprecian diferencias significativas en cuanto al género en las pruebas de Inglés de las PAAU: ni en la Selectividad oficial de junio de 2001 ni en la prueba *Cavemen?* realizada en clase (sig.: 0,101 y 0,337), pero sí aparecen en el caso del C-test y las calificaciones en la 2ª Evaluación, puesto que 0,05 es el límite para la significación. No obstante, debemos tener en cuenta que en este análisis hemos introducido cambios. Por una parte, la muestra ha cambiado y por otra, hemos considerado el C-test globalmente y no desglosado en subtests.

Veamos la información que aporta la Tabla 11.7.

Tabla 11.7. Pruebas de los efectos inter-sujetos

Fuente	Variable dependiente	Suma de cuadrados tipo III	Gl	Media cuadrática	F	Significación
Modelo corregido	2ª Evaluación	16,010(a)	1	16,010	4,307	,041
	Selectividad 2001	7,684(b)	1	7,684	2,749	,101
	Cavemen?	2,605(c)	1	2,605	,933	,337
	Ctestt10	8,700(d)	1	8,700	4,563	,036
Intersección	2ª Evaluación	2650,060	1	2650,060	712,921	,000
	Selectividad 2001	2567,765	1	2567,765	918,565	,000
	Cavemen?	2792,980	1	2792,980	1000,377	,000
	Ctestt10	2105,376	1	2105,376	1104,119	,000
Género	2ª Evaluación	16,010	1	16,010	4,307	**,041**
	Selectividad 2001	7,684	1	7,684	2,749	,101
	Cavemen?	2,605	1	2,605	,933	,337
	Ctestt10	8,700	1	8,700	4,563	**,036**
Error	2ª Evaluación	289,940	78	3,717		
	Selectividad 2001	218,042	78	2,795		
	Cavemen?	217,770	78	2,792		
	Ctestt10	148,733	78	1,907		
Total	2ª Evaluación	3660,000	80			
	Selectividad 2001	3417,833	80			
	Cavemen?	3627,078	80			
	Ctestt10	2794,390	80			
Total corregida	2ª Evaluación	305,950	79			
	Selectividad 2001	225,726	79			
	Cavemen?	220,375	79			
	Ctestt10	157,434	79			

a R cuadrado = ,052 (R cuadrado corregida = ,040)
b R cuadrado = ,034 (R cuadrado corregida = ,022)
c R cuadrado = ,012 (R cuadrado corregida = -,001)
d R cuadrado = ,055 (R cuadrado corregida = ,043)

En consecuencia, a la luz de estos datos, cabe confirmar la hipótesis inicialmente planteada en el primer caso y rechazarla en el segundo. En una primera aproximación, tomando la muestra completa (162 sujetos) y desglosando el C-test en subtests, el análisis del ANOVA informa de que no se aprecian diferencias significativas en cuanto a la actuación de los géneros en el C-test.

Pero cuando se analiza la incidencia del género en las cuatro pruebas valoradas en nuestra investigación, y la muestra queda reducida a la mitad que incluye a los sujetos de mayor competencia, tomando el C-test en su conjunto, el modelo lineal general precisa que, a pesar de que los promedios obtenidos por las mujeres son ligeramente superiores en todas las pruebas aplicadas, en el C-test y

en la valoración de la 2ª Evaluación los estadísticos muestran diferencias significativas en la actuación de varones y mujeres.

Sería interesante, de cara a futuras investigaciones, indagar en las causas de estas diferencias que afectan a los sujetos con mayor competencia, y que podrían ir desde el uso que varones y mujeres hacen de las estrategias (Phakiti 2003), hasta la incidencia del tema (Lumley y O'Sullivan 2005), pasando por la motivación (Mori y Gobel 2006) o el tipo de tarea propuesta.

11.3. Incidencia del IES de procedencia de los sujetos

En las hipótesis de partida de esta investigación planteamos también la posible incidencia de la procedencia de los sujetos que participaron en el estudio. Los sujetos de la muestra proceden de cuatro IES de la red de centros públicos de la Comunidad de Madrid que presentan características sociodemográficas muy distintas. El número de los sujetos de cada IES varía, desde los 45 del IES Ágora de Alcobendas hasta los 36 del IES Humanejos de Parla, siempre superando los 30 sujetos en cada subgrupo.

En este apartado intentaremos determinar la validez de la hipótesis 7:

> "No se prevé que existan diferencias de funcionamiento del C-test al aplicar la variable IES"

11.3.1. Entorno de los IES en que se realizó el estudio

A pesar de que el capítulo 8 incluye el perfil de los cuatro centros a los que pertenecen los sujetos de la muestra, antes de abordar las estadísticas de los resultados obtenidos por cada uno de ellos en las pruebas aplicadas, insistiremos en la variedad de los entornos socioeconómicos en que se encuentran ubicados.

Del cuestionario retrospectivo que completaron los sujetos se extraen también datos significativos acerca de cómo afecta el entorno a las oportunidades de aprendizaje de los sujetos (véase el apartado 12.5.1 del capítulo 12). Las tablas de

frecuencias de los conocimientos previos de los sujetos muestran, por ejemplo, que los alumnos de Madrid y Alcobendas son los que más han disfrutado de estancias en el extranjero para perfeccionar la lengua. Aunque estas oportunidades extra, externas a la escuela no sean determinantes para el rendimiento y competencia de los sujetos, una vez más queremos hacer constar que no es nuestro objetivo hacer una mera comparación de resultados, sino cerciorarnos de la validez y fiabilidad del C-test en todos los contextos como prueba que discrimina a los sujetos en función de su competencia lingüística.

Se ha constatado que la correlación entre las distintas pruebas aplicadas es alta y significativa en los diferentes IES, por tanto, las diferencias que puedan aparecer al contrastarlos se explican en función de los datos sociales.

11.3.2. Análisis estadístico de los promedios de cada centro

La Tabla 11.8 expone los resultados obtenidos por cada centro en las pruebas aplicadas. Además del número de sujetos (N), en la tabla aparece la media de cada prueba y la desviación típica, es decir, la dispersión de las puntuaciones. Para facilitar el análisis, todos los datos están en escala de 0 a 10.

Tabla 11.8. Informe

Centro		C-test	2ª Evaluación	Cavemen?	Selectividad 2001
Alcobendas	Media	5,1444	6,09	5,8756	6,1759
	N	45	45	45	29
	Desv. Típ.	1,43266	2,485	2,07204	1,70433
Pinto	Media	4,5634	4,60	4,8232	5,9525
	N	41	40	41	16
	Desv. Típ.	1,49211	1,780	1,82024	1,35552
Madrid	**Media**	**6,1300**	**6,23**	**6,3412**	**7,1731**
	N	**40**	**40**	**40**	**26**
	Desv. Típ.	**1,30014**	**1,687**	**1,80544**	**1,36779**
Parla	Media	4,5750	4,94	4,6347	5,1500
	N	36	36	36	10
	Desv. Típ.	1,04809	1,672	1,87147	1,97625
Total	Media	5,1142	5,50	5,4485	6,3252
	N	162	161	162	81
	Desv. Típ.	1,47011	2,068	2,01101	1,68147

Lo primero que llama nuestra atención en la lectura de la tabla es que los promedios más altos en todas las pruebas se consiguen en el IES San Isidoro, de Madrid capital (los datos aparecen destacados en negrita).

El IES San Isidoro presentó 26 sujetos de los 40 totales a las PAAU (un 65%) y obtuvo una media de 7,17 puntos. En el C-test logró una media de 6,13, muy semejante a la de la 2ª Evaluación (6,2) y la de *Cavemen?* (6,3). Este grupo de sujetos es también el que muestra mayor uniformidad en los promedios de las pruebas analizadas. Cabe señalar que este dato indica fiabilidad y coherencia en la actuación, es decir, menores diferencias entre los resultados obtenidos en unas pruebas y otras. Además la desviación típica indica que no hay gran dispersión de puntuaciones.

Los datos nos llevan a pensar que a medida que mejora la competencia en la lengua, mejora también la actuación en el C-test (Klein-Braley 1984). Aunque Sussmich (1984) considera que el C-test es adecuado para todos los niveles, desde principiante hasta avanzado, parece que son los más hábiles en la lengua los que encuentran menos obstáculos en su realización. Lógicamente, estos sujetos manejarán mejor las estrategias. Y es posible que destaquen la validez aparente de la prueba, pero al ser un cuestionario anónimo, no contamos con ese dato.

El IES Ágora, de Alcobendas, obtiene también buenos resultados y presentó a las PAAU un 64,4 % de los alumnos del grupo analizado, cifra muy semejante a la del IES San Isidoro. En el otro extremo, los resultados obtenidos por el IES Humanejos, de Parla y el IES Vicente Aleixandre, de Pinto.

En cuanto a la desviación típica, vemos que el grupo de Parla es el más homogéneo en las puntuaciones en el C-test, pero el que más dispersión presenta en las calificaciones de Selectividad, a la que tan sólo se presentaron 10 sujetos.

Aún sin pretender comparar los resultados de los diferentes IES participantes, atendiendo a los promedios, parece claro que las características socio-económicas del entorno en que se encuentra ubicado el IES condicionan la actuación de los sujetos. Las cifras corroboran las diferencias Norte-Sur también en el ámbito de nuestra Comunidad. Pero el frío análisis estadístico ha de ser matizado teniendo en cuenta las circunstancias concretas de los centros educativos, su entorno socio-económico, cultural, etc. (véase apartado 8.4.1, capítulo 8). En caso contrario recibiríamos una información sesgada y parcial de la realidad.

11.3.3. Análisis de varianza univariante de los resultados de los centros

A pesar de las diferencias constatadas en los promedios debemos completar este estudio con el análisis de varianza univariante, puesto que la disparidad de promedios no implica necesariamente la existencia de diferencias significativas en la actuación de los centros.

En el análisis de varianza univariante se aprecian diferencias significativas entre centros en los resultados del C-test, concretamente entre el IES Ágora de Alcobendas y el IES San Isidoro, de Madrid capital. Las pruebas post-hoc muestran una significación de 0,005 en Bonferroni y 0,006 en Games-Howell (Tabla 11.9).

Tabla 11.9. Pruebas post hoc

Comparaciones múltiples

Variable dependiente: CTESTTOTAL

	(I) Centro	(J) Centro	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior
Bonferroni	Alcobendas	Pinto	5,72	2,890	,297	-2,00	13,44
		Madrid	-9,94*	2,908	,005	-17,72	-2,17
		Parla	5,58	2,993	,385	-2,42	13,57
	Pinto	Alcobendas	-5,72	2,890	,297	-13,44	2,00
		Madrid	-15,67*	2,974	,000	-23,61	-7,72
		Parla	-,14	3,057	1,000	-8,31	8,02
	Madrid	Alcobendas	9,94*	2,908	,005	2,17	17,72
		Pinto	15,67*	2,974	,000	7,72	23,61
		Parla	15,52*	3,075	,000	7,31	23,74
	Parla	Alcobendas	-5,58	2,993	,385	-13,57	2,42
		Pinto	,14	3,057	1,000	-8,02	8,31
		Madrid	-15,52*	3,075	,000	-23,74	-7,31
Games-Howell	Alcobendas	Pinto	5,72	3,156	,275	-2,55	14,00
		Madrid	-9,94*	2,959	,006	-17,70	-2,19
		Parla	5,58	2,752	,187	-1,65	12,80
	Pinto	Alcobendas	-5,72	3,156	,275	-14,00	2,55
		Madrid	-15,67*	3,107	,000	-23,82	-7,51
		Parla	-,14	2,911	1,000	-7,80	7,51
	Madrid	Alcobendas	9,94*	2,959	,006	2,19	17,70
		Pinto	15,67*	3,107	,000	7,51	23,82
		Parla	15,52*	2,697	,000	8,43	22,61
	Parla	Alcobendas	-5,58	2,752	,187	-12,80	1,65
		Pinto	,14	2,911	1,000	-7,51	7,80
		Madrid	-15,52*	2,697	,000	-22,61	-8,43

Basado en las medias observadas.

*. La diferencia de medias es significativa al nivel ,05.

En la última parte de este capítulo (apartado 11.3.5), no obstante, veremos los resultados del análisis de varianza univariante aplicado a los dos factores en cuestión: género y centro.

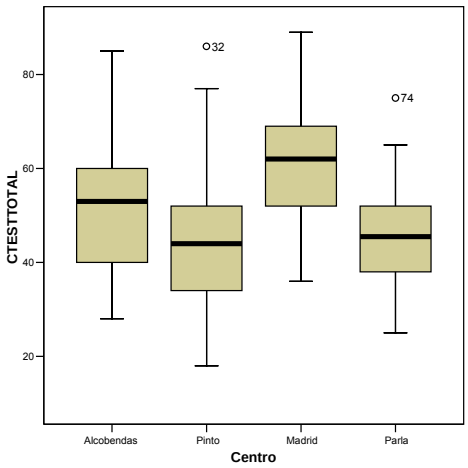
11.3.4. Repercusiones de la variable IES de procedencia en el C-test

A continuación centramos nuestro análisis en el C-test, buscando respuesta a última parte de la pregunta de investigación: "¿Qué influencia ejercen las variables género, formación previa y centro de estudios de los sujetos en los resultados obtenidos?", que ha de confirmar o rechazar la hipótesis 7:

> "No se prevé que existan diferencias de funcionamiento del C-test al aplicar la variable IES".

Veamos un gráfico general de los resultados del C-test por centros:

Figura 11.1. Diagrama de cajas: Promedios del C-test por centros



El diagrama de cajas (Fig. 11.1) muestra lo ya constatado en este apartado; que el promedio más alto se logra en el IES de Madrid capital, seguido por el de Alcobendas, mientras que los IES de Parla y Pinto obtienen promedios bastante inferiores y muy próximos entre sí. En ambos centros se detecta la presencia de una puntuación extrema (*outlier*) que destaca del resto.

Esto nos lleva a rechazar la hipótesis 7: "no hay diferencias de funcionamiento en el C-test al aplicar la variable IES" cuando hablamos de centros de distinto

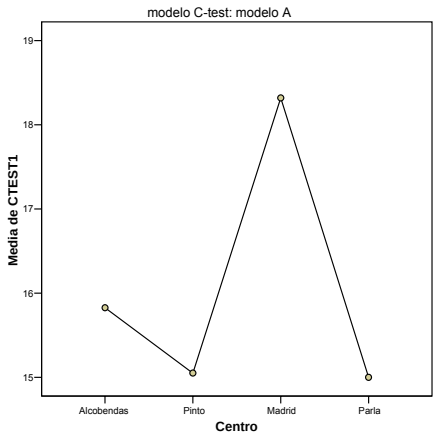
estatus. Sin embargo, podríamos confirmarla en centros de estatus similar, puesto que es el IES de Madrid capital el que se diferencia de los demás. A pesar de que al aplicar la variable de centro (por su ubicación y/o características) sí hay diferencias significativas en los resultados obtenidos en el C-test, éstas son comunes a todas las pruebas aplicadas y no sólo al C-test. El C-test sigue las mismas pautas de funcionamiento que cualquier otra prueba, y discrimina igual entre los alumnos, independientemente de las características del IES en que se aplique, como indica la alta correlación entre las pruebas. No obstante, en el siguiente epígrafe aplicaremos el análisis de varianza univariante a las variables “género” e “IES”.

En la caja correspondiente al IES de Alcobendas vemos que la mediana no se corresponde exactamente con la media y que hay mayor dispersión de puntuaciones entre los sujetos, sobre todo en los que están por debajo de la media (Fig. 11.1). Sin embargo, el grupo de Parla es muy homogéneo, es el que presenta menor dispersión de puntuaciones, la mediana y la media prácticamente coinciden.

No consideramos necesario el análisis más concreto de los resultados que se obtienen en el C-test en cada centro para los distintos modelos y subtests, puesto que no aportan datos de interés para nuestra investigación.

Tanto en el modelo A como en el B, y en todos los subtests, los promedios más altos corresponden al IES San Isidoro. En el C-test 1 del modelo A, la diferencia entre promedios supera los 3 puntos (de 15,00 puntos en Parla a 18,31 en Madrid, en escala de 0 a 25). En los C-tests 3 y 4 hay aún mayor distancia entre promedios. Los estadísticos descriptivos para el modelo B son semejantes a los del A. Como ejemplo, la comparativa de promedios por centros en el subtest 1 del modelo A.

Figura 11.2. Comparativa de promedios por centros: C-test 1 modelo A



11.3.5. Análisis de varianza univariante de ambas variables

Una vez analizados ambos factores por separado, afrontamos el análisis de varianza univariante siguiendo el modelo lineal general. Tomaremos el total del C-test como variable dependiente y veremos cómo afectan el género y el centro de los sujetos a la prueba. Una primera aproximación es el resumen de promedios en el C-test, desglosado en centros y género.

Tabla 11.10. Estadísticos descriptivos

Centro	Género	Media	Desv. Típ.	N
Alcobendas	Varón	49,13	12,258	16
	Mujer	52,59	15,340	29
	Total	51,36	12,275	45
Pinto	Varón	46,07	12,363	14
	Mujer	45,41	16,308	27
	Total	45,63	14,921	41
Madrid	Varón	62,22	12,882	9
	Mujer	61,03	13,235	31
	Total	61,30	13,001	40
Parla	Varón	43,90	10,249	20
	Mujer	48,13	10,595	16
	Total	45,78	10,472	36
Total	Varón	48,63	12,979	59
	Mujer	52,55	15,455	103
	Total	51,12	14,683	162

* Los datos de la tabla están en escala de 0 a 100

En la tabla aparece, además, el número total de sujetos de cada subgrupo (N) y la desviación típica. Podemos hacer una lectura de la tabla que no tenga en cuenta el género, sino sólo los promedios totales de cada centro en el C-test. Esta información ya la obtuvimos en la Tabla 11.8. El IES San Isidoro destaca frente al resto de los centros con una media de 61,30 en escala 0 a 100.

El desglose por género aporta precisión. Nos muestra que, a pesar de los resultados totales, la actuación de las mujeres no siempre es mejor que la de los hombres. De hecho, en dos centros; el IES V. Aleixandre de Pinto y el S. Isidoro de Madrid, los varones obtienen resultados ligeramente superiores en el C-test (véase Fig. 11.3). No en vano, el C-test es la prueba de las estudiadas que presenta menores diferencias genéricas. Estudios posteriores, fuera ya del alcance de esta tesis, podrían encaminarse a determinar las causas de este hecho.

Por otra parte, la dispersión de puntuaciones es, curiosamente, siempre superior en las mujeres (aunque en el IES Humanejos se obtienen valores muy cercanos).

En la siguiente tabla vemos las pruebas de los efectos intersujetos que determinan el grado de significación de las variables género y centro.

Tabla 11.11. Pruebas de los efectos intersujetos. Variable dependiente: CTESTTOTAL

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	6705,226[a]	7	957,889	5,267	,000
Intersección	361675,713	1	361675,713	1988,769	,000
Centro	5010,739	3	1670,246	9,184	,000
Género	73,734	1	73,734	,404	,525
Centro * género	198,383	3	66,128	,364	,779
Error	28006,305	154	181,859		
Total	458116,000	162			
Total corregida	34711,531	161			

a. R cuadrado = ,193 (R cuadrado corregida = ,156)

Se aprecia que no hay interrelación entre las variables centro y género, son independientes. Mientras que sí hay diferencias significativas en los resultados de los centros, no las hay en el género.

Al hacer las comparaciones por pares (Tabla 11.12), teniendo en cuenta los centros, encontramos que el IES San Isidoro, de Madrid capital, es el que presenta diferencias significativas con los otros tres que forman parte del estudio (sig. = 0,008 con Alcobendas y 0,000 con Pinto y Parla). Recordemos la acusada diferencia entre el promedio de los varones del IES San Isidoro (62,22 puntos) y del IES Humanejos de Parla (43,90 puntos) en el C-test, que se lee en la Tabla 11.10. Entre los centros de Alcobendas, Pinto y Parla no se aprecian diferencias significativas.

Tabla 11. 12. Comparaciones por pares: Centros

Comparaciones por pares

Variable dependiente: CTESTTOTAL

(I) Centro	(J) Centro	Diferencia entre medias (I-J)	Error típ.	Significación[a]	Intervalo de confianza al 95 % para diferencia[a]	
					Límite inferior	Límite superior
Alcobendas	Pinto	5,116	3,056	,577	-3,052	13,285
	Madrid	-10,772*	3,306	,008	-19,607	-1,936
	Parla	4,843	3,086	,712	-3,405	13,092
Pinto	Alcobendas	-5,116	3,056	,577	-13,285	3,052
	Madrid	-15,888*	3,384	,000	-24,932	-6,844
	Parla	-,273	3,170	1,000	-8,745	8,198
Madrid	Alcobendas	10,772*	3,306	,008	1,936	19,607
	Pinto	15,888*	3,384	,000	6,844	24,932
	Parla	15,615*	3,411	,000	6,499	24,731
Parla	Alcobendas	-4,843	3,086	,712	-13,092	3,405
	Pinto	,273	3,170	1,000	-8,198	8,745
	Madrid	-15,615*	3,411	,000	-24,731	-6,499

Basadas en las medias marginales estimadas.

*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Bonferroni.

Por último, mostramos los resultados de las pruebas de Tuckey, Bonferroni y Games-Howel (Tabla 11.13, en la página siguiente).

La Figura 11.3 refleja gráficamente el resumen de los datos analizados en este capítulo.

Figura 11.3. Medias estimadas del C-test por géneros en los distintos IES

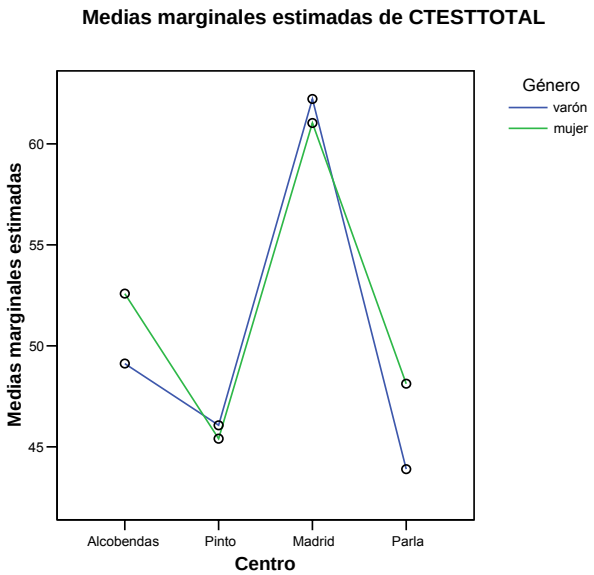


Tabla 11.13. Comparaciones múltiples

Comparaciones múltiples

Variable dependiente: CTESTTOTAL

	(I) Centro	(J) Centro	Diferencia entre medias (I-J)	Error típ.	Significación	Intervalo de confianza al 95%. Límite inferior	Límite superior
DHS de Tukey	Alcobendas	Pinto	5,72	2,912	,206	-1,84	13,28
		Madrid	-9,94*	2,930	,005	-17,56	-2,33
		Parla	5,58	3,015	,254	-2,25	13,41
	Pinto	Alcobendas	-5,72	2,912	,206	-13,28	1,84
		Madrid	-15,67*	2,997	,000	-23,45	-7,88
		Parla	-,14	3,080	1,000	-8,14	7,86
	Madrid	Alcobendas	9,94*	2,930	,005	2,33	17,56
		Pinto	15,67*	2,997	,000	7,88	23,45
		Parla	15,52*	3,098	,000	7,48	23,57
	Parla	Alcobendas	-5,58	3,015	,254	-13,41	2,25
		Pinto	,14	3,080	1,000	-7,86	8,14
		Madrid	-15,52*	3,098	,000	-23,57	-7,48
Bonferroni	Alcobendas	Pinto	5,72	2,912	,307	-2,06	13,50
		Madrid	-9,94*	2,930	,005	-17,78	-2,11
		Parla	5,58	3,015	,398	-2,48	13,64
	Pinto	Alcobendas	-5,72	2,912	,307	-13,50	2,06
		Madrid	-15,67*	2,997	,000	-23,68	-7,66
		Parla	-,14	3,080	1,000	-8,38	8,09
	Madrid	Alcobendas	9,94*	2,930	,005	2,11	17,78
		Pinto	15,67*	2,997	,000	7,66	23,68
		Parla	15,52*	3,098	,000	7,24	23,80
	Parla	Alcobendas	-5,58	3,015	,398	-13,64	2,48
		Pinto	,14	3,080	1,000	-8,09	8,38
		Madrid	-15,52*	3,098	,000	-23,80	-7,24
Games-Howell	Alcobendas	Pinto	5,72	3,156	,275	-2,55	14,00
		Madrid	-9,94*	2,959	,006	-17,70	-2,19
		Parla	5,58	2,752	,187	-1,65	12,80
	Pinto	Alcobendas	-5,72	3,156	,275	-14,00	2,55
		Madrid	-15,67*	3,107	,000	-23,82	-7,51
		Parla	-,14	2,911	1,000	-7,80	7,51
	Madrid	Alcobendas	9,94*	2,959	,006	2,19	17,70
		Pinto	15,67*	3,107	,000	7,51	23,82
		Parla	15,52*	2,697	,000	8,43	22,61
	Parla	Alcobendas	-5,58	2,752	,187	-12,80	1,65
		Pinto	,14	2,911	1,000	-7,51	7,80
		Madrid	-15,52*	2,697	,000	-22,61	-8,43

Basado en las medias observadas.

*. La diferencia de medias es significativa al nivel ,05.

CAPÍTULO 12: ANÁLISIS DE LA VALIDEZ APARENTE DEL C-TEST

12.1. Introducción

El aspecto externo que presenta una prueba es, en palabras de Bachman (1990:289), “a very important consideration in test use”. Como en tantas otras cosas y situaciones, la primera información que percibimos acerca de una prueba es su apariencia física. A partir de ella los sujetos reaccionarán de una u otra forma. Y si el alumno no tiene confianza en la prueba como instrumento de medida, no hará un esfuerzo serio en su realización (Jafarpur 1995). A pesar de que algunos autores, como Anastasi (1982), la infravaloren, pensamos que la validez aparente de un examen no es una mera cuestión de preferencias o gustos personales, ni de estética o moda, sino un rasgo de las pruebas que debemos estudiar con detenimiento.

El C-test presenta un aspecto que contrasta con el de otras pruebas, a pesar de la popularidad de algunos tipos de pruebas de cierre en la enseñanza del Inglés. Resulta novedoso, distinto, y esta circunstancia puede ser, en ocasiones, un motivo de rechazo inicial hacia la prueba. No en vano, el ser humano tiende a ofrecer resistencia al cambio en todos los ámbitos de la vida.

Desde la creación del C-test (Klein-Braley y Raatz 1981) distintas voces se han alzado para cuestionar su validez aparente (Jafarpur 1995; Weir 1988; Bradshaw 1990). El formato del C-test ha sido tachado de fragmentario, “*puzzle-like*”, e inadecuado para su propósito.

Dentro del proceso de validación del C-test, una vez probada su validez criterial concurrente y de constructo en el capítulo 9, en esta tesis nos disponemos a analizar la validez aparente de la prueba a partir de los resultados del cuestionario retrospectivo de opinión administrado a los 162 alumnos de la muestra (véase el capítulo 8, apartado 8.4.4) para confirmar o rechazar la hipótesis 5:

> "Por su novedad y su carácter fragmentario, algo confuso al principio, puede conducir al rechazo. El C-test carece de validez aparente."

12.2. La validez aparente del C-test en los estudios piloto

El capítulo 7 de la tesis desarrolla los primeros contactos con el C-test a través de la aplicación de dos pruebas piloto. Como hemos visto, en la primera investigación (Estudio piloto I) no se estudió la validez aparente de manera formal, pero sí se sondeó de manera informal en el aula, escuchando las manifestaciones directas y espontáneas de los alumnos acerca de la prueba. Fueron el punto de partida para que en trabajos posteriores nos planteáramos la creación de un cuestionario de opinión que permitiera la valoración objetiva de este aspecto.

En la Prueba piloto II se diseñó el primer modelo de cuestionario de opinión para el alumnado. Muy semejante ya al modelo definitivo utilizado en nuestra investigación (véase el Apéndice), se basaba en el ideado por Jafarpur (1995).

No obstante, el cuestionario retrospectivo fue tomado únicamente como elemento informativo que recopilaba las impresiones del alumnado.

En las dos ocasiones se apreció que los alumnos, a pesar de enfrentarse a una prueba nueva y diferente, lejos de rechazarla, manifestaban su aceptación de la misma. En la investigación definitiva sobre el C-test era preciso disponer de elementos de juicio objetivos y fiables. Se decidió administrar un cuestionario ya que este procedimiento permite recopilar cuantiosa información de manera rápida y sencilla. Su análisis estadístico posterior servirá para determinar la validez aparente de la prueba.

12.3. El cuestionario: partes y orígenes

Para la confección del cuestionario retrospectivo de opinión acerca de la prueba tomamos como referencia el propuesto por Jafarpur (1995: 207).

El autor planteó un cuestionario sobre la validez aparente del C-test. Lo administró a un grupo de alumnos y a sus profesores. En ambos casos obtuvo

resultados negativos, sobre todo por parte de los alumnos (64%), que le llevaron a considerar al C-test como carente de validez aparente: “C-testing does not fulfil the exigency of this requirement” (ibíd.).

Como mencionamos en la introducción del capítulo, Jafarpur (1995) considera que si los alumnos no valoran bien una prueba, no la toman en serio y, en consecuencia, no se esfuerzan en su realización. En su investigación descubrió que ni siquiera los profesores confiaban en la técnica. Aunque la valoraron mejor que los alumnos (respondieron positivamente al 57% de las preguntas), únicamente dos de los profesores manifestaron que la aceptarían como criterio selectivo en pruebas de acceso a la universidad. El propio autor califica las opiniones del colectivo como “conservadoras”. No obstante, la muestra de profesores era insuficiente para la extracción de conclusiones, ya que estaba constituida por sólo 5 sujetos.

En nuestro estudio consideramos prioritaria la opinión de los alumnos y decidimos enfocar en ellos la investigación, puesto que el número de profesores de Inglés que han colaborado constituye una muestra de nuevo insuficiente para extraer conclusiones fiables. Futuros planteamientos podrían incluir la exploración de la opinión del profesorado sobre el C-test.

El cuestionario de Jafarpur constaba de 10 preguntas, a las que el sujeto debía contestar de forma afirmativa o negativa. Admitía la posibilidad de explicar la respuesta en algunos casos. A continuación mostramos las preguntas que planteaba el cuestionario para el alumno (Jafarpur 1995):

QUESTIONNAIRE ON C-TESTING

1. What do you think of this as a test of English?
2. Do you think it is a good test?
3. Do you think this test measures English proficiency only?
4. If not, what else does it measure?
5. Do you think it is a fair test of English?
6. Why so, or why not?
7. What do you think of the representativeness of this test?
8. What do you think of the completeness of this test?
9. Would you want your acceptance at university to depend on this test?
10. Why?

Las ventajas del uso de cuestionarios en términos de economía de esfuerzo, tiempo y dinero pueden verse gravemente menguadas si su diseño no es el apropiado (Dörnyei 2003). De ahí nuestro interés por lograr un cuestionario válido y fiable.

Puesto que el contenido de las preguntas del de Jafarpur nos parecía adecuado (opinión general acerca de algunos rasgos de las pruebas como su representatividad, validez, fiabilidad, etc.), decidimos mantenerlas básicamente en el nuestro. Pero cambiamos el modelo de respuesta, que resultaba muy limitado, por otro que incluyera la posibilidad de gradación y facilitara el tratamiento estadístico de los datos.

Se elaboró una primera versión del cuestionario muy similar al de Jafarpur (1995) y también en lengua inglesa. Finalmente optamos por traducirla al español para evitar los sesgos debidos a posibles problemas de comprensión. Además, incluimos algunos cambios, ya comentados en el capítulo 8, apartado 8.3.5:

- En primer lugar se añadió una parte previa en la que solicitábamos información personal (*biodata*). En concreto, la edad y formación en la lengua objeto de estudio. Se respetó el anonimato para garantizar la plena libertad del sujeto al expresar sus opiniones.
- La segunda parte del cuestionario se centra en la valoración personal del sujeto. Plantea una pregunta abierta y el resto pide al sujeto una estimación en escala del 1 al 5 (escala de Likert) acerca de las dificultades surgidas en su realización y sobre su percepción de los rasgos del C-test como instrumento de evaluación de la lengua (qué mide, si es adecuado, completo, si reflejará bien sus conocimientos, etc.).
- Una tercera parte pide opinión sobre la posible utilización del C-test en pruebas selectivas (PAAU), como alternativa o complemento a otras pruebas.

  Concluye agradeciendo la colaboración de los alumnos.

Así pues, el cuestionario final, que puede consultarse en el Apéndice, quedó configurado de tal modo que podemos identificar en él tres partes bien diferenciadas:

1. Información personal
2. Valoración de la prueba en sí misma:
    2.1. Dificultades encontradas
    2.2. Qué mide el C-test
    2.3. Rasgos que lo definen
3. Valoración de la prueba con respecto a la Selectividad

Si el C-test como prueba de evaluación carece de validez aparente para los sujetos, esto provocará su rechazo hacia ella y se traducirá en una pobre valoración de la prueba.

Optamos por aplicar el cuestionario exclusivamente a los alumnos participantes en la investigación, puesto que, en nuestro caso, de nuevo el escaso número de profesores no nos permitiría extrapolar resultados. La valoración del C-test por parte del profesorado de Inglés supondría el planteamiento de otro estudio diferente aunque complementario al de esta tesis.

Para garantizar la total libertad de los sujetos al emitir sus juicios y opiniones se decidió respetar el anonimato (Dörnyei 2003).

12.4. Valoración global de las dificultades planteadas por el C-test

La simple lectura de los cuestionarios antes de someterlos al tratamiento estadístico resulta interesante e ilustrativa. Evidencia que, al completarlos, los sujetos hicieron un importante ejercicio de reflexión acerca de su propio aprendizaje. Este hecho nos produce una satisfacción inicial, porque evidencia que, en general, los cuestionarios no fueron completados “a la ligera”.

La primera pregunta del cuestionario es la única de tipo abierto, conscientes de la dificultad de este tipo de preguntas para ser codificadas de manera fiable, pero a la vez de la riqueza cualitativa que aportan a la investigación (Dörnyei 2003: 47).

Esta pregunta abierta, aunque guiada, plantea si se han encontrado dificultades en la realización del C-test. Se refiere a problemas de tipo general, pero de la reacción de los alumnos a esta pregunta deducimos fallos en su comprensión y/o en su planteamiento, ya que las respuestas de algunos sujetos aluden a sus problemas

concretos para la resolución correcta de la prueba. La mayoría, por ejemplo, manifiesta que tuvo dificultades (aunque quizá deberían decir dudas o desconocimiento de algunas omisiones) e indica que el vocabulario fue el mayor problema. Algunos especifican más, y señalan la ortografía como motivo de error (nº 85[74]: "sé a veces qué palabra es, pero no sé como se escribe").

Merece la pena detenernos en el análisis cualitativo, ya que pese a los posibles problemas de comprensión, en las respuestas a la pregunta abierta "*¿Has encontrado dificultades para realizarlo (el C-test)? ¿de qué tipo?*" encontramos algunas claves del diseño del C-test:

- Los alumnos consideran que el *contexto* es fundamental (nº 91, 94, 141) y constatan sus limitaciones por desconocimiento del vocabulario "no sólo del que hay que rellenar" (nº 90) del texto correspondiente.
- Señalan la frecuencia en las omisiones, que tachan de excesiva, como motivo de error: "es muy difícil comprender el texto si te quitan tantas palabras" (nº 84 y comentarios semejantes en 1, 21, 26).
- Creen que para cada omisión hay varias posibilidades de respuesta correcta, probablemente porque en las pruebas de cierre tradicionales a veces ocurre así, y les lleva a confusión: "hay palabras que las confundes con otras que empiezan igual" (nº 43, 53, 59, 86, 103).
- Son conscientes de que la segunda parte de la prueba (omisiones 51 a 100) supone una dificultad añadida y comentan: "las palabras que no tenían guiones eran mucho más difíciles" (comentario muy repetido en los cuestionarios).
- Aluden al tema del texto como fuente de dificultad: "Dependiendo de cada texto he tenido mayor o menor dificultad, ya que hay temas de los textos que eran más fáciles" (nº 10).
- Acerca de la confusión, que tanto se ha achacado al C-test en la literatura, también encontramos algún comentario. Un sujeto menciona el formato de la prueba: "La forma de ponerlo es muy confusa" (nº 76) y otro la novedad del examen: "estaba perdido, quizás no esté acostumbrado" (nº 8).

[74] Puesto que los cuestionarios son anónimos les fue asignado un número de forma aleatoria (del 1 al 162) para su identificación en el tratamiento estadístico.

- Algunos alumnos mencionan memoria e *imaginación* como ingredientes necesarios para la resolución de la prueba (nº 52 y 65). Cuando se refieren a la memoria, entendemos que aluden de nuevo al vocabulario. Por otra parte, llama nuestra atención que se mencione la imaginación, pero pensamos que esta percepción está en sintonía con la alta correlación entre el C-test y las pruebas subjetivas.
- Por último, un alumno puntualiza: "no es un examen muy complejo si tienes claras las estructuras de la lengua inglesa" (nº 38), comentario que hace pensar que el sujeto reconoce la prueba como buen instrumento de evaluación de la competencia global en la lengua.

A pesar de las dificultades que mencionan los alumnos, como veremos en el análisis de los porcentajes correspondientes a cada pregunta del cuestionario, la valoración de la prueba es positiva.

12.5. Análisis estadístico

El estudio del cuestionario para determinar la validez aparente del C-test se realizó a partir de dos procedimientos estadísticos:

- la confección de tablas de frecuencias
- el procedimiento de análisis factorial.

12.5.1. Tablas de frecuencias

El análisis de las tablas de frecuencias para cada pregunta planteada en el cuestionario permite la valoración de los porcentajes obtenidos en ellas.

De la primera parte del cuestionario (información personal) sólo mostramos los estadísticos correspondientes a los conocimientos previos del sujeto, puesto que la edad es una variable muy homogénea en los sujetos de la muestra.

En la Tabla 12.1 observamos que sólo 12 alumnos de los 162 totales (7,4%) ha disfrutado de alguna estancia lingüística en países de lengua inglesa. Por el

contrario, el 45,1% limita su conocimiento de la lengua a lo aprendido en la enseñanza reglada (colegio de Educación Primaria e IES) y el 47% la ha completado con formación extra en academias, escuelas, clases particulares, etc.

Tabla 12.1. Frecuencias: Conocimientos previos

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Enseñanza reglada	73	45,1	45,1	45,1
	Extra academias	77	47,5	47,5	92,6
	Estancias en el extranjero	12	7,4	7,4	100,0
	Total	162	100,0	100,0	

Si contemplamos los datos desglosados atendiendo a los IES de procedencia de los sujetos (Tablas 12.2) obtenemos una información complementaria muy útil para valorar las diferencias en los resultados de las pruebas cuando se aplica la variable IES, como queda reflejado en el capítulo 11 de la tesis.

Tabla 12.2a. Conocimientos previos: IES Ágora (Alcobendas)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Enseñanza reglada	21	46,7	46,7	46,7
	Extra. Academias	20	44,4	44,4	91,1
	Estancias en el extranjero	4	8,9	8,9	100,0
	Total	45	100,0	100,0	

Tabla 12.2b. Conocimientos previos: IES Vicente Aleixandre (Pinto)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Enseñanza reglada	17	41,5	41,5	41,5
	Extra. Academias	22	53,7	53,7	95,1
	Estancias en el extranjero	2	4,9	4,9	100,0
	Total	41	100,0	100,0	

Tabla 12.2c. Conocimientos previos por IES: San Isidoro (Madrid)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Enseñanza reglada	14	35,0	35,0	35,0
	Extra. Academias	21	52,5	52,5	87,5
	Estancias en el extranjero	5	12,5	12,5	100,0
	Total	40	100,0	100,0	

Tabla 12.2d. Conocimientos previos: IES Humanejos (Parla)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Enseñanza reglada	21	58,3	58,3	58,3
	Extra. Academias	14	38,9	38,9	97,2
	Estancias en el extranjero	1	2,8	2,8	100,0
	Total	36	100,0	100,0	

En las Tablas 12.2 observamos que los porcentajes de alumnos que han completado su formación en Inglés fuera de los centros escolares oscilan entre el 38 % de Parla y el 53 % de Pinto. En cuanto a las estancias en el extranjero, de nuevo el menor porcentaje está en los alumnos del IES Humanejos (2,8 %), frente al del IES San Isidoro (12,5 %), en el otro extremo, y muy por debajo del promedio de la muestra (7,4%). Estos datos se explican por las diferencias socioeconómicas de las poblaciones en que están situados los IES que forman parte del estudio.

Centraremos nuestro análisis de frecuencias en la parte central del cuestionario, en la que, mediante la escala de Likert, exploramos la valoración que los sujetos hacen de la prueba y sus características como instrumento de evaluación. Casi la totalidad de los alumnos (95%) manifestó que había encontrado dificultades para resolver el C-test. Como hemos mencionado en el apartado 12.3, la pregunta pretendía aludir a cuestiones de tipo general, de comprensión de la prueba, pero de las respuestas de los sujetos deducimos que se interpretó de manera diferente.

Por tanto, aunque prácticamente todos manifiestan que la prueba les supuso dificultades (Tabla 12.3), entendemos que no más que cualquier otro examen o prueba.

Tabla 12.3. Frecuencias: Dificultades con el C-test

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	155	95,7	95,7	95,7
	No	5	3,1	3,1	98,8
	valores perdidos	2	1,2	1,2	100,0
	Total	162	100,0	100,0	

En la tercera pregunta del cuestionario el sujeto debía valorar en una escala de Likert del 1 al 5 qué mide el C-test. A continuación veremos los promedios, las tablas de frecuencia y los diagramas de barras de los distintos aspectos a que aludía el cuestionario: gramática, ortografía, conocimiento general de la lengua, fluidez, y léxico.

Atendiendo a los promedios, presentados en la Tabla 12.4, el léxico y la ortografía (*spelling*) son las variables que obtienen mayor puntuación al preguntar a los sujetos qué mide el C-test, no obstante también el conocimiento general en la lengua, la fluidez y la gramática logran puntuaciones cercanas a los 3 puntos.

Por tanto, se identifica al C-test como prueba en que la variable “vocabulario” tiene un peso específico (véanse las frecuencias en las Tablas 12.5 y diagramas de barras, Fig. 12.1). Un 55 % de los sujetos asigna el valor máximo de la escala de valoración al léxico y un 37 % a la ortografía, éstos son también los dos aspectos que muestran la distribución de frecuencias más irregular.

Tabla 12.4. Promedios. Estadísticos descriptivos

	Media	Desviación típica	N del análisis
Gramática	2,96	1,074	162
Ortografía	3,94	1,061	162
General	3,36	1,193	162
Fluidez	3,32	1,135	162
Léxico	4,34	1,023	162
Adecuado	2,96	1,318	162
Completo	2,81	1,127	162
Válido	2,56	1,216	162

Las Tablas 12.5 reflejan todos estos datos. En la 12.5a vemos que sólo 11 sujetos consideran que el C-test mida primordialmente conocimientos gramaticales (5 en la escala de Likert). Las puntuaciones se concentran en los valores centrales 2, 3 y 4 (que suman un 83,3 % del total, correspondiendo al 3 el 35,2 %).

Tabla 12.5a. Frecuencias: Gramática

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	16	9,9	9,9	9,9
	2	37	22,8	22,8	32,7
	3	57	35,2	35,2	67,9
	4	41	25,3	25,3	93,2
	máximo (escala de Likert)	11	6,8	6,8	100,0
	Total	162	100,0	100,0	

Tabla 12.5b. Frecuencias: Ortografía

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	6	3,7	3,7	3,7
	2	8	4,9	4,9	8,6
	3	36	22,2	22,2	30,9
	4	52	32,1	32,1	63,0
	Máximo (escala de Likert)	60	37,0	37,0	100,0
	Total	162	100,0	100,0	

Tabla 12.5c. Frecuencias: Conocimiento general de la lengua

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	6	3,7	3,7	3,7
	2	35	21,6	21,6	25,3
	3	49	30,2	30,2	55,6
	4	43	26,5	26,5	82,1
	máximo (escala de Likert)	28	17,3	17,3	99,4
	valores perdidos	1	,6	,6	100,0
	Total	162	100,0	100,0	

Tabla 12.5d. Frecuencias: Fluidez

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	9	5,6	5,6	5,6
	2	26	16,0	16,0	21,6
	3	54	33,3	33,3	54,9
	4	54	33,3	33,3	88,3
	máximo (escala de Likert)	18	11,1	11,1	99,4
	valores perdidos	1	,6	,6	100,0
	Total	162	100,0	100,0	

En la Tabla siguiente (12.5e) veremos que 90 sujetos consideran que el C-test mide sobre todo los conocimientos de léxico y asignan al vocabulario el valor máximo de la escala. El 82,8 % del total de los alumnos se aglutina en los valores 4 y 5.

Tabla 12.5e. Frecuencias: Léxico

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	3	1,9	1,9	1,9
	2	7	4,3	4,3	6,2
	3	17	10,5	10,5	16,7
	4	44	27,2	27,2	43,8
	máximo (escala de Likert)	90	55,6	55,6	99,4
	valores perdidos	1	,6	,6	100,0
	Total	162	100,0	100,0	

A continuación, las Figuras 12.1 muestran de forma gráfica, en diagramas de barras, los datos anteriores.

Nos fijaremos especialmente en la distribución de frecuencias reflejada en los diagramas correspondientes al léxico y la ortografía, que asciende a medida que lo hace la escala de Likert. Sin embargo, los demás reflejan la tendencia central previamente comentada.

Fig. 12.1a. Diagrama de barras: Gramática

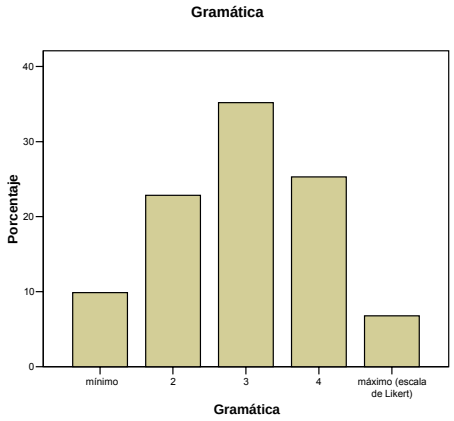


Fig. 12.1b. Diagrama de barras: Ortografía

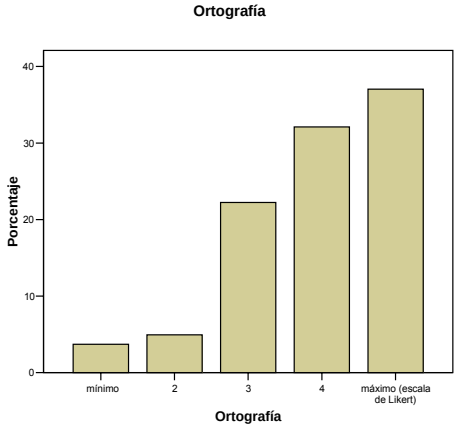


Fig. 12.1c. Diagrama de barras: C.General

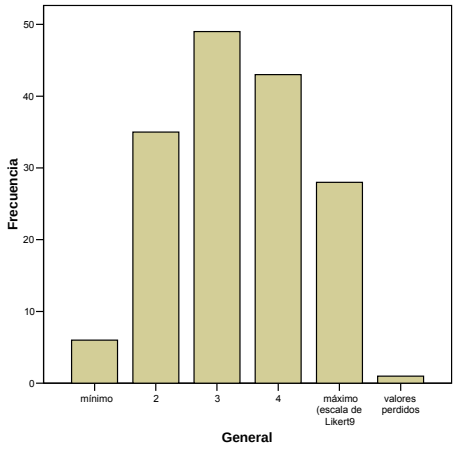


Fig. 12.1d. Diagrama de barras: Fluidez

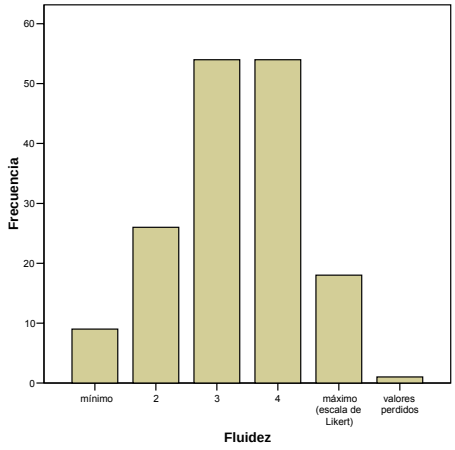
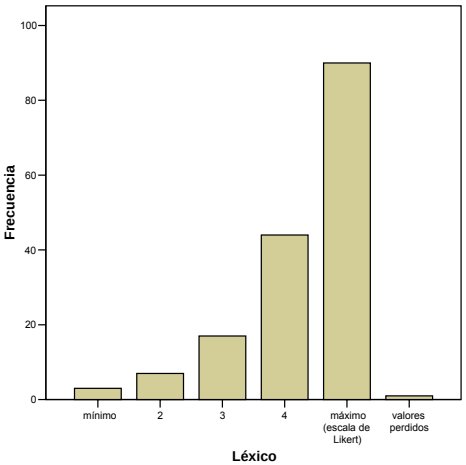


Fig. 12.1e. Diagrama de barras: Léxico



Seguiremos con el análisis de la valoración que hacen los sujetos en cuanto a las características que podrían definir al C-test: adecuado, completo y válido como instrumento de evaluación.

A partir de los porcentajes obtenidos (Tablas 12.6), podemos decir que los alumnos consideran al C-test como prueba apropiada (el 45% otorga un 3 en la escala a este rasgo), más que completa (el 32 %) y válida (34,6).

Atendiendo a los promedios obtenidos por estos tres rasgos (Tabla 12.4), el C-test consigue una valoración media bastante alta, “aprueba”, a juicio del alumnado que se enfrenta a ella por primera vez (el rasgo “adecuado” consigue el mayor promedio, un 2,96 en escala de 1 a 5).

La Tabla 12.6a, correspondiente al rasgo “adecuado”, indica que el 69 % de las puntuaciones está en los valores más altos de la escala de Likert (3, 4 y 5).

Tabla 12.6a. Rasgos del C-test: Adecuado

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	20	12,3	12,3	12,3
	2	29	17,9	17,9	30,2
	3	74	45,7	45,7	75,9
	4	27	16,7	16,7	92,6
	Máximo (escala de Likert)	9	5,6	5,6	98,1
	Valores perdidos	3	1,9	1,9	100,0
	Total	162	100,0	100,0	

Tabla 12.6b. Rasgos del C-test: Completo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	14	8,6	8,6	8,6
	2	55	34,0	34,0	42,6
	3	53	32,7	32,7	75,3
	4	31	19,1	19,1	94,4
	Máximo (escala de Likert)	8	4,9	4,9	99,4
	Valores perdidos	1	,6	,6	100,0
	Total	162	100,0	100,0	

En la Tabla 12.6c llama la atención el alto porcentaje de sujetos que asignan la puntuación mínima a la validez como rasgo de la prueba, probablemente por falta de comprensión o interpretación errónea de la terminología.

Tabla 12.6c. Rasgos del C-test: Válido

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mínimo	37	22,8	22,8	22,8
	2	39	24,1	24,1	46,9
	3	56	34,6	34,6	81,5
	4	23	14,2	14,2	95,7
	Máximo (escala de Likert)	6	3,7	3,7	99,4
	Valores perdidos	1	,6	,6	100,0
	Total	162	100,0	100,0	

En la parte final del cuestionario planteamos la posibilidad de que el C-test fuera una alternativa a la prueba actual de Inglés de Selectividad (Tabla 12.7). Los sujetos manifestaron mayoritariamente su negativa (76,5 %). Esta reacción era previsible y podría deberse simplemente al miedo de enfrentarse a una prueba imprevista a sólo dos meses de las PAAU oficiales y después de una preparación enfocada a otro tipo de prueba. Cuando se propone el C-test sólo como complemento que podría completar el diseño de la actual prueba los porcentajes se equilibran bastante (Tabla 12.8), aunque sigue existiendo cierto recelo inevitable y resistencia al cambio, que, en nuestra opinión, puede enmascarar o sesgar la opinión real del sujeto.

Tabla 12.7. El C-test como alternativa a la prueba de Inglés de Selectividad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	37	22,8	22,8	22,8
	No	124	76,5	76,5	99,4
	valores perdidos	1	,6	,6	100,0
	Total	162	100,0	100,0	

Tabla 12.8. El C-test como complemento de la prueba de Inglés de Selectividad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	80	49,4	49,4	49,4
	No	81	50,0	50,0	99,4
	valores perdidos	1	,6	,6	100,0
	Total	162	100,0	100,0	

12.5.2. Análisis factorial

El análisis factorial es una técnica de reducción de datos que permite encontrar grupos homogéneos de variables a partir de un conjunto de ellas que se consideran independientes. De este modo se pueden explicar mejor las respuestas de los sujetos.

Aplicamos este procedimiento de análisis al cuestionario porque queremos comprobar si es posible resumir toda la información disponible sobre la opinión de los sujetos acerca del C-test mediante un número reducido de factores.

Para comenzar volvemos a los promedios obtenidos por cada uno de los factores estudiados (Tabla 12.4) en escala del 1 al 5. Esta información nos da una primera aproximación de conjunto. Como hemos visto, destacan el léxico (4,34) y la ortografía (3,94) como factores más valorados, es decir que los sujetos piensan que el C-test mide, sobre todo, aspectos relacionados con el vocabulario. Le siguen otros aspectos como el conocimiento general de la lengua (3,36) y la fluidez (3,32), relativos al dominio de la lengua.

Continuamos analizando el KMO. El estadístico KMO (Kaiser-Meyer-Olkin) varía entre 0 y 1. La medida de adecuación muestral (0,7) indica que sí es adecuado realizar el análisis factorial (Tabla 12.10). Si se hubiera obtenido un valor menor que 0,5 no habría sido pertinente continuar con el procedimiento. Un valor de significación mayor que 0,05 en la prueba de esfericidad de Bartlett tampoco lo habría recomendado. Por tanto, a partir de estos datos realizaremos un primer análisis factorial de dos componentes.

Tabla 12.10. Estadístico KMO y prueba de Bartlett.

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,717
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	291,901
	Gl	28
	Sig.	,000

La comunalidad de una variable es la proporción de su varianza que puede ser explicada por el modelo factorial obtenido. La Tabla 12.11 contiene las comunalidades asignadas inicialmente a las variables y las reproducidas por la solución factorial (extracción). En nuestro caso la ortografía es la variable peor explicada (30,6 % de la varianza).

Tabla 12.11. Comunalidades

	Inicial	Extracción
Gramática	1,000	,440
Ortografía	1,000	,306
General	1,000	,462
Fluidez	1,000	,555
Léxico	1,000	,672
Adecuado	1,000	,648
Completo	1,000	,732
Válido	1,000	,484

Método de extracción: Análisis de Componentes principales.

En la Tabla 12.12, los autovalores iniciales expresan la cantidad de varianza total explicada por cada factor. Vemos que con dos factores se explica el 53,7% de la varianza.

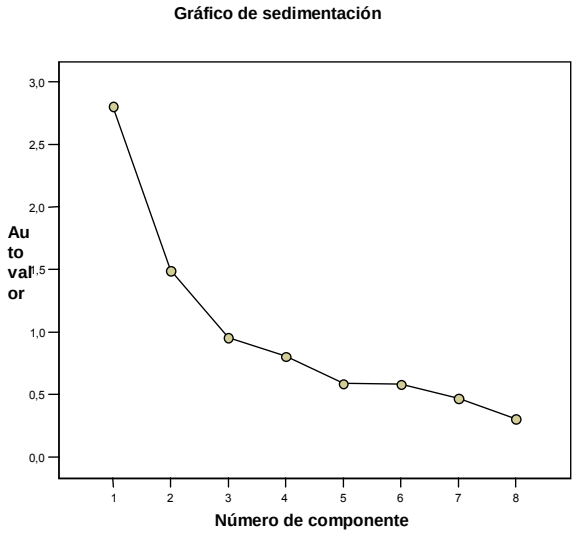
El gráfico de sedimentación (Figura 12.2), en consonancia con la Tabla anteriormente citada, también muestra que destacan dos componentes, y a partir del tercero se observa que apenas contribuyen a explicar la varianza. Por tanto no es necesario continuar realizando el análisis factorial con tres componentes.

Tabla 12.12. Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Componente	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	**2,808**	**35,104**	**35,104**	**2,808**	**35,104**	**35,104**
2	**1,491**	**18,642**	**53,746**	**1,491**	**18,642**	**53,746**
3	,955	11,935	65,681			
4	,805	10,064	75,744			
5	,585	7,318	83,062			
6	,580	7,249	90,311			
7	,470	5,877	96,188			
8	,305	3,812	100,000			

Método de extracción: Análisis de Componentes principales.

Figura 12.2. Gráfico de sedimentación de Cattell



Por último, la matriz de componentes (Tabla 12.13) contiene las correlaciones entre las variables originales y cada uno de los factores.

Se puede apreciar que el primer factor o componente está constituido por las variables *ortografía* y *léxico*, y refleja la dimensión de "vocabulario" en la prueba.

El segundo factor está formado por la *fluidez* y el *conocimiento general de la lengua* y reflejan el "dominio" de la lengua. Estos resultados coinciden con lo ya observado en el análisis de las tablas de frecuencias del apartado 12.5.1.

Tabla 12.13. Matriz de componentes (a)

	Componente	
	1	2
Gramática	,637	-,185
Ortografía	**,173**	,525
General	,629	**,257**
Fluidez	,556	**,496**
Léxico	**,177**	,800
Adecuado	,783	-,186
Completo	,825	-,225
Válido	,584	-,379

Método de extracción: Análisis de componentes principales.
a 2 componentes extraídos

12.6. Conclusiones

A partir de los datos obtenidos mediante el cuestionario retrospectivo de opinión y después del estudio de las tablas de frecuencias y del análisis factorial, comprobamos que los sujetos identifican al C-test como prueba que mide principalmente dos factores: vocabulario y destrezas generales.

Del análisis, más cualitativo que cuantitativo, de las respuestas a la primera pregunta del cuestionario, de tipo abierto, podemos también extraer algunas conclusiones. Los alumnos fueron capaces de captar algunos aspectos clave del C-test, tales como:

- La importancia del contexto en la recuperación de las omisiones.
- La influencia del texto de partida (tema, características, etc.) en el grado de dificultad del C-test.
- La complicación que supone el paso de omisiones guiadas a no guiadas en la segunda parte de la prueba.
- La importancia del vocabulario en la prueba.
- La relación del C-test con otras pruebas de cierre en que la frecuencia de las omisiones es menor.
- La falta de familiarización con este tipo de prueba.

Es evidente que los sujetos reconocen los rasgos que aportan dificultad a la prueba, incluidos los derivados de su aspecto (frecuencia de las omisiones, falta de familiarización, etc.), pero la admiten y reconocen su validez como instrumento de medida. Son muy pocos los que aluden al aspecto externo del C-test como fuente de confusión, si bien algunos más señalan su novedad como problema añadido.

Quizá sobrevaloran el papel del conocimiento del vocabulario para resolver correctamente un C-test e infravaloran el de los conocimientos generales de la lengua. Lo que indica que, por su aspecto, el C-test puede parecer una prueba que mide exclusivamente el vocabulario.

Al contrario que Weir (1988), Jafarpur (1995) y Bradshaw (1990), entre otros, en este estudio hemos verificado que no hay rechazo hacia el C-test por su formato. El grado de satisfacción expresado por los sujetos supera incluso las expectativas de los propios creadores de la prueba. A pesar de su novedad, la prueba es, en general, bien valorada por los alumnos (según indican los promedios) como reflejo de su conocimiento de la lengua inglesa, especialmente en lo relativo al léxico. No obstante, se aprecia resistencia al cambio, que se expresa en la respuesta negativa a las preguntas que sugieren la inclusión de un C-test como alternativa a la actual prueba de Selectividad o como parte de ella, si bien hemos de atender a los posibles sesgos que afectan a estas cuestiones.

Por tanto, el cuestionario retrospectivo ha puesto de manifiesto la validez aparente del C-test y, en consecuencia, se cuestiona y rechaza la hipótesis 5:

> "Por su novedad y su carácter fragmentario, algo confuso al principio, puede conducir al rechazo. El C-test carece de validez aparente."

Por otra parte, en este estudio se ha constatado que el hecho de que los cuestionarios sean anónimos, por una parte facilita la libre expresión de la opinión de los sujetos, pero por otra, limita el análisis, ya que impide hacer inferencias a partir de las correlaciones entre los resultados obtenidos en el C-test y la valoración que el sujeto hace de la prueba. En investigaciones posteriores este punto ha de ser tenido en cuenta.

CONCLUSIONES Y SÍNTESIS DE RESULTADOS

Introducción

Este trabajo se inscribe en el marco general de la Evaluación de la Lengua, y más concretamente en el de la Evaluación del Inglés como Lengua Extranjera. En él se ha llevado a cabo el análisis del funcionamiento del C-test, subtipo de las pruebas de cierre. Se ha constatado que esta prueba de redundancia reducida presenta ventajas, sobre todo de tipo práctico, frente a los *clozes* tradicionales. Por ello, se está consolidando como instrumento de medida de la competencia general en Inglés como Lengua Extranjera.

Se ha prestado especial atención a algunas de sus características; como la validez, la fiabilidad y la factibilidad. Partiendo de las investigaciones que demuestran la validez de constructo (Eckes y Grotjahn 2006; Klein-Braley y Raatz 1981, 1984; Spolsky 1973; Oller 1979) y contenido (Hughes 1989; Bachman *et al*. 1996) del C-test, en las que entronca directamente nuestro estudio, nos hemos centrado en su validez criterial y aparente.

Nuestro trabajo empírico comienza con el diseño y aplicación de un C-test (formado por cuatro subtests: C-test 1, C-test 2, C-test 3 y C-test 4) a alumnos de 2º curso de Bachillerato.

El estudio de la validez del C-test ha abordado varios aspectos. En primer lugar se ha analizado el funcionamiento del C-test partiendo de sus características intrínsecas: modelos y subtests. Se ha intentado acotar las variables que determinan la dificultad o facilidad de los ítems: factores textuales, de formato, tipo de término de las omisiones, etc.

En segundo lugar, para probar su validez criterial concurrente como instrumento de medida hemos tomado como referencia principal otra prueba, la de Inglés de las PAAU, ya instaurada y aplicada a gran escala en España. Hemos trabajado con una prueba modelo PAAU realizada en el aula por todos los sujetos de la muestra y con los resultados de la PAAU de Inglés de junio de 2001. Se han revisado las correlaciones entre el C-test, estas pruebas y la valoración que hacen los profesores de Inglés de la evolución y competencia de los sujetos en la 2ª Evaluación del curso escolar.

Mediante el procedimiento de regresión lineal se han explorado las relaciones entre las distintas partes o subtests que forman el C-test (C-test 1, C-test 2, C-test 3 y C-test 4) y las otras pruebas aplicadas (VDs): *Cavemen?*, la Selectividad de junio de 2001 y las calificaciones en la 2ª Evaluación.

Se ha analizado la validez aparente del C-test tomando como punto de partida el cuestionario retrospectivo de opinión aplicado a los sujetos de la muestra, a partir de dos procedimientos estadísticos: la confección de tablas de frecuencias y el análisis factorial.

La fiabilidad del C-test ha sido estudiada principalmente mediante el método de "análisis por mitades" y calculando el Alfa de Cronbach.

Nuestro trabajo ha corroborado la factibilidad de la prueba, reconocida ampliamente en la literatura (Süssmilch 1984; Döryei y Katona 1992; Klein-Braley 1997; Connelly 1997; Babaii y Ansary 2001) y su aplicabilidad, probablemente incluso en formato electrónico.

Finalmente, atendiendo a las características de la muestra, detectamos la existencia de dos factores demográficos que pueden suponer diferencias en la actuación de los sujetos en el C-test: el género y el IES de procedencia. Abordamos el género con el análisis de los promedios desglosados por géneros, el ANOVA y el modelo lineal general. En cuanto a los IES, de muy diferente ubicación dentro de la Comunidad de Madrid, se revisaron las características del entorno, los promedios y se realizó un análisis de varianza univariante.

A lo largo de nuestro trabajo, siguiendo el orden de presentación de los datos, se han confirmado las hipótesis 1, 3, 4 y 6, mientras que las hipótesis 2, 5, y 7 inicialmente planteadas, han sido rechazadas.

Conclusiones

Las conclusiones se han organizado en 7 apartados, de acuerdo con las líneas de investigación seguidas en nuestro trabajo.

Empezaremos por exponer, en el apartado A, los resultados relativos a la validez de la prueba atendiendo a distintos aspectos. En primer lugar, en el apartado A.1 se analizan los promedios obtenidos en el C-test teniendo en cuenta sus características: estructura en subtests y formato de las omisiones. El apartado A.2 aborda la incidencia de las variables textuales en el grado de dificultad de la prueba. Seguidamente, en el A.3 se analiza la validez criterial concurrente del C-test a partir de sus correlaciones con la prueba de Inglés de las PAAU (oficial y en el aula) y con las calificaciones de Inglés en la 2ª Evaluación. En este apartado abordaremos también las diferencias en las correlaciones del C-test con distintos tipos de pregunta, en definitiva, con la parte objetiva y subjetiva de la prueba de Selectividad. El apartado A.4 mostrará las relaciones entre los subtests y el resto de las pruebas aplicadas (VDs) mediante el análisis de regresión lineal. Por último, en el apartado A.5 exploraremos el cuestionario retrospectivo, que aporta datos empíricos para valorar la validez aparente de la prueba.

La fiabilidad del C-test se estudia en el apartado B mediante el método de "análisis por mitades", el Alfa de Cronbach y el análisis de las correlaciones con las otras pruebas.

A continuación, en el apartado C examinaremos la incidencia de dos factores externos a la prueba: el género de los sujetos y la ubicación del IES.

Más adelante, en el D se analizarán las implicaciones pedagógicas de todo el trabajo y en el E se incluyen algunos consejos prácticos para la creación de C-tests.

Finalmente, en el apartado F se presenta una síntesis de los resultados más relevantes del estudio, para concluir con la propuesta de posibles futuras líneas de investigación en el apartado G.

A. Validez del C-test

En este epígrafe se exponen los resultados de nuestras investigaciones en torno a la validez de la prueba, estructuradas en varios apartados, desde la validez criterial concurrente hasta la aparente. Se tendrán en cuenta los factores propios del diseño del C-test (estructura en subtests, factores textuales, etc.) y los promedios obtenidos.

A.1. Características intrínsecas del C-test aplicado y análisis de promedios

Comenzaremos con el análisis de las características o aspectos intrínsecos de la prueba diseñada y aplicada en esta investigación que inciden en la actuación de los sujetos y en los resultados obtenidos en ella.

El diseño de la prueba (100 ítems) se estructura en torno a cuatro textos a partir de los cuales se crean cuatro subtests de 25 omisiones cada uno. De este modo conseguimos un buen número de ítems, lo que asegura mayor validez y fiabilidad en las pruebas de cierre, según Farhady y Keramati (1996).

En los dos últimos subtests se introduce un cambio de formato al retirar la ayuda de las omisiones. Se diseñaron dos modelos, A y B, alternando el orden de los textos.

Los resultados obtenidos en los promedios del C-test (5,112 puntos en escala de 0 a 10) y los subtests son los adecuados para una prueba de tipo normativo con un grado de dificultad medio, y ponen de manifiesto el poder discriminatorio del C-test (Klein-Braley 1984). Los promedios de los subtests varían en función de su grado de dificultad, motivado por distintos factores, pero los histogramas reflejan siempre una distribución normal. Cuando se trata de omisiones no guiadas (subtests 3 y 4) los promedios descienden sensiblemente y la dispersión de puntuaciones es mayor, a pesar de la familiarización con la técnica, lo que indica que el formato de omisiones no guiadas aumenta la dificultad de la prueba.

El análisis de las correlaciones entre el C-test y los subtests muestra que todos los subtests correlacionan de forma significativa con el total del C-test (con valores entre 0,727 y 0,877). Las correlaciones entre los subtests no presentan valores tan

altos, concretamente el C-test 3 presenta una correlación muy baja con el C-test 1, lo que hace pensar de nuevo en el formato como principal factor de dificultad, aunque no el único.

El estudio de los promedios en los subtests de los dos modelos de C-test (A y B) permitió constatar que un mismo texto puede presentar una diferencia de aproximadamente 3 ó 4 puntos en los promedios (sobre un total de 25) dependiendo del tipo de omisión (con o sin pistas). La ayuda que proporciona el número de letras de cada omisión facilita la inferencia. De este modo, se confirma la hipótesis 4; los cambios en el formato influyen directamente en los resultados obtenidos, cuando se incluye el número de letras que corresponde a cada omisión se facilita la tarea del alumno.

La incidencia del formato en la recuperación de los ítems se muestra en los estadísticos de frecuencias de la recuperación de términos omitidos. Destaca el aumento de valores perdidos cuando no se aportan pistas, síntoma de lo que hemos denominado efecto psicológico de “desánimo” o desmotivación en los sujetos, que ni siquiera intentan resolverlos. Este tipo de acercamiento proporciona también algunas pautas de recuperación de los términos funcionales y léxicos.

El análisis de promedios revela que las mayores dificultades para recuperar el texto inicial surgieron en los subtests creados a partir del texto *American imperialism* independientemente del formato aplicado, es decir, tanto con omisiones guiadas (10,37) como no guiadas (7,47). Además, los histogramas de *American imperialism* muestran cómo cambian la distribución de puntuaciones y la desviación estándar con formato guiado y no guiado. Mientras que con omisiones guiadas la distribución es normal, con no guiadas se aprecia un sesgo positivo que viene dado porque las frecuencias más altas corresponden a los valores más bajos de la tabla.

En cuanto a la desviación estándar, es menor en el caso de omisiones guiadas. Para explicarlo, en los apartados siguientes dirigiremos nuestra atención a las características del texto sobre el que se diseñaron los subtests.

A.2. Incidencia de factores textuales en el grado de dificultad de la prueba

En distintos momentos hemos comentado en esta tesis la importancia de los textos a partir de los cuales se crea el C-test. El análisis de los promedios revela la existencia de factores textuales que inciden directamente en la dificultad de la prueba. Entre ellos, mencionaremos el tema del texto, la variación y densidad léxicas, y el tipo de término afectado por la mutilación. Aunque algunos autores (Dörnyei y Katona 1992) mencionan también incidencia de la longitud de las oraciones, en nuestros C-tests se descartó este factor y, por tanto, no se consideró pertinente su análisis.

Acerca del tema hemos de decir que la familiarización y el interés que suscita en los sujetos puede suponer una ayuda eficaz (Sasaki 2000). Los propios alumnos lo reflejan en el cuestionario retrospectivo. Sin embargo, resulta difícil cuantificar hasta qué punto influyen en la realización del C-test.

En cuanto a la variación y densidad léxicas (Laufer y Nation 1995; Schmitt 2000) hemos visto que a medida que éstas aumentan lo hace también la dificultad del C-test. Según Dörnyei y Katona (1992: 197) y Babaii y Ansary (2001: 217) los textos fáciles son los que mejor funcionan para crear C-tests.

En el apartado anterior hemos señalado que en los subtests creados a partir del texto *American imperialism* se obtienen los promedios más bajos (10,37 y 7,47 en omisiones guiadas y no guiadas, respectivamente). Como veremos a continuación, debido a sus características léxicas derivadas, entre otros factores, del tema del texto, resultó ser el más difícil.

El texto presenta un tema de carácter político-histórico, quizá más lejano y de menor interés para el alumno que los otros. Si atendemos al vocabulario, veremos que en él abundan los términos de carácter léxico largos y abstractos, como *leadership, development, engagement, cooperation, prosperity, peaceful,* etc., algunos de ellos afectados por la mutilación (Laufer 1997). Así pues, la variable temática pudo incidir en la comprensión del texto y en la motivación para intentar recuperarlo correctamente (Sasaki 2000).

Además, *American imperialism* muestra también los valores más altos en variación (70,37) y densidad léxicas (60,18). Y los porcentajes se disparan al tener

en cuenta el tipo de término afectado por la mutilación: el 76% de las omisiones corresponde a términos léxicos y tan sólo el 24% a funcionales.

La literatura muestra que los términos funcionales se recuperan mejor que los léxicos (Klein-Braley 1985, Dörnyei y Katona 1992, Farhady y Keramati 1996) en las pruebas de cierre. En general es así, pero en nuestro estudio comprobamos que es fundamental notar, además, la incidencia de la redundancia del texto, la longitud de la palabra omitida y su frecuencia en la lengua.

A partir del análisis de las varianzas podemos decir que los términos léxicos se recuperan muy bien o muy mal, dependiendo del dominio de la lengua, mientras que en la recuperación de términos funcionales la dispersión de puntuaciones es menor.

Estos resultados nos hacen confirmar la hipótesis 3. En ella planteábamos que la recuperación de los términos funcionales sería más fácil que la de los que tienen contenido léxico. Hemos constatado que, en general, los términos de función se recuperan mejor por ser un número limitado en la lengua, pero también debido a su frecuencia de uso y a su tamaño (generalmente son cortos). Por otra parte, los términos léxicos se recuperan bien si son frecuentes en la lengua o redundantes en el texto. La frecuencia es, por tanto, un factor que pondera de una manera significativa en el índice de facilidad/dificultad de los ítems. Así pues, a pesar de todo, la mera clasificación en términos léxicos y gramaticales nos parece insuficiente para explicar su comportamiento en el C-test.

El análisis de las correlaciones entre la recuperación de términos léxicos y funcionales en ambos modelos de C-test nos permite afirmar que las variables “léxico” y “función” están muy asociadas en la prueba.

Aunque en la literatura encontramos opiniones favorables a la supresión de los términos excesivamente fáciles o difíciles porque no discriminan entre los sujetos (Grotjan 1987; Kamimoto 1993; Jafarpur 1999), consideramos que este tipo de términos debe mantenerse, ya que los muy fáciles pueden resultar motivadores para los sujetos menos expertos y los difíciles un reto para los aventajados. Además, los estudios muestran que su eliminación no cambia los resultados (Jafarpur 1999).

A.3. Validez criterial concurrente del C-test

Antes de abordar los resultados del análisis de las correlaciones entre el C-test y la prueba de Inglés de las PAAU revisaremos algunos aspectos relativos a la estructura de esta última prueba.

La prueba de Inglés de las PAAU se estructura en torno a dos partes de igual valor (5 puntos), una de carácter objetivo y otra subjetiva. La parte objetiva está compuesta por tres preguntas, una de verdadero o falso, otra de vocabulario y una tercera de contenido gramatical. Una pregunta abierta y una redacción integran la subjetiva.

En la investigación se trabajó con los resultados de dos pruebas semejantes; una aplicada en el aula (*Cavemen?*), lo que permitió su desglose, y la PAAU oficial de junio de 2001, a la que sólo se presentó la mitad de los sujetos de la muestra.

Los resultados de este estudio confirman los resultados de Herrera (1999): algunas preguntas objetivas de la prueba no discriminan. Comparando los promedios e histogramas de la parte objetiva y subjetiva de la prueba *Cavemen?* (6,2 y 4,69 respectivamente) observamos que la parte subjetiva resultó más difícil, pero tiene mayor potencia discriminatoria.

Resulta llamativo constatar que el C-test correlaciona mejor con la parte subjetiva. Hemos de rechazar, por tanto, la hipótesis 2 que presuponía que, atendiendo a sus características, por ser una prueba de elementos discretos, el C-test debería correlacionar mejor con pruebas de tipo objetivo. Se ha comprobado que, por su carácter de prueba holística en función del contexto, el C-test correlaciona mejor con pruebas de tipo subjetivo y holístico, como las preguntas abiertas y las redacciones.

Soyoung Lee (1996), en la línea de Hanania y Shikhani (1986) y Fotos (1991), propone las pruebas de cierre como alternativa a los ensayos. Como veremos, los resultados de nuestra investigación permiten proponer al C-test como alternativa a otras pruebas, tanto de tipo objetivo como subjetivo.

El análisis de la validez criterial concurrente del C-test pone de manifiesto que el C-test correlaciona de forma significativa con otras pruebas estandarizadas que miden la competencia global en lengua inglesa. Se obtienen unos valores de

correlación muy semejantes (0,722 entre el C-test y la PAAU oficial de junio de 2001, 0,750 con *Cavemen?* y 0,723 con la 2ª Evaluación). Se confirma así la hipótesis 1, eje fundamental de nuestra investigación: El C-test correlaciona bien tanto con pruebas estandarizadas (PAAU) como con la valoración de los respectivos profesores de Inglés con respecto a la evolución en la asignatura.

Entre las dos pruebas PAAU (*Cavemen?* y la prueba de Inglés de Selectividad de junio de 2001) se aprecia una buena correlación (0,654), aunque es todavía mejor con la 2ª Evaluación (0,805) y con el C-test (0,750). Hay que tener en cuenta la significación y trascendencia de la prueba oficial y las características de su aplicación (trascendencia, ansiedad, etc.).

La correlación entre el C-test y la PAAU oficial es también un indicador de la validez predictiva del C-test (0,722). A la luz de estos resultados podemos proponer el uso de C-tests en la preparación de la prueba de Inglés de Selectividad vigente. Aunque el promedio de la PAAU (6,32) es superior al del C-test (5,75), la diferencia se justifica por las características de la PAAU.

Para completar y corroborar los resultados obtenidos en el análisis de promedios y correlaciones, se analizó la relación entre los subtests del C-test y las otras pruebas de nuestra investigación mediante el procedimiento de regresión lineal.

Con respecto a la 2ª Evaluación, determinamos que el C-test 4 es el que mejor predice los resultados de los sujetos, y que el C-test 1, por el contrario, no contribuye a explicar la varianza. De nuevo, en el caso de *Cavemen?* es el subtest 4 el mejor predictor de la VD y, el subtest 1, el peor. Este patrón se repite parcialmente también en la Selectividad oficial; el C-test 4 sigue siendo el que mejor explica la varianza, aunque en este caso el C-test 2 queda fuera del modelo. Así pues, el C-test 4 es el que presenta mejor funcionamiento como predictor de las tres variables dependientes estudiadas. Se explica por varias razones: en primer lugar, en el subtest 4 ya se ha producido una suficiente familiarización con la técnica a pesar de las omisiones no guiadas, en segundo lugar, este subtest está diseñado a partir de los dos textos en los que, casualmente, el nivel de densidad y variación léxicas es menor: *Women doctors* y *Evolution,* en los modelos A y B respectivamente.

A.4. Análisis de regresión lineal

En esta investigación se ha utilizado el procedimiento de regresión lineal para explorar y cuantificar las relaciones entre los subtests que forman el C-test (C-test 1, C-test 2, C-test 3 y C-test 4) y las otras pruebas aplicadas (*Cavemen?*, PAAU de junio de 2001 y 2ª Evaluación).

La prueba de regresión lineal corrobora el buen funcionamiento del C-test como predictor de los resultados obtenidos en Inglés en la 2ª Evaluación y en las pruebas de Inglés de Selectividad (oficiales o no). Concretamente, el subtest 4 es el que mejor predice los resultados de los sujetos en todos los casos, lo que no implica que se pudiera prescindir de los otros subtests.

Podría resultar llamativo que el subtest que mejor explica la varianza tanto de las calificaciones en la 2ª Evaluación como de las PAAU aplicadas sea uno de los que tiene omisiones no guiadas (C-test 4). Pero estos resultados coinciden con los obtenidos al analizar las correlaciones de los subtests con el total de la prueba: el C-test 4 consigue la mejor correlación con el C-test (0,877), aunque no los promedios más altos.

El diseño global de la prueba explica el buen funcionamiento del subtest 4 como predictor de todas las variables dependientes. Cuando se aplicó el C-test, éste era un formato nuevo para el alumno, y a través de la práctica se produjo un aprendizaje. En el subtest 3 la tarea de recuperar los textos se complica al comenzar las omisiones no guiadas, pero en el subtest 4 se domina la técnica, para quedar únicamente las dificultades derivadas de los propios textos: *Women doctors* (modelo A) y *Evolution* (modelo B). En este caso, son textos que no presentan dificultades de contenido, y su densidad y variación léxicas son las menores del C-test. Su buen funcionamiento corrobora la tesis de Dörnyei y Katona (1992: 197); los textos más fáciles son los más adecuados para la técnica del C-test.

El análisis de regresión lineal aplicado contribuyó a respaldar los resultados obtenidos en los análisis anteriores.

A.5. Validez aparente del C-test: cuestionario retrospectivo

El aspecto externo del C-test y la novedad de la técnica han provocado reacciones de rechazo en los expertos (Weir 1988; Bradshaw 1990; Jafarpur 1995), que cuestionan la validez aparente de la prueba. Sin embargo, mediante el análisis de los resultados del cuestionario retrospectivo de opinión hemos probado la validez aparente del C-test aplicado.

En sus respuestas al cuestionario los sujetos de la muestra manifiestan su aceptación del C-test y evidencian haber encontrado algunas claves de la prueba. En primer lugar, descubren la importancia del contexto para la recuperación de las omisiones, además son conscientes del papel del vocabulario, al que sobrevaloran. Por otra parte, reconocen aspectos que repercuten en el grado dificultad de la prueba, como la frecuencia de las omisiones, el que sean guiadas o no, el tema del texto, etc. Finalmente, descubren en la prueba un buen instrumento de medida de su competencia en lengua inglesa, aunque mencionan la novedad de la técnica y las dificultades encontradas en su resolución.

El análisis de frecuencias correspondiente a la valoración del C-test como instrumento de evaluación muestra que la mayoría de los sujetos considera que la prueba mide principalmente aspectos relativos al vocabulario (léxico y ortografía), aunque el conocimiento general de la lengua obtiene también un 3,36 en escala de 1 a 5 y la fluidez un 3,32. Los sujetos destacan que el C-test es una prueba adecuada (el 68% de los sujetos de la muestra otorgan más de un 3 a este rasgo en la escala de Likert de 1 a 5, siendo 5 la puntuación máxima), pero también completa (el 56% la valora en más de 3 puntos en la misma escala) y válida (52,5%). Estos datos contrastan con el estudio de Jafarpur (1995), en el que el 64% de los alumnos y el 57% de los profesores valoraron negativamente al C-test.

Ahora bien, cuando se planteó la posibilidad de que el C-test fuera la alternativa a la prueba de Inglés de Selectividad el 76% manifestaron su negativa. El porcentaje bajó considerablemente si se proponía al C-test sólo como parte o complemento de la prueba vigente. Estos porcentajes muestran en los alumnos la resistencia al cambio que Jafarpur (1995) señaló como rasgo del colectivo de los profesores, pero que es propio del ser humano, y el miedo a que en el breve tiempo

de unos meses cambiara el planteamiento de la prueba oficial en que se basaba su preparación a lo largo del curso.

Mediante el análisis factorial tratamos de encontrar grupos homogéneos de variables, y descubrimos que con sólo dos factores se explica el 53,7% de la varianza. El primer factor se refiere a la dimensión del vocabulario y está formado por las variables "léxico" y "ortografía". El segundo, formado por la "fluidez" y el "conocimiento general de la lengua" refleja el dominio de la lengua.

Por tanto, podemos concluir que los sujetos de la muestra identifican al C-test como prueba que mide el vocabulario y la competencia general en lengua inglesa. No lo rechazan por su formato, aunque reconocen la dificultad que supone su novedad. Queda rechazada la hipótesis 5, que planteaba que el C-test carece de validez aparente y achacaba el rechazo al C-test por parte de los alumnos como consecuencia de su novedad y carácter fragmentario.

B. Fiabilidad

El estudio de la fiabilidad del C-test muestra la consistencia entre distintas actuaciones del mismo sujeto. Para cuantificar la fiabilidad de la prueba se utilizó el método de "análisis por mitades", se calculó el Alfa de Cronbach y se revisó el análisis de las correlaciones con otras pruebas. Se rechazó el método *test-retest* por los posibles sesgos que habría introducido en la investigación (familiarización con la técnica y desmotivación).

La estructura del C-test aplicado permitió su división en dos mitades equivalentes para después asignar dos puntuaciones a cada alumno. El análisis de los estadísticos descriptivos evidenció la fiabilidad de la prueba puesto que se obtuvieron valores muy semejantes en los promedios (levemente superiores en la segunda mitad, fruto de la práctica), las puntuaciones máximas y mínimas, el error típico y la desviación. La correlación entre los resultados de ambas mitades fue muy significativa (0,816)

El C-test consigue un buen Alfa de Cronbach (0,794) y en total consonancia con los coeficientes de fiabilidad encontrados en el estudio de Dörnyei y Katona (1992: 193), se respalda así la teoría de Klein Braley y Raatz (1984: 140) de que en

los C-tests se encuentran coeficientes de validez y fiabilidad aceptables, incluso en los demasiado fáciles o difíciles para los sujetos. En este caso, el C-test presenta una dificultad media, como mencionamos en el apartado A.

Por otra parte, las correlaciones entre pruebas que miden el mismo constructo reflejan también consistencia en la actuación. Por tanto, si volvemos a los resultados reflejados en el apartado anterior constatamos una vez más la fiabilidad de la prueba.

En cuanto a la fiabilidad del corrector, necesaria para que una prueba sea fiable (Hughes 1989, 1994), hemos de decir que en el C-test las propias características del diseño de la prueba, que deja poco margen a la subjetividad, aseguran un alto grado de fiabilidad. El criterio de corrección elegido también garantiza la objetividad, puesto que sólo la palabra exacta se considera válida. Por tanto, la subjetividad queda limitada a las decisiones del profesor en la fase de diseño de la prueba cuando realiza la selección de los textos a partir de los cuales se creará el C-test. Una vez elegidos, las normas son estrictas (Klein-Braley 1997; Klein-Braley y Raatz 1981, 1984).

C. Incidencia de las variables género e IES

Los resultados del análisis del ANOVA, el modelo lineal general y los promedios indican cómo inciden las variables externas: género e IES de procedencia de los sujetos en su actuación en el C-test.

C.1. Incidencia del género de los sujetos en el C-test

Aunque los promedios muestran que las mujeres obtienen mejores resultados que los varones (recordemos que, como media, las mujeres recuperaron aproximadamente cuatro omisiones del C-test más que los varones), el análisis del ANOVA indica que no hay diferencias significativas en la actuación de ambos géneros. Por el contrario, el análisis multivariante sí refleja la existencia de diferencias de género en el C-test y en la 2ª Evaluación, cuando se toma el C-test

sin desglosar en subtests, con la muestra de los 81 sujetos que se presentaron a las PAAU oficiales. A pesar todo, las diferencias en los promedios son menores en el C-test que en el resto de las pruebas analizadas. También debemos mencionar que la desviación típica es ligeramente superior en las mujeres.

Así pues, los resultados de la variable género indican que no hay diferencias en la actuación de los géneros en el C-test si se toman los subtests individualmente y la muestra de sujetos completa. Pero cuando se reduce la muestra a los sujetos que se presentan a la Selectividad y se analiza el C-test globalmente, sí las hay, aunque otros factores pudieran justificarlas (motivación, interés vocacional, futuros estudios, etc.). Lo que nos lleva a confirmar la hipótesis 6: “No habrá diferencias significativas al aplicar la variable de género” en el primer caso, y a rechazarla en el segundo.

C.2. IES de procedencia de los sujetos

En cuanto a la variable IES de procedencia de los sujetos, nuestro análisis reveló diferencias en los promedios, fruto, entre otras variables, de las características del entorno socio-económico de los centros educativos (que repercuten en las oportunidades de aprendizaje de los sujetos fuera del IES). Puesto que la disparidad de promedios no implica diferencias significativas en la actuación, se realizó un análisis de varianza univariante que también mostró diferencias entre centros en los resultados del C-test, en concreto entre el IES San Isidoro de Madrid (en las pruebas post-hoc, sig. 0,005 con la corrección de Bonferroni y 0,006 aplicando la corrección de Games-Howell) y los de la periferia. Por otra parte, aunque se aprecian diferencias significativas inter-centros en los resultados de la prueba (debidas probablemente a circunstancias socio-económicas que no podemos valorar en este trabajo), al comparar el comportamiento de todas las pruebas aplicadas, vemos que el C-test funciona igual que cualquier otra prueba. Ha quedado demostrado que discrimina entre los sujetos atendiendo a su dominio de la lengua, independientemente de las características del centro educativo.

Hemos de rechazar la hipótesis: “No se prevé que existan diferencias de funcionamiento del C-test al aplicar la variable IES” cuando nos referimos a centros de distinto estatus. Cabría confirmarla en centros de estatus semejante.

Al aplicar el análisis de varianza univariante a las dos variables “género” e “IES” siguiendo el modelo lineal general, tomamos el total del C-test como variable dependiente y pudimos ver cómo afectan el género y el centro de los sujetos a la prueba. Descubrimos que ambas variables son independientes, y que el IES San Isidoro, de Madrid capital, presenta diferencias significativas con los otros tres que forman parte del estudio (sig. = 0,008 con Alcobendas y 0,000 con Pinto y Parla).

D. Implicaciones pedagógicas

Desde la Introducción de esta tesis se planteó la clara orientación pedagógica de la investigación que se ha llevado a cabo. Partimos del interés por encontrar instrumentos prácticos, válidos y fiables para la evaluación de la competencia en Inglés como Lengua Extranjera en la práctica docente.

El C-test ha demostrado reunir los rasgos que requiere un instrumento tal de evaluación. Se ha trabajado, sobre todo, la validez criterial concurrente y la aparente del C-test, su fiabilidad y factibilidad. Todas estas características han quedado empíricamente probadas en nuestra investigación.

El C-test es una prueba versátil y tiene algo que ofrecer a cada uno de los implicados en el proceso de enseñanza-aprendizaje del inglés.

Responde de forma efectiva a las demandas y necesidades del profesorado de idiomas en materia de evaluación. Para el profesorado, el C-test se revela como una prueba válida y fiable para medir la competencia global de los sujetos en Inglés, como muestran las elevadas correlaciones con otras pruebas y el análisis de regresión lineal realizado. Pero no sólo eso, su factibilidad resulta fundamental. La economía de tiempo y esfuerzo que ofrece la prueba a este colectivo, por su facilidad de diseño y corrección es difícilmente igualable por otras pruebas. Momento de especial importancia en la creación de C-tests es el de selección de textos, puesto que las características textuales inciden en el grado de dificultad de la prueba (tema, familiarización, tipo de término omitido, variación y densidad léxicas, etc.). Una vez elegidos recomendamos seguir rigurosamente las indicaciones de Klein-Braley y Raatz (1984). Tanto si se aportan pistas para la recuperación de las omisiones como si no, se consiguen C-tests válidos y fiables.

Por otra parte, su carácter de prueba objetiva facilita la corrección, pero no le impide ser buen predictor de la actuación del alumno en pruebas de tipo subjetivo, como los ensayos.

No sólo funciona bien como prueba de evaluación de distintos tipos (inicial de nivel, de control, evaluación final), también puede ser utilizado como actividad de aprendizaje en el aula (reflexión sobre la lengua, vocabulario, etc.), de repaso y revisión, incluso como ejercício de autoevaluación. Su diseño es muy apropiado para su explotación mediante las Nuevas Tecnologías, de hecho ya ha comenzado a utilizarse en versiones electrónicas, sobre todo en niveles universitarios.

Es obvio que la técnica del C-test resulta muy rentable; podríamos decir que el C-test da mucho y pide muy poco al profesor.

Ahora bien, a pesar de las virtudes de la prueba, no parece aconsejable limitar la evaluación del Inglés al uso de C-tests, pero sí incluir este tipo de prueba de cierre en baterías de pruebas y en pruebas estandarizadas, como la actual Selectividad.

En el contexto de las PAAU vigentes, ampliamente cuestionadas (Herrera 2005, García Laborda 2005, Sanz y Fernández 2005) recomendamos la inclusión de un C-test en lugar de otras preguntas que han demostrado su bajo poder discriminatorio (como la prueba de “verdadero o falso”).

En cuanto al alumnado, destacaremos la validez aparente de la prueba. El análisis del cuestionario retrospectivo de opinión muestra que se valora positivamente al C-test a pesar de su aspecto novedoso y algo confuso al principio, contrastando con los estudios de Bradshaw (1990) y Jafarpur (1995). Su resolución plantea un reto motivador al sujeto.

E. Consejos para la creación de C-tests

Tomando como referencia la investigación presentada en esta tesis, probadas las características básicas de la prueba (validez, fiabilidad, factibilidad, etc.), desde este trabajo animamos a docentes de Inglés como Lengua Extranjera y expertos en evaluación a la creación y aplicación de C-tests, tanto en el ámbito del aula (de cualquier nivel) como en el de exámenes estandarizados.

En las clases de idiomas el C-test es un ejercicio interesante y motivador para reflexionar sobre la lengua, para introducir o fijar vocabulario, como autoevaluación después de haber trabajado un texto concreto, etc. Las posibilidades de aplicación dependen más de la creatividad del profesor, que puede adaptarla a su contexto y necesidades, que de la propia prueba. Y por supuesto, como ha demostrado nuestra investigación, funciona muy bien como prueba de evaluación de la competencia lingüística global del alumno.

Por eso, partiendo de las indicaciones de expertos, tales como Hughes (1989: 37) y Klein-Braley (1997: 64) y de la propia experiencia en la elaboración de C-tests, presentamos las siguientes sugerencias para la creación de C-tests.

La técnica del C-test permite seguir las recomendaciones de Hughes (1989); proponer una tarea precisa y controlada, clara y sin ambigüedades para el alumno, completa en cuanto al número de ítems pero no agobiante.

La secuencia para diseñar un buen C-test incluye los pasos siguientes:

- Pre-selección de un buen número de textos, preferiblemente auténticos, pero siempre adaptados al nivel de competencia de los sujetos. La selección final incluirá 4 ó 5 de nivel y tema adecuado, ordenados según su grado de dificultad. El C-test funcionará incluso con textos fáciles, debido a su potencia discriminatoria y carácter motivador.
- Mutilación de los textos elegidos aplicando las normas y parámetros de Klein-Braley y Raatz (1984).
- Administración de la prueba a un grupo de control (nativos o buenos conocedores de la lengua objeto de estudio), que ha de conseguir puntuaciones en torno al 90% de aciertos.

Si los resultados muestran la validez y fiabilidad de la prueba, pueden aplicarse a los grupos a que van dirigidos, siempre cuidando los siguientes aspectos, comunes a cualquier otro tipo de prueba de evaluación:

- Familiarizar al alumno con la técnica y formato de la prueba.
- Cuidar la tipografía, el orden y la claridad.

- Proporcionar instrucciones claras y precisas, orales y escritas. También es aconsejable aportar un modelo de realización del C-test cuando es totalmente nuevo para los sujetos.
- Dejar tiempo suficiente en la administración.
- Procurar que las condiciones de administración del examen sean las adecuadas: duración de la prueba, lugar, luz, condiciones acústicas, temperatura, silencio, etc.

El buen criterio del profesor debe guiar siempre el diseño de las pruebas. Dependiendo del contexto, del nivel de competencia del alumnado al que van dirigidos y los objetivos que se pretenda lograr, se utilizará un tipo u otro de texto (materiales auténticos, adaptados, previamente trabajados, nuevos, etc.) para la creación de C-tests.

F. Síntesis de los resultados más relevantes del estudio

A continuación resumiremos los resultados más significativos de esta tesis, que, no obstante, deben ser interpretados con cautela y debidamente contrastados en investigaciones posteriores:

1. Los promedios globales obtenidos en el C-test aplicado (5,112 puntos) muestran su potencia discriminatoria como prueba de tipo normativo, de dificultad media (Klein-Braley 1984). Los promedios de los subtests varían en función de su grado de dificultad, motivado por distintos factores.

2. Entre los factores que inciden en el grado de dificultad de la prueba destacan las características textuales y el formato de las omisiones. El formato de omisiones no guiadas (subtests 3 y 4) aumenta la dificultad de la prueba, pues los promedios descienden sensiblemente y la dispersión de puntuaciones aumenta, a pesar de la familiarización con la técnica. Pero se consiguen pruebas válidas y fiables independientemente del formato. En cuanto al texto, se hace hincapié en la apropiada selección (Dörnyei y Katona 1992, Babaii y

Ansary 2001), pues aunque se logran C-tests válidos en textos difíciles (Klein Braley y Raatz 1984), el funcionamiento de la prueba mejora en los fáciles y adecuados al nivel de los sujetos, que resultan además motivadores. En nuestro estudio se ha analizado principalmente la incidencia del tema, variación y densidad léxicas, y el tipo de palabra afectada por la mutilación (términos léxicos y funcionales). El conocimiento e interés por el tema aumenta la motivación y facilita la tarea. A mayor variación y densidad léxicas, mayor dificultad en la resolución del C-test. Aunque se recuperan mejor los términos funcionales, reconocemos la influencia de otros factores, como la frecuencia de uso, el tamaño y grado de abstracción de los términos léxicos.

3. El análisis de las correlaciones del C-test con otras pruebas indica su validez criterial concurrente al medir la competencia global de los alumnos españoles de 2º de Bachillerato en Inglés como Lengua Extranjera. Los valores de correlación del C-test con las PAAU oficiales (0,722), con el modelo aplicado en el aula (0,750) y con la valoración de los profesores de Inglés en la 2ª Evaluación (0,723) son muy semejantes, y están en la línea de los obtenidos en otros estudios similares (Dörnyei y Katona 1992). Si tenemos en cuenta las dos partes (objetiva y subjetiva) que forman la PAAU de Inglés vigente vemos que el C-test correlaciona mejor con la parte subjetiva, que es también la que mejor discrimina entre los sujetos. Con las preguntas abiertas de *Cavemen?* se obtuvo una correlación de 0,616 y con la redacción 0,665. Obviamente, el C-test también ha mostrado en nuestro estudio su excelente correlación con la Gramática (0,672). Por todo ello, atendiendo a los resultados obtenidos y garantizada su validez, el C-test puede ser utilizado en el aula como preparación de las PAAU, como mera actividad de aprendizaje o autoevaluación, o bien podría formar parte de cualquier prueba estandarizada a gran escala, como la PAAU.

4. La fiabilidad o consistencia del C-test queda patente tanto en el Alfa de Cronbach (0,794) como en el análisis *split-half*, en el que la correlación entre los resultados de las dos mitades fue de 0,816. Asimismo, las características del diseño de la prueba, de tipo objetivo, a la vez que facilitan la tarea de la

corrección, aseguran un alto grado de fiabilidad del corrector pues no hay lugar para la subjetividad.

5. Contrastando con el punto de vista reflejado en los estudios de Bradshaw (1990) y Jafarpur (1995), nuestra investigación confirma la validez aparente del C-test. Aparece reflejada en los datos de nuestro cuestionario, que indaga en la opinión de los sujetos acerca de la prueba. Las escalas de Likert y el análisis factorial muestran que, a pesar de su novedad, los alumnos aceptan al C-test como prueba válida para medir su competencia en lengua inglesa, y reconocen la importancia de las claves contextuales para recuperar el texto y la ayuda que aportan las omisiones guiadas.

6. Siguiendo la pauta de otros estudios, puede observarse que los promedios en el C-test son más altos en el grupo de mujeres que en el de los varones. No obstante, el análisis del ANOVA revela que no hay diferencias significativas en la actuación de ambos géneros cuando se toman los subtests individualmente. Resultado que contrasta con el obtenido mediante el análisis multivariante, que refleja la existencia de diferencias entre géneros en el C-test y en la 2ª Evaluación, si se toma el C-test globalmente para el grupo de sujetos de mayor competencia lingüística, es decir, los que se presentaron a las PAAU oficiales.

7. Hay diferencias según el centro en función del medio social. Se aprecian en los promedios obtenidos en el C-test y en el análisis de varianza univariante. Pueden explicarse por las características de los entornos socio-económicos en que se encuentran ubicados los IES participantes en el estudio, que suponen la existencia o no de oportunidades de aprendizaje fuera de la escuela. Si se aplica el procedimiento de análisis de varianza univariante conjuntamente a las variables “género” e “IES” siguiendo el modelo lineal general, se observa la independencia de ambas variables, y que el IES San Isidoro, de Madrid capital, presenta diferencias significativas con los otros tres que forman parte del estudio (sig. = 0,008 con Alcobendas y 0,000 con Pinto y Parla), sus elevados promedios destacan siempre.

8. A lo largo de nuestro trabajo el C-test se ha revelado como una prueba práctica y rentable, en la línea de autores como Süssmilch (1984), Dörnyei y Katona (1992), Klein-Braley (1997), Connelly (1997), Babaii y Ansary (2001). Su factibilidad y versatilidad como instrumento de evaluación no admite discusión. No es que el C-test valga "para todo" (Klein-Braley 1997), pero sí muestra rasgos tan ventajosos para el desarrollo de la tarea de evaluación de la lengua que merecen ser tomados en cuenta.

G. Propuesta de posibles futuras líneas de investigación

No podemos concluir nuestro trabajo sobre el C-test sin reconocer la amplitud de las vías de investigación que quedan todavía abiertas en torno a esta prueba de evaluación.

Proponemos continuar el trabajo en los siguientes campos:

- Desde la lingüística contrastiva, comparando su funcionamiento en sujetos de otras lenguas maternas y en el aprendizaje de Segundas Lenguas de características diferentes al Inglés.
- Desde la validez de constructo. Queda abierto el reto de conocer qué mide realmente el C-test, pues hasta ahora tan sólo se han apuntado ideas que es necesario acotar y precisar.
- Con el estudio de la validez del C-test en sujetos con distintos niveles de competencia en la lengua (niveles básicos, ESP, etc.).
- Abordando su validez aparente mediante el análisis de la opinión del profesorado de idiomas.
- En el análisis de sus posibles aportaciones en el aula de Lenguas Extranjeras como instrumento de aprendizaje más que de evaluación.
- Valorando el papel de las estrategias de aprendizaje de lenguas que se utilizan en su resolución.
- En la investigación y desarrollo de la aplicabilidad de la prueba en el ámbito de las Nuevas Tecnologías (Internet).

BIBLIOGRAFÍA

Aarts, F. y J. Aarts (1988) *English Syntactic Structures: Functions and Categories in Sentence Analysis.* New York: Prentice-Hall.

Adair, J. G. (1984) The Hawthorne Effect: A Reconsideration of the Methodological Artifact. *Journal of Applied Psychology* 69 (2), 334-345.

Adair, J. G. *et al.* (1989) Hawthorne Control Procedures in Educational Experiments: A Reconsideration of Their Use and Effectiveness. *Review of Educational Research* 59 (2), 215-228.

Alarcos, E. (1994) *Gramática de la Lengua Española*. Madrid: Espasa Calpe S.A.

Alcaraz, E. y J. Ramón (1980) *La evaluación del inglés. Teoría y práctica*. Madrid: SGEL, S.A.

Alcina J. y J. M. Blecua (1975) *Gramática española*. Barcelona: Ed. Ariel.

Alderson, J. C. (1979) The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly* 13, 219-23.

Alderson, J. C. (1980) Native and non-native speaker performance on cloze tests. *Language Learning* 30, 59-76.

Alderson, J. C. (1990) Testing Reading Comprehension Skills (Part I). *Reading in a Foreign Language* 6 (2), 425-438.

Alderson, J. C. (1991) Language Testing in the 1990´s: How far have we come? How much further have we to go? En Anivan, S., ed. *Current developments in language testing*, 1-26. Singapore: Regional Language Center.

Alderson, J. C. (1995) Assessing Student Performance in the ESL Classroom. *TESOL Quarterly* 29 (1), 184-187.

Alderson, J. C. (2000) *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. C. y J. Banerjee (2001) Language Testing and Assessment (Part I) *Language Teaching: The International Abstracting Journal* 213-236. Cambrige: Cambrige University Press.

Alderson, J. C. y G. Buck (1993) Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing. *Language Testing* 10 (1), 1-26.

Alderson, J. C. y L. Hamp-Lyons (1996) TOEFL preparation courses: a study of washback. *Language Testing* 13 (3), 280-297.

Alderson, J. C. y D. Wall (1993a) Does washback exist? *Applied Linguistics* 14 (2), 115-129.

Alderson, J. C. y D. Wall (1993b) Examining washback: The Sri Lankan Impact Study. *Language Testing* 10 (1), 41-69.

Allan, D. (1999) Testing and Assessment. *English Teaching Professional* 11, 19.

Álvarez Méndez, J. M. (2001) E*valuar para conocer, examinar para excluir*. Madrid: Ediciones Morata.

Álvarez Méndez, J. M. (2003) *La evaluación a examen*. Ensayos críticos. Madrid: Miño y Dávila Eds.

Andrés Cortés, J. (2004) Análisis Lingüístico de Términos Comparados en Inglés Técnico Agrícola. Tesis doctoral sin publicar. Madrid: Universidad Complutense.

Andrews, S. *et al.* (2002) Targeting washback –a case-study. *System* 30, 207-223.

Amengual Pizarro, M. (2003) Análisis de la fiabilidad en las puntuaciones holísticas de ítems abiertos. Tesis doctoral sin publicar. Madrid: Universidad Complutense.

Amengual Pizarro, M. (2005) Posibles sesgos en el examen de Selectividad. En Herrera Soler, H. y J. García Laborda, eds. *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*, 121-148. Valencia: Ed. Universidad Politécnica de Valencia.

Amengual Pizarro, M. *et al.* (2001) Discrepancy in ratings of second language performance. *La lingüística española a finales del siglo XX. Ensayos y propuestas*, Tomo I, 23-29. AESLA. Universidad de Alcalá.

Amengual Pizarro, M. y H. Herrera Soler (2001) Rater's assumptions about form and content. En *Trabajos en Lingüística aplicada*, 63-71. AESLA. Barcelona.

Amengual Pizarro, M. y H. Herrera Soler (2003) What is it that raters are judging? *Las lenguas en un mundo global*, 319. AESLA. Universidad de Jaén.

Arnaud, P. J. L. (1984) The lexical richness of L2 written productions and the validity of vocabulary tests. En Culhane, T. *et al*., eds. *Practice and problems in language testing*. *Occasional Papers* 29, 14-28. Colchester: University of Essex.

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.

Bacha, N. (2001) Writing evaluation: what can analytic versus holistic essay scoring tell us? *System* 29, 371-383.

Bachman, L. F. (1982) The trait structure of cloze test scores*. TESOL Quarterly* 16, 61-70.

Bachman, L. F. (1985) Performance on cloze tests with fixed ratio and rational deletions*. TESOL Quarterly* 19, 535-56.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000) Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing* 17 (1), 1-42.

Bachman, L. F. (2004) *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

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

Bachman, L. *et al*. (1996) The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing* 13 (3), 125-149.

Bailey, K. (1996) Working for washback: a review of the washback concept in language testing. *Language Testing* 13 (3), 257-259.

Bello, A. (1984) *Gramática de la Lengua Castellana.* Madrid: EDAF, S.A.

Bensoussan, M. y R. Ramraz (1984) The fill-in test: a modified multiple-choice cloze technique to test reading comprehension of English as a foreign language. En Culhane *et al*., eds. *Practice and problems in language testing. Occasional Papers* 29, 44-65. Colchester: University of Essex.

Bialystok, E. (1998) Coming of Age in Applied Linguistics. *Language Learning* 48 (4), 497-518.

Bocanegra, A. (2001) El aula de lenguas segundas/extranjeras como contexto para la generación y procesamiento del aducto. En *La Lingüística Aplicada a finales del siglo XX. Ensayos y propuestas.* Tomo 1, 31-43. AESLA. Universidad de Alcalá.

Bocanegra A. y P. Franco (2003) El aprendizaje estratégico de los estudiantes principiantes y avanzados de inglés como lengua extranjera. En *Las lenguas en un mundo global*, 320. AESLA. Jaén.

Bogaards, P. (2000) Testing L2 Vocabulary Knowledge at a High Level: the Case of the *Euralex French Tests*. *Applied Linguistics* 21 (4), 490-516.

Bradshaw, J. (1990) Test-takers' reactions to a placement test. *Language Testing* 7 (1), 13-30.

Braine, G. (2001) When an exit test fails. *System* 29, 221-234.

Broadfoot (2005) Dark alleys and blind bends: testing the language of learning. *Language Testing* 22 (2), 123-141.

Brown, J. D. (1983) A closer look at cloze: Validity and reliability. En J. W. Oller, Jr. ed, *Issues in language testing research,* 237-50. Rowley, Massachusetts: Newbury House Publishers.

Brown, J. D. (1988) *Understanding Research in Second Language Learning*. Cambridge: Cambridge University Press.

Brown, J. D. (1988) Tailored cloze: Improved with classical item analysis techniques. *Language Testing* 5, 19-31.

Brown, J. D. (1993) What are the characteristics of *natural* cloze tests? *Language Testing* 10 (2), 93-116.

Brumfit, C. J. (2001) *Individual freedom in language teaching*. Oxford: Oxford University Press.

Butler, C. (1985) *Statistics in Linguistics*. Oxford: Blackwell.

Butler, F. A. y R. Stevens (2001) Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas. *Language Testing* 18 (4), 409-427.

Bybee, J. (1995) Diachronic and Typological Properties of Morphology and their Implications for Representation. *Morphological Aspects of Language Processing*, 226-246. Feldman L. B., ed. Hillsdale (NJ): Lawrence Erlbaum.

Bygate, M. (2004) Some current trends in applied linguistics: Towards a generic view. *AILA Review* 17, 6-22.

Cabré T. y A. Adelstein (2001) ¿Es la terminología lingüística aplicada? En *Trabajos en Lingüística aplicada*, 387-393. AESLA. Barcelona.

Cameron, D. (2005) Language, Gender, and Sexuality: Current Issues and New Directions. *Applied Linguistics* 26 (4), 482-502.

Canale, M. (1988) The measurement of communicative competence. *Annual Review of Applied Linguistics* 8, 67-84.

Canale, M. y M. Swain (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1-47.

Carrol, J. B. (1982) Language testing –is there another way? En Heaton, J. B., ed. *Language Testing*. UK: Modern English Publications Ltd.

Carroll, J. B. (1987) Review of "C-tests in der Praxis", AKS Rundbrief 13/14 Bochum, 1985. *Language Testing* 4, 99-106.

Carter, R. (1987) *Vocabulary: Applied Linguistic Perspective* (2nd Edition) London: Routledge.

Carter, R. (1988) Vocabulary, cloze and discourse: an applied linguistic view. En Carter, R. y M. McCarthy, eds. *Vocabulary and Language Teaching.* London: Longman.

Carter, R. y M. McCarthy, eds. (1988) *Vocabulary and Language Teaching.* London: Longman.

Catford, J. C. (1998) Language Learning and Applied Linguistics: a Historical Sketch. *Language Learning* 48 (4), 465-496.

Chalhoub-Deville, M. (1997) Theoretical Models, Assessment Frameworks and Test Construction. *Language Testing* 14 (1), 3-22.

Chalhoub-Deville, M. (2003) Second language interaction: current perspectives and future trends. *Language Testing* 20 (4), 369-383.

Chamot, A. U. y J. M. O'Malley (1994) Language Learner and Learning Strategies. En Ellis, N. C., ed. *Implicit and explicit learning of languages*. London: Academic Press.

Chapelle, C. A. (1994) Are C-tests valid measures for L2 vocabulary research? *Second Language Research* 10, 157-187.

Chapelle, C. A. y R. G. Abraham (1990) Cloze Method: What Difference does it make? *Language Testing* 7 (2), 121-146.

Chapelle, C. A. et al. (2003) Validation of a web-based ESL test. *Language Testing* 20 (4), 409-439.

Chaudron, C. *et al*. (2001) La composición como comunicación: influencia en el desarrollo general del inglés como segunda lengua en un contexto de instrucción. *La lingüística española a finales del siglo XX. Ensayos y propuestas*. Tomo I, 54-62. AESLA. Universidad de Alcalá.

Chavez-Oller, M. A. *et al.* (1985) When are cloze items sensitive to constraints across sentences? *Language Learning* 35, 181-206.

Chihara, T. *et al*. (1977) Are cloze items sensitive to constraints across sentences? *Language Learning* 27, 63-73.

Cohen, A. D. *et al*. (1984) The C-test in Hebrew. Research Note. *Language Testing* 1, 221-225.

Connelly, M. (1997) Using C-Tests in English with Post-Graduate Students. *English for Specific Purposes* 16 (2), 139-150.

Cook, V. (1996) (2nd Edition) *Second Language Learning and Language Teaching*. London: Arnold.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cronbach, L. J. (1971) (2nd Edition) Test validation. En Thorndike, R.L. Ed. *Educational Measurement*. Washington, DC: American Council of Education.

Culhane, T. *et al*. Eds. (1984) *Practice and problems in language testing*. *Occasional Papers* 29. Colchester: University of Essex.

Cumming, A. (1990) Expertise in evaluating second language compositions. *Language Testing* 7 (1) 31-51.

Cumming, A. (1996) Introduction: The Concept of Validation in Language Testing. En Cumming, A. y R. Berwick, eds. *Validation in Language Testing*, Modern Languages in Practice 2, 1-14. Clevendon: Multilingual Matters Ltd.

Cumming, A. y R. Berwick, eds. (1996) *Validation in Language Testing*. Modern Languages in Practice 2. Clevendon: Multilingual Matters Ltd.

Cushing, S. (1994) Effects of training on raters of ESL compositions. *Language Testing*11 (2) 197-223.

Dastjerdi, H. V. y M. R. Talebinezhad (2006) Chain-preserving deletion procedure in cloze: a discoursal perspective. *Language Testing* 23 (1) 58-72.

Davies, A. (1997) The limits of ethics in language testing. *Language Testing* 14 (3), 235-241.

Davies, A. (1997) Demands of being professional in language testing. *Language Testing* 14 (3), 328-339.

Davies, A. (2003) Three heresies of language testing research. *Language Testing* 20 (4), 355-367.

Davidson, F. (1996) *Principles of Statistical Data Handling*. California: SAGE Publications, Inc.

Decoo, W. (2003) Language Methods and CALL: Redefining our Relations. *Computer Assissted Language Learning* 16 (4), 269-274.

Denton, J., Lewis, R. y A. Siles Suárez (1996) *Themes for 1º Bachillerato*. Burlington Books.

Doménech, J. M. (2001) Fundamentos de diseño y estadística. UD 13: Correlación y regresión lineal. Barcelona: Signo.

Dörnyei, Z. (2003) Questionnaires in Second Language Research: Construction, Administration, and Processing. Mahwah, NJ.: Lawrence Erlbaum Associates.

Dörnyei, Z. y L. Katona (1992) Validation of the C-Test amongst Hungarian EFL Learners. *Language Testing* 9, 187-206.

Doughty, C. I. y M. H. Long, eds. (2003) *The Handbook of Second Language Acquisition*. Oxford: Blackwell Publishing Ltd.

Douglas, D. y C. Chapelle (1993) *A New Decade of Language Testing Research.* Alexandria, Virginia: TESOL.

Eckardt, A. y B. Voss (2006) The UNIcert® initiative -an Update. [Documento de Internet disponible en http://www.acad.polyu.edu.hk/~02900821r/ilta/rr-2.htm].

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

Eco, U. (1977) *Come si fa una tesi di laurea. Le materie umanistiche*. Milano: Tascabili Bompiani.

Elder, C. (1997) What does test bias have to do with fairness? *Language Testing* 14 (3), 261-277.

Ellis, N. C. y A. Beaton (1993) Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning* 43 (4), 559-617.

Ellis, R. (1985) *Understanding Second Language Acquisition*. Oxford: Oxford University Press.

Ellis, N. C., ed. (2000) (1st Edition 1994) *Implicit and explicit learning of languages*. London: Academic Press.

Esteban, M., Herrera, H. y M. Amengual (2001) ¿Puede el C-test ser una alternativa a otras pruebas en la enseñanza del inglés como segunda lengua? *La lingüística española a finales del siglo XX. Ensayos y propuestas*. Tomo I, 169-175. AESLA. Universidad de Alcalá.

Esteban, M., Herrera, H. y M. Amengual (2000) Niveles de correlación entre el C-test y las pruebas de Inglés de Selectividad. Comunicación al XIX Congreso Nacional de AESLA. Universidad de León.

Esteban, M. y H. Herrera (2003) El C-test: instrumento apropiado para la evaluación de la competencia en inglés como lengua extranjera. *Las lenguas en un mundo global*, 323. 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.

Falk, B. (1984) Can grammatical correctness and communication be tested simultaneously? En Culhane,T. *et al*., eds. *Practice and problems in language testing. Occasional Papers* 29, 90-96. Colchester: University of Essex.

Farhady, H. (1979) The disjunctive fallacy between discrete-point and integrative tests. *TESOL Quarterly* 13, 347-357.

Farhady, H. y M. N. Keramati (1996) A text-driven method for the deletion procedure in cloze passages. *Language Testing* 13 (2), 191-207.

Feldmann, U. y B. Stemmer (1987) Thin___ aloud a___ retrospective da___ in C-te___ taking: diffe___ languages- diff___ learners- sa___ approaches?. En Faerch, C. y G. Kasper, eds. *Introspection in Second Language Research*. Clevendon: Multilingual Matters Ltd.

Fernández Álvarez, M. e I. Sanz Sáiz (2005) Breve historia del examen de Selectividad. En Herrera Soler, H. y J. García Laborda, eds. *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*, 19-26. Valencia: Ed. Universidad Politécnica de Valencia.

Fernández Álvarez, M. e I. Sanz Sáiz (2005) Metodología para el diseño de una prueba de Inglés en Selectividad. En Herrera Soler, H. y J. García Laborda, eds. *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*, 41-62. Valencia: Ed. Universidad Politécnica de Valencia.

Fernández Toledo, P. (2001) Uso de estrategias discursivas y de género en la comprensión lectora de inglés como lengua extranjera. En *Perspectivas recientes sobre el discurso*, 152. AESLA Universidad de León.

Fotos, S. (1991) The Cloze Test as an Integrative Measure of EFL Proficiency: A Substitute for Essays on College Entrance Examinations. *Language Learning* 41 (3), 313-336.

Fox. J. (2004) Test decisions over time: tracking validity. *Language Testing* 21 (4), 437-465.

Freedle, R. y I. Kostin (1999) Does the test matter in a multiple-choice test of comprehension? The case for the construct validity of TOEF's minitalks. *Language Testing* 16 (1), 2-32.

Fries, C. C. (1945) *Teaching and Learning English as a Second Language*. Michigan: University Press.

Fukkink, R. G. *et al.* (2001) Deriving Word Meaning from Written Context: A Multicomponential Skill. *Language Learning* 51 (3), 477-496.

Fulcher, G. (1997) An English language placement test: issues in reliability and validity. *Language Testing* 14 (2), 113-138

Fulcher, G. (1999a) Assessment in English for Academic Purposes: Putting Content Validity in Its Place. *Applied Linguistics* 20 (2), 221-236.

Fulcher, G. (1999b) Ethics in Language Testing. *TAE SIG Newsletter* 1 (1), 1-4. [Documento de Internet disponible en http://taesig.8m.com/news1.html].

Gamaroff, R. (2000) Rater reliability in language assessment: the bug of all bears. *System* 28, (31-53).

García Hoz, V. (1992) *Enseñanza y aprendizaje de las lenguas modernas*. Madrid: Ediciones Rialp, S.A.

García Laborda, J. (2005) Un análisis cualitativo de la Selectividad de Inglés abierto a la esperanza. En Herrera Soler, H. y J. García Laborda, eds. *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*, 27-40. Valencia: Ed. Universidad Politécnica de Valencia.

García Laborda, J. y L. G. Bejarano (2005) Análisis de la necesidad de creación de páginas web para la evaluación y baremación de estudiantes internacionales: una experiencia internacional. En *Actas de la XXII Edición del Congreso Internacional de la Asociación Española de Lingüística Aplicada* (AESLA).

García Laborda, J. y E. Enríquez Carrasco (2005) Expectativas institucionales del proyecto HIELO/HIEO en Internet (e intranet) en la baremación inicial (diagnóstico) de Lenguas para Fines Específicos a gran escala. *Las TIC en el aula*, 100-110. Madrid: UNED.

Gibbons, J. y E. Ramírez (2004) *Maintaining a Minority Language. A Case Study of Hispanic Teenagers.* Clevendon: Multilingual Matters Ltd.

Giné, N. y A. Parcerisa (2000) *Evaluación en la educación secundaria*. *Elementos para la reflexión y recursos para la práctica*. Barcelona: Ed. GRAÓ.

Gipps, C. (1994) *Beyond Testing.* London: The Falmer Press.

González-Cascos, E. (2000) La evaluación de la L2. En Ruiz, J. *et al*. (Coord.) *Estudios de metodología de la lengua inglesa.* Valladolid: Centro Buendía, Universidad de Valladolid.

Graña López, B. (1997) Frecuencia y procesamiento léxico. *Revista Española de Lingüística Aplicada* 12, 27-41.

Green, A. B. y C. J. Weir (2004) Can placement test inform instructional decisions? *Language Testing* 21 (4), 467-494.

Grotjahn, R. (1986) Test validation and cognitive psychology: some methodological considerations. *Language Testing* 3 (2), 159-85.

Grotjahn, R. (1987) On the Methodological Basis of Introspective Methods. En Faerch, C. y G. Kasper, eds. *Introspection in Second Language Research*. Clevendon: Multilingual Matters Ltd.

Goulden, R., Nation, P. y J. Read (1990) How Large Can a Receptive Vocabulary Be? *Applied Linguistics* 11, 431-359.

Hadley G. S. y J. E. Naaykens (2006) An Investigation of the Selective Deletion Cloze Test as a Valid Measure of Grammar-Based Proficiency in Second Language Learning. [Documento de Internet disponible en http://www.nuis.ac.jp/~hadley/publication/nucloze/NUCLOZE.htm].

Halliday, M. A. K. (1987) *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M. A. K. y R. Hasan (1987) *Cohesion in English*. London: Longman.

Hamilton, J. *et al*. (2001) Teachers perceptions of on-line rater training and monitoring. *System* 29, 505-520.

Hamp-Lyons, L. (1991) Scoring procedures for ESL contexts. En Hamp-Lyons, L., ed. *Assessing Second Language Writing in Academic Contexts*, 241-276. Norwood, NJ: Ablex.

Hamp-Lyons, L. (1997) Washback, impact and validity: ethical concerns. *Language Testing* 14 (3), 295-303.

Hatch, E. y C. Brown (1995) *Vocabulary, Semantics and Language Education.* Cambridge: Cambridge University Press.

Heaton, J. B., ed. (1982) *Language Testing*. UK: Modern English Publications Ltd.

Heinlenman, L. (1983) The use of a cloze procedure in foreign language placement. *The Modern Language Journal* 67, 121-6.

Henricksen, B. (1999) Three dimensions on vocabulary development. *Studies in Second Language Acquisition* 21 (2), 303-317.

Herrera Soler, H. (1999) Is the English test in the Spanish University Entrance Examination as discriminating as it should be? *Estudios Ingleses de la Universidad Complutense* 7, 89-107.

Herrera Soler, H. (2000-2001) The effect of gender and working place of raters on University Entrance Examination scores. *RESLA* 14, 161-179.

Herrera Soler, H. (2005) El test de elección múltiple: herramienta básica en la Selectividad. En Herrera Soler, H. y J. García Laborda, eds. *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*, 65-96. Valencia: Ed. Universidad Politécnica de Valencia.

Herrera Soler, H. *et al*. (1999) Lectura de una prueba de selectividad desde una perspectiva pitagórica. *La lingüística española a finales del siglo XX. Ensayos y propuestas* Tomo I, 177-183. Universidad de Alcalá.

Herrera Soler, H. y C. Martínez Arias (2002) A new insight into examinee behaviour in a multiple-choice test: a quantitative approach. *Estudios Ingleses de la Universidad Complutense* 10, 113-137.

Huddleston, R. (1988) *English Grammar.* Cambridge: Cambridge University Press.

Hughes, A. (1989) *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Huhta, A. y R. Randell (1995) Multiple-choice Summary: A Measure of Test Comprehension. En Cumming, A. y Berwick, R., eds. *Validation in Language Testing*. Modern Languages in Practice 2, 94-110. Clevendon: Multilingual Matters Ltd.

Huhta, A. *et al*. (2006) Discursive construction of a high-stakes test: the many facets of a test-taker. *Language Testing* 23 (3), 326-350.

Ikeguchi, C. B. (1998) Do different C-tests discriminate proficiency levels of EL2 learners? *JAALT Testing and Evaluation SIG Newsletter* 2 (1), 3-8. [Documento de Internet disponible en http://www.jalt.rg/test/ike_1.htm].

ILTA (2000) *Code of Ethics for ILTA*. [Documento de Internet disponible en http://www.Dundee.ac.uk./languagestudies/1test/ilta/ilta_test2.html].

In'nami, Y. (2006) The effects of text anxiety on listening test performance. *System* 34, 317-340

Jafarpur, A. (1995) Is C-testing superior to cloze? *Language Testing* 12 (2) 194-216.

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

Johnson, R. K. (1982) Questioning some assumptions about cloze testing. En Heaton, J. B., ed. (1982) *Language Testing*. UK: Modern English Publications Ltd.

Johnstone, R. (2002) Research on language teaching and learning: 2001. *Language Teaching* 35, 157-181.

Jonz, J. (1987) Textual cohesion and second language comprehension. *Language Learning* 37, 409-38.

Jonz, J. (1991) Cloze item types and second language comprehension. *Language Testing* 8 (1), 1-22.

Kamimoto, T. (1989) C-tests and stylistic variation. Unpublished M. A. TEFL Dissertation. University of Reading. [Documento de Internet disponible en www.melta.org.my/modules/sections/12.doc].

Kamimoto, T. (2001) An examination of Nation's (1990) Vocabulary Levels Test. Paper presented at JALT. [Documento de Internet disponible en http://www1.harenet.ne.jp/~waring/vocab/colloquium/tad2001.htm].

Kees de Bot (2004) Applied linguistics in Europe. *AILA Review* 17, 57-68.

Katona L. y Z. Dörnyei (1993) The C-test: A Friendly Way to Test Language Proficiency. *English Teaching FORUM on line* 31 (2), 35.

Katona L. y Z. Dörnyei (2004) What the C-test is. [Documento de Internet disponible en http://effortlessacquisition.blogspot.com/2004/10/what-c-test-is.html].

Klein-Braley, C. y U. Raatz (1984) A survey of research on the C-test. *Language Testing* 1, 134-146.

Klein-Braley, C. (1985) A cloze-up on the C-test: a study in the construct validation of authentic tests. *Language Testing* 2, 76-104.

Klein-Braley, C. (1984) Advanced prediction of difficulty with C-tests. En Culhane, T. *et al*., eds. *Practice and problems in language testing*. *Occasional Papers* 29, 97-112. Colchester: University of Essex.

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

Kokkota, V. (1988) Letter-deletion procedure: a flexible way of reducing text redundancy. *Language Testing* 5 (1), 115-119.

Köler, W. (1972) *Psicología de la forma. Su tarea y últimas experiencias*. Madrid: Biblioteca Nueva.

Köler, W. (1998) *El problema de la* p*sicología de la forma.* Madrid: Facultad de Filosofía. Universidad Complutense*.*

Kroll, F. *et al*. (2002) The development of lexical fluency in a second language. *Second Language Research* 18 (2), 137-171.

Lado, R. (1961) *Language Testing.* London: Longman.

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

Laufer, B. *et al*. (2004) Size and strength: do we need both to measure vocabulary knowledge? *Language Testing* 21 (2), 202-226.

Laufer, B. y P. Nation (1995) Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics* 16 (3), 307-322.

Laufer, B. y P. Nation (1999) A vocabulary-size test of controlled productive ability. *Language Testing* 16 (1), 33-51.

Laufer, B. y Hulstijn, J. (2001) Incidental Vocabulary Acquisition in a Second Language: The Construct of Task-Induced Involvement. *Applied Linguistics* 22 (1), 1-26.

Lawley, J. y R. Fernández (1998) *Exam Strategies.* Madrid: Alhambra Longman, S.A.

Lawson, M. J. and D. Hogden (1996) The vocabulary-learning strategies of foreign-language students. *Language Learning* 46, 101-135.

Lee, S. (1996) The Concurrent Validity of Cloze Test with Essay Test among Korean Students. *Texas Papers in Foreign Language* 2 (2), 57-69.

Lee, S. H. (2003) ESL learner's vocabulary use in writing and the effects of explicit vocabulary instruction. *System* 31, 537-561.

Lee Y. P. (1985) Investigating the validity of the Cloze Score. En *New Directions in Language Testing*, 137-147. Lee *et al*., eds. Oxford: Pergamon Press.

Lee, Y. P. *et al*., eds. (1985) *New Directions in Language Testing*. Papers presented at the International Symposium on Language Testing, Hong Kong. Oxford: Pergamon Press.

Linn, R. L. (1993) *Educational Measurement*. American Council on Education. (3rd edition). Phoenix: Oryx Press.

Little, D. (2002) The European Language Portfolio: structure, origins, implementation and challenges. *Language Teaching* 35, 182-189.

Liu, M. (2006) Anxiety in Chinese EFL students at different proficiency levels. *System* 34, 301-316.

Lynch, B. K. (1997) In search of the ethical test. *Language Testing* 14 (3), 315-327.

Lynch, B. K. (2001) Rethinking assessment from a critical perspective. *Language Testing* 18 (4), 351-372.

Lumley, T. y T. F. McNamara (1995) Rater characteristics and rater bias: implications for training. *Language Testing* 12 (1), 54-71.

Lumley, T. y B. O'Sullivan (2005) The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing* 22 (4), 415-437.

Mackey, A. y S. M. Gass (2005) *Second Language Research*. *Methodology and Design*. Mahwah, NJ.: Lawrence Erlbaum Associates.

MacNamara, T. (1997) *Measuring Second Language Performance.* London y New York: Longman.

MacNamara, T. (1998) Policy and social considerations in language assessment. *Annual Review of Applied Linguistics* 18, 304-319.

MacNamara, T. (2001a) Rethinking alternative assessment. *Language Testing* 18 (4), 329-332.

MacNamara, T. (2001b) Language assessment as social practice: challenges for research. *Language Testing* 18 (4), 333-349.

MacNamara, T. (2003) Looking back, looking forward: rethinking Bachman. *Language Testing* 20 (4), 466-473.

Marcos Llinàs, M. (2006) Variables afectivas en el aprendizaje de una lengua. Tesis doctoral sin publicar. Mallorca: Universitat de les Illes Balears.

MEC (1992) *Secundaria Obligatoria. Lenguas extranjeras*. Secretaría de Estado de Educación. ISBN: 84-369-2186-0.

MEC (1993) *Documentos de apoyo a la evaluación. Educación Secundaria*. Dirección General de Renovación Pedagógica. Subdirección General de Ordenación Académica.

Meara, P. (1996). The dimensions of lexical competence. En G. Brown, K Malmkjaer y J. Williams, eds. *Performance and Competence in Second Language Acquisition*, 35-53. Cambridge: Cambridge University Press.

Meara, P. (1997) Towards a new approach to modelling vocabulary acquisition. En Schmitt N. y M. McCarthy, eds. *Vocabulary: Description, acquisition and pedagogy*, 109-121. Cambridge: Cambridge University Press.

Meara, P. (1999) The Vocabulary Knowledge Framework. [Documento de Internet disponible en http://www.swan.ac.uk/cals/calsres/vlibrary/pm96d.htm].

Meara, P. (2002) The rediscovery of vocabulary. *Second Language Research* 18 (4), 393-407.

Meara, P. y T. Fitzpatrick (2000) Lex30: an improved method of assessing productive vocabulary in a L2. *System* 28, 19-39.

Melka, F. (1997) Receptive vs. productive aspects of vocabulary. En Schmitt, N. y M. McCarthy, eds. *Vocabulary: Description, acquisition and pedagogy*, 84-102. Cambridge: Cambridge University Press.

Messick, S. (1989) Validity. En Linn, R. Ed. *Educational Measurement* (3rd edition) 13-103. American Council of Education, Washington: Macmillan.

Messick, S. (1996) Validity and washback in language testing. *Language Testing* 13 (3), 241-256.

Moliner, M. (2000) Diccionario de uso del español. Madrid: Gredos.

Monroy, R. (2000) Paradigmas de investigación y su incidencia en la enseñanza de lenguas extranjeras. En Ruiz, J. *et al*. (Coord.) *Estudios de metodología de la lengua inglesa.* Valladolid: Centro Buendía, Universidad de Valladolid.

Moon, R. (1997) Vocabulary connections: multi-words items in English. En Schmitt N. y M. McCarthy, eds. *Vocabulary: Description, acquisition and pedagogy*, 40-63. Cambridge: Cambridge University Press.

Moya Santoyo, J. (2002) *Historia de la Psicología. Autores más influyentes*. Madrid: PS Editorial.

Murtagh, L. (2003) Retention and Attrition of Irish as a Second Language. [Documento de Internet disponible en http://www.ite.ie/lmurtagh/RAISLeng.htm].

Nattinguer, J. R. y J. S. DeCarrico (1992) *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Nation, I. S. P. y H. Kyongho. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System* 23 (1), 35-41.

Nation, I.S.P. (1995) The word on words: An interview with Paul Nation. Interviewed by N. Schmitt. *The Language Teacher* 19 (2), 5-7.

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

Nation, I.S.P. (1993) Using dictionaries to estimate vocabulary size: essential but rarely followed procedures. *Language Testing* 10 (1), 27-40.

Nation, I. S. P. (2001) *Learning vocabulary in another Language.* Cambridge: Cambridge University Press.

Norton, B. (1998) Accountability in Language Testing. En Corson, D. y C. Clapham, eds. *Language Testing and Assessment*. Vol. 7 of the Encyclopedia of Language and Education, 313-322. Amsterdam: Kluwer Academic Publishers.

Nunan, D. (1992) *Research Methods in Language Learning*. Cambridge: Cambridge University Press.

Oh, J. (1992) The Effects of L2 Reading Assessment Methods on Anxiety Level. *TESOL Quarterly* 26 (1), 172-176.

Oller, J. W. Jr. (1973a) Cloze tests of second language proficiency and what they measure. *Language Learning* 23, 105-18.

Oller, J. W. Jr. (1973b) Discrete-point tests versus tests of integrative skills. En Oller, J. y J. Richards, eds. *Focus on the learner*, 184-199. Rowley, Massachusetts: Newbury House Publishers.

Oller, J. W. Jr. (1979) *Language tests at school*. London & New York: Longman.

Oller, J. W., Jr. (1983) Evidence for a general language proficiency factor: an expectancy grammar. En Oller, J. W., Jr., ed. *Issues in Language Testing Research*, 3-10. Rowley, Massachusetts: Newbury House.

Oller, J. W., Jr. (1995) Adding Abstract to Formal and Content Schemata: Results of Recent Work in Peircean Semiotics. *Applied Linguistics* 16 (3), 273-305.

Oxford, R. L. (1990) *Language Learner Strategies: What every teacher should know*. New York: Newbury House.

Pajares, R. *et al*. (2004) *El proyecto PISA 2000: Aproximación a un modelo de evaluación*. Madrid: Ministerio de Educación, Cultura y Deporte. Secretaría General Técnica. Instituto Nacional de Evaluación y Calidad del Sistema Educativo (INECSE).

Papalia, D. E. y S. Wendkos (1988) *Psicología*. Mexico: Mc. Graw-Hill.

Phakiti, A. (2003) A Closer Look at Gender and Strategy Use in L2 Reading. *Language Learning* 53 (4), 649-702.

Pica, T. (2000) Tradition and transition in English language teaching methodology. *System* 28,1-18.

Pilliner, A. E. G. (1968) Subjective and objective testing. En Davies (1968) *Language Testing Symposium. A Psycholinguistic Perspective*, 19-35. London: Oxford University Press.

Porter, D. (1978) Cloze procedure and equivalence. *Language Learning* 28, 333-41.

Prapphal, K. (1994) A study of the C-test and the X-Test Performed by First-Year Science-Oriented University Students. *PASAA* 24, 16-23.

Prapphal, K. (2006) An investigation of the General and Academic English X-Tests in Measuring Grammatical Competence of Thai Science Students. [Documento de Internet disponible en http://pioneer.chula.ac.th/~pkanchan/html/x-tests.htm].

Qi, L. (2005) Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing* 22 (2), 142-173.

Quirk, R. y S. Greenbaum (1973) *A University Grammar of English*. Longman Group UK Ltd.

Raatz, U. (1984) The factorial validity of C-tests. En Culhane, T. *et al*., eds. *Practice and problems in language testing*. *Occasional Papers* 29, 124-139. Colchester: University of Essex.

Raatz, U. (1985) Better theory for better tests. *Language Testing 2*, 60-75.

Raatz, U. y C. Klein-Braley (2002) Introduction to language testing and to C-Tests. En *University Language Learning and the C-test*. Coleman, J. *et al*., eds. AKS-Verlag, Bochum. [Documento de Internet disponible en http://www.uni-duisburg.de/FB3/ANGLING/FORSCHUNG/HOWTODO.HTM].

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].

Rea, M. P. (1984) Language tests as indicators of academic achievement. En Culhane, T. *et al*., eds. *Practice and problems in language testing*. *Occasional Papers* 29, 140-158. Colchester: University of Essex.

Rea, M. P. (1985) Language Testing and the Communicative Curriculum (1). En *New Directions in Language Testing*, 15-31. Lee *et al*., eds. Oxford: Pergamon Press.

Rea-Dickins, P. (2001) Mirror, mirror on the wall: identifying processes of classroom assessment. *Language Testing* 18 (4), 429-462.

Rea-Dickins, P. (2004) Understanding teachers as agents of assessment. *Language Testing* 21 (3), 249-258.

Read, J. (1993) The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10 (3), 355-371.

Read, J. (1997) Vocabulary and testing. En Schmitt, N. y M. McCarthy, eds. *Vocabulary: Description, acquisition and pedagogy*, 303-320. Cambridge: Cambridge University Press.

Read, J. (2000) *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. y C. A. Chapelle (2001) A framework for second language vocabulary assessment. *Language Testing* 18 (1), 1-32.

Rietveld, T. y R. van Hout. (2005) *Statistics in Language Research: Analysis of Variance*. Berlin: Mouton de Gruyter.

Ruiz, J. M. y otros (2000) *Estudios de metodología de la lengua inglesa.* Valladolid: Centro Buendía, Universidad de Valladolid.

Saif, S. (2006) Aiming for positive washback: a case study of internacional teaching assistants. *Language Testing* 23 (1) 1-34.

Sanz Saiz, I. (1999) El examen de Selectividad a examen. *GRETA Revista para profesores de Inglés* 7 (2), 16-29.

Sanz Sáiz, I. y M. Fernández Álvarez (2005) La validez del examen de Inglés en Selectividad. En Herrera Soler, H. y J. García Laborda, eds. *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*, 149-164. Valencia: Ed. Universidad Politécnica de Valencia

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.

Schmidt, R. (1994) Implicit Learning and the Cognitive Unconscious: Of Artificial Grammars and SLA. En *Implicit and explicit learning of languages*. N.C. Ellis, ed., London: Academic Press.

Schmitt, N. (1997) Vocabulary learning strategies. En Schmitt, N. y M. McCarthy, eds. *Vocabulary: Description, acquisition and pedagogy*, 199-227. Cambridge: Cambridge University Press.

Schmitt, N. (1998a) Quantifying word association responses: what is native-like? *System* 26 (3), 389-401.

Schmitt, N. (1998b) Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning* 48 (2), 281-317.

Schmitt, N. (1999) The relationship between TOEFL vocabulary items and meaning, association, collocation, and word class knowledge. *Language Testing* 16, 189-216.

Schmitt, N. (2000) *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Schmitt, N. y M. McCarthy Eds. (1997) *Vocabulary: Description, acquisition and pedagogy*, 84-102. Cambridge: Cambridge University Press.

Sheerin, P. H. (2000) Some reflections on the university-entrance exam in English. En Ruiz, J. *et al*. (Coord.) *Estudios de metodología de la lengua inglesa.* Valladolid: Centro Buendía, Universidad de Valladolid

Skehan, P. (1989) *Individual Differences in Second-Language Learning*. UK: Edward Arnold.

Skehan, P. (1991) Progress in language testing: the 1990s. En Alderson, J. C. y B. North, eds. *Language testing in the 1990s: the communicative legacy*, 3-21. London: MacMillan.

Shohamy, E. (1983) Interrater and intrarrater reliablility of the oral interview and concurrent validity with cloze procedure in Hebrew. En J. Oller, Jr., ed. *Issues in language testing research*, 229-236. Rowley, Massachusetts: Newbury House.

Shohamy, E. (1984a) Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1, 147-161.

Shohamy, E. (1984b) Input and output in language testing. En Culhane, T. *et al*., eds. *Practice and problems in language testing*. *Occasional Papers* 29, 159-176. University of Essex. Colchester.

Shohamy, E. *et al*. (1996) Test impact revisited: washback effect over time. *Language Testing* 13 (3), 298-317.

Shohamy, E. (1997) Testing methods, testing consequences: are they ethical, are they fair? *Language Testing* 14 (3), 340-349.

Shohamy, E. (2000) The relationship between language testing and second language acquisition, revisited. *System* 28, 541-553.

Shohamy, E. (2001a) *The power of tests*. London: Longman.

Shohamy, E. (2001b) Democratic assessment as an alternative. *Language Testing* 18 (4), 373-391.

Shohamy, E. y T. Reves (1985) Authentic language tests: where from and where to? *Language Testing* 2 (1), 48-59.

Schoonen, R. (2005) Generalizability of writing scores: an application of structural equation modelling. *Language Testing* 22 (1), 1-30.

Sigott, G. y J. Kobrel (1993) Validating the X-Test. *Language Testing Update* 14, 53-58.

Singleton, D. (1999) *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Snellings, P. *et al*. (2004) Validating a test of second language written lexical retrieval: a new measure of fluency in written language production. *Language Testing* 21 (2), 174-201.

Sökmen, A. J. (1997) Current trends in teaching second language vocabulary. En Schmitt N. y M. McCarthy, eds. *Vocabulary: Description, acquisition and pedagogy*, 40-63. Cambridge: Cambridge University Press.

Spence-Brown, R. (2001) The eye of the beholder: autenticity as an embedded assessment task. *Language Testing* 18 (4), 463-481.

Spolsky, B. (1973) What does it mean to know a language; or how do you get someone to perform his competence? En Oller, J. y J. Richards, eds. *Focus on the learner*, 164-176. Rowley Massachusetts: Newbury House Publishers.

Spolsky, B. (1985) The limits of authenticity in language testing. *Language Testing* 2 (1), 31-40.

Spolsky, B. (1990) The prehistory of TOEFL. *Language Testing* 7 (1), 98-115.

Spolsky, B. (1995) *Measured words*. Oxford: Oxford University Press.

Spolsky, B. (1997) The ethics of gatekeeping tests: what have we learnt in a hundred years? *Language Testing* 14 (3), 242-247.

Storey, P. (1997) Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing* 14 (2), 214-231.

Stubbs, J. y Tucker, G. (1974) The cloze test as a measure of English proficiency. *Modern Language Journal* 58, 239-41.

Suau (1998-99) La estrategia de inferencia léxica en textos de economía y empresa: aplicación y automatización por un grupo de estudiantes universitarios españoles. *RESLA* 13, 37-47.

Süssmilch, E. (1984) Language testing with immigrant children. En Culhane, T. *et al*., eds. *Practice and problems in language testing. Occasional Papers* 29, 167-176. Colchester: University of Essex.

Swenson, L. C. (1980) *Teorías del aprendizaje*. Barcelona: Paidós.

Takeuchi, O. (2003) What can we learn from good foreign language learners? A qualitative study in the Japanese foreign language context. *System* 31, 385-392.

Tamburini, F. y Paci, S. (2002) *Web-based language learning: authoring and assessment technologies*. International Conference on Information and Communication Technologies in Education - ICTE2002, 871-877. Badajoz. [Documento de Internet disponible en http://137.204.243.238:8000/Tamburini/ICTE_2002.pdf].

Taylor, L. (2005) Washback and impact. *ELT Journal* 59 (2), 154-155.

Taylor, W. L. (1953) Cloze procedure: A new tool for measuring readability. *Journalism Quarterly* 30, 414-38.

Thatcher, P. (2000) Acquisition and learning –theory matters. *IRAL* 38, 161-174.

Tribble, C. (2000) Designing evaluation into educational change processes. *ELT Journal* 54 (4), 319-327.

Wall, D. *et al*. (1994) Evaluating a placement test. *Language Testing* 11, 321 completar

Wall, D. (1996) Introducing new tests into traditional systems: insights from general education and from innovation theory. *Language Testing* 13 (3), 334- 354.

Wall, D. (2000) The impact of high stakes testing on teaching and learning: can this be predicted or controlled? *System* 28, 499-509.

Waring, R. (1998) Receptive and productive foreign language vocabulary size II. Manuscrito sin publicar. [Documento de Internet disponible en http://www.1.harenet.ne.jp/-waring/vocabindex.html].

Watts, F. y A. García Carbonell (2005) Control de calidad en la calificación de la prueba de lengua inglesa de Selectividad. En Herrera Soler, H. y J. García Laborda, eds. *Estudios y criterios para una Selectividad de calidad en el examen de Inglés*, 99-119. Valencia: Ed. Universidad Politécnica de Valencia.

Weigle, S. C. (1994) Effects of training on raters of ESL compositions. Language Testing 11 (2) 197-223.

Weir, C. (1988) *Communicative Language Testing*. Exeter Linguistic Studies Vol. 11. Exeter: University of Exeter.

Weir, C. J. (1990) *Communicative Language Testing*. Hemel Hempstead: Prentice Hall.

Weir, C. y Roberts, J (1994) *Evaluation in ELT*. Oxford: Blackwell.

Widdowson, H. G. (2000) On the Limitations of Linguistics Applied. *Applied Linguistics* 21 (1), 3-25.

Wolter, B. (2002) Assessing proficiency through word associations: is there still hope? *System* 30, 315-329.

Wood, R. (1991) *Assessment and Testing. A Survey or Research*. Cambridge: Cambridge University Press.

Wragg, E. C. (2003) *Evaluación y aprendizaje en la escuela secundaria*. Barcelona: Paidós.

Yamashita, J. (2003) Processes of taking a gap-filling test: comparisons of skilled and less skilled EFL readers. *Language Testing* 20 (3), 267-293.

Zimmerman, C. B. (1997) Historical trends in second language vocabulary instruction. En Coady, J. y T. Huckin, eds. *Second language vocabulary acquisition*, 5-19. Cambridge: Cambridge University Press.

BASES DE DATOS CONSULTADAS

LLBA (Language and Linguistic Behaviour Abstracts)

ERIC (Educational Research International Centre)

PROGRAMAS INFORMÁTICOS

SPSS (Statistical Package for the Social Sciences) 8.1, 9.1 y 12.5 for Windows

Apéndice 1. Modelo para la realización de C-tests.

DIRECTIONS FOR TAKING C-TESTS

At the bottom of this page is a sample of a new kind of test.
The test is made by copying several varied short texts and deleting the second half of every second word, beginning with word two in sentence two.

Your job will be to restore the texts by replacing the missing elements. Only entirely correct restorations are counted as correct.

SAMPLE C-TEST

BREAKFAST AROUND THE WORLD

Breakfast is an important meal because it gives you energy to start the day. When (1) y_____ do (2) n_____ have a (3)go_____ breakfast, (4)y_____ feel (5)hun_____ and (6)e_____ cakes, (7)bisc____ or (8)swe_____ before lunchtime.

ANSWERS

1. you
2. not
3. good
4. you
5. hungry
6. eat
7. biscuits
8. sweets

Apéndice 2. Estudio piloto I

C-TEST 1

NAME
DATE

Complete the following text filling in the blanks with the appropriate letters. Every dash corresponds to a single letter.

BREAKFAST AROUND THE WORLD

Breakfast is an important meal because it gives you energy to start the day. When (1) y-- do (2) n-- have (3) - good (4) break----, you (5) f--- hungry (6) a-- eat (7) ca---, biscuits (8) o- sweets (9) be---- lunchtime. (10) Th-- type (11) o- food (12) i- bad (13) f-- you (14) bec---- it (15) i- not (16) v--- nutritious (17) a-- has (18) l--- of (19) su--- and (20) f--.
Breakfast (21) i- not (22) t-- same (23) i- every (24) coun---.For(25) ex-----, many British (26) peo--- have (27) to--- or (28) ce---- and (29) - cup (30) o- tea. (31) Ot---- prefer (32) - traditional (33) break---- of (34) ba--- and (35) eg--. In (36) ot--- Northern European (37) coun-----, for (38) ex----- Germany (39) a-- Sweden, (40) peo--- eat (41) c--- meat (42) a-- cheese (43) w--- bread (44) a-- coffee. (45) I- Nigeria (46) h-- soup (47) i- very (48) co----. Many Brazilians (49) e-- different (50) trop---- fruit and cold meat for breakfast.
However, in many parts of the world, people only eat a small dish of rice for breakfast.

1.y	11.o	21.i	31.ot	41.c
2.n	12.i	22.t	32.	42.a
3.	13.f	23.i	33.break	43.w
4.break	14.bec	24.coun	34.ba	44.a
5.f	15.i	25.ex	35.eg	45.i
6.a	16.v	26.peo	36.ot	46.h
7.ca	17.a	27.to	37.coun	47.i
8.o	18.l	28.ce	38.ex	48.co
9.be	19.su	29.	39.a	49.e
10.th	20.f	30.o	40.peo	50.trop

C-TEST 2

NAME
DATE

Complete the following text filling in the blanks with the appropriate letters. Every dash corresponds to a single letter.

MEALS IN BRITAIN

A traditional English breakfast is a very big meal, sausages, bacon, eggs, tomatoes, mushrooms...But (1) now----- many (2) peo--- just (3) h--- cereal (4) w--- milk (5) a-- sugar, (6) o- toast (7) w--- marmalade, (8) j-- or (9) ho---. Marmalade (10) a-- jam (11) a-- not (12) t-- same! (13) Marm----- is (14) ma-- from (15) oran--- and (16) j-- is (17) ma-- from (18) o---- fruit. (19) T-- traditional (20) break---- drink (21) i- tea, (22) wh--- people (23) h--- with (24) c--- milk. (25) So-- people (26) h--- coffee, (27) of--- instant (28) co----, which (29) i- made (30) w--- just (31) h-- water. (32) Ma-- visitors (33) t- Britain (34) f--- this (35) co---- disgusting!
(36) F-- many (37) peo--- lunch (38) i- a (39) qu--- meal. (40) I- cities (41) th--- are (42) - lot (43) o- sandwich (44) b---, where (45) off--- workers (46) c-- choose (47) t-- kind (48) o- bread (49) th-- want, (50) br--- , white or roll, and then all sorts of salad and meat or fish to go in the sandwich. Pubs often serve good, cheap food, both hot and cold. School children can have a hot meal at school, but many just take a snack from home, a sandwich, a drink, some fruit and perhaps some crisps.

1.now	11.a	21.i	31.h	41.th
2.peo	12.t	22.wh	32.ma	42.
3.h	13.marm	23.h	33.t	43.o
4.w	14.ma	24.c	34.f	44.b
5.a	15.oran	25.so	35.co	45.off
6.o	16.j	26.h	36.f	46.c
7.w	17.ma	27.of	37.peo	47.t
8.j	18.o	28.co	38.i	48.o
9.ho	19.t	29.i	39.qu	49.th
10.a	20.break	30.w	40.i	50.br

C-TEST 3

NAME
DATE

Complete the following text filling in the blanks with the appropriate letters. Every dash corresponds to a single letter.

PROFILE OF A GENERATION

Something is going on inside them. They are like volcanoes about to erupt, with a passion only youth can feel. Searching for self expression whilst enjoying the hotel they call home, life is not so bad for these well-fed citizens of tomorrow.
They (1) a-- full (2) o- the (3)j--- of (4) li--. They (5) lo-- their (6) pa----- but (7) a-- not (8) su-- what (9) th-- country (10) ha- to (11) of--- them (12) i- the (13) fut---.
Arguments (14) w--- parents: 8 (15) o-- of 10 (16) r-- with (17) pa----- about (18) t-- time (19) th-- have (20) t- be (21) ho-- at (22) n----. 7 out (23) o- 10 argue (24) w--- parents (25) ab--- not (26) hel---- around (27) t-- house. 6 (28) o-- of 10 (29) g---- complain (30) th-- their (31) bro----- do (32) n-- have (33) t- do (34) - thing, (35) wh--- they (36) a-- expected (37) t- do (38) house----. “I (39) g-- away (40) w--- everything (41) a- home, (42) s-- (my sister) (43) d--- not. (44) - am (45) n-- a (46) chauv-----. It (47) i- just (48) n-- my (49) ro--.” 5 out (50) o- 10 row with parents about their studies. Parents are loved, ...with the knowledge that they are only human, a difficult fact for some teenagers to face.

1.a	11.of	21.ho	31.bro	41.a
2.o	12.i	22.n	32.n	42.s
3.j	13.fut	23.o	33.t	43.d
4.li	14.w	24.w	34.	44.
5.lo	15.o	25.ab	35.wh	45.n
6.pa	16.r	26.hel	36.a	46.chauv
7.a	17.pa	27.t	37.t	47.i
8.su	18.t	28.o	38.house	48.n
9.th	19.th	29.g	39.g	49.ro
10.ha	20.t	30.th	40.w	50.o

C-TEST 4

NAME
DATE

Complete the following text filling in the blanks with the appropriate letters. Every dash corresponds to a single letter.

ARGUMENTS

Read what an adult comments about the beliefs, ambitions and feelings of typical teenagers in Spain.

Teenagers (1) ar--- with (2) th--- parents (3) f-- lots (4) o- different (5) rea----. Some (6) ar--- about (7) clo----, others (8) ab--- jobs (9) i- the (10) ho---, and (11) ot---- about (12) th--- friends. (13) - remember (14) wh-- I (15) w-- young (16) - argued (17) ab--- everything. (18) Wh-- my (19) m-- asked, (20) “H--- you (21) tid--- your (22)r---?” I (23) al---- replied, (24) “- am (25) go--- to (26) ti-- it (27) tom-----!” Of (28) co----, tomorrow (29) ne--- came (30) a-- we (31) h-- the (32) inev------ argument. (33) N-- I (34) h--- children (35) a-- they (36) h--- to (37) d- things (38) ar---- the (39) ho---, usually (40) tom-----! Sometimes (41) w- argue (42) a-- then (43) - remember (44) t-- arguments (45) - had (46) w--- my (47) mo---- and (48) sm--- about (49) h-- little (50) th---- have changed.

It helps me to understand them.

1.ar	11.ot	21.tid	31.h	41.w
2.th	12.th	22.r	32.inev	42.a
3.f	13.	23.al	33.n	43.
4.o	14.wh	24.	34.h	44.t
5.rea	15.w	25.go	35.a	45.
6.ar	16.	26.ti	36.h	46.w
7.clo	17.ab	27.tom	37.d	47.mo
8.ab	18.wh	28.co	38.ar	48.sm
9.i	19.m	29.ne	39.ho	49.h
10.ho	20.h	30.a	40.tom	50.th

Apéndice 3. Estudio piloto II

C-TEST 1 NAME
DATE

Now, complete the following texts filling in the blanks with the appropriate letters.

1. LEARN TO COMMUNICATE

To be fluent in several languages is no longer considered a rare talent, but a necessity to succeed and communicate in the world in which we now live. Many (1)peo_____ believe (2)th_____ once (3)y_____ are (4)pa_____ childhood, (5)lear_____ a (6)n_____ language (7)i_____ too (8)diff_____. This (9)i_____ not (10)tr_____.
Whether (11)y_____ want (12)t_____ learn English, French, Spanish (13)o_____ Polish there (14)a_____ schools (15)a_____ courses (16)gea_____ for (17)yo_____ needs (18)a_____ specifically (19)ai_____ at (20)ad_____ learning. (21)Ad_____ learning (22)i_____ pro-active; (23)y_____ are (24)invo_____ with (25)t_____ language from the beginning and encouraged to talk, whatever your ability. There are a variety of methods available.

ANSWERS

1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

2. THE HISTORIC VOYAGE OF CHISTOPHER COLUMBUS

In 1992, more than thirty countries celebrated the 500th anniversary of the world´s most famous transatlantic voyage.
Christopher Columbus is (26)cred_____ with (27)“disco_____” the New World (28)o_____ that (29)hist_____ trip (30)i_____ 1492. In (31)fa_____, of (32)cou_____, some 20 million (33)nat_____ people (34)we_____ already (35)th_____ before (36)h_____ stepped (37)ash_____. Many (38)histo_____ also (39)cl_____ that (40)t_____ Vikings saw (41)i_____ first. (42)B_____ it (43)w_____ Columbus who (44)ma_____ Europe aware (45)o_____ the (46)exis_____ of (47)t_____ vast American (48)cont_____ and (49)w_____ started (50)t_____ adventure which has never stopped since: the exploration, conquest and settlement of this newfound land.

ANSWERS

26	31	36	41	46
27	32	37	42	47
28	33	38	43	48
29	34	39	44	49
30	35	40	45	50

3. COPING WITH ADDICTION

Alcoholics Anonymous (AA), founded 60 years ago, is increasingly familiar to the general public as a network where ex-drinkers get together for the “talk therapy” that helps them to cope with their drink problem.
Show (51)busi_____ stars (52)ha_____ contributed (53)t_____ this (54)famil_____ by (55)brea_____ their (56)anon_____ and (57)refe_____ publicly (58)t_____ the (59)w_____ AA has (60)hel_____ them. (61)Tal_____ on Radio Four (62)rece_____, the (63)ac_____ Sir Anthony Hopkins confessed (64)th_____ he (65)h_____ been a (66)to_____ mess (67)bef_____ recovering (68)h_____ sobriety (69)thr_____ AA. Elton John is (70)ano_____ one (71)a_____ there (72)a_____ some AA (73)meet_____ in London (74)kn_____ to (75)b_____ startlingly glamorous and packed with celebrities. But AA is not looking for money from anyone. And it is not, curiously, looking for publicity either.

ANSWERS

51	56	61	66	71
52	57	62	67	72
53	58	63	68	73
54	59	64	69	74
55	60	65	70	75

4. THE NEW GENERATION OF FARMYARD CLONES

“I MAKE all my sheep here”. Bill Ritchie gestured (76)tow_____ an (77)ann_____ of (78)h_____ laboratory (79)wh_____ he (80)h_____ used a (81)sp_____ of (82)elect_____ to (83)viv_____ two (84)ce_____ that (85)gr_____ into Morag (86)a_____ Megan, the (87)sis_____ who (88)ma_____ front-page (89)head_____ earlier (90)th_____ year (91)f_____ being (92)“man-_____”.
As (93)t_____ first (94)fr_____ of a (95)tech_____ that (96)c_____ make (97)mill_____ of (98)iden_____ sheep, (99)th_____ innocent Welsh (100)moun_____ sheep sparked a major controversy about the rights and wrongs of such research, which is currently extended to create supersheep and cloned cattle.

ANSWERS

76	81	86	91	96
77	82	87	92	97
78	83	88	93	98
79	84	89	94	99
80	85	90	95	100

Apéndice 4. Instrucciones para las profesoras de Inglés de los grupos participantes en el estudio

Esta experiencia forma parte de una tesis doctoral que estudia un tipo nuevo de ***Cloze test***, el **C-test**, y cómo correlaciona con otros tipos de examen, en concreto con las pruebas de Selectividad y con el progreso del alumno en la asignatura.

No se va a estudiar el nivel de los alumnos en los distintos IES, sino la relación entre los resultados de las distintas pruebas de cada alumno (un examen del tipo de los de Selectividad y un C-test). No obstante, vamos a aplicar estos exámenes en grupos de 2º Bachillerato de diferentes IES de Madrid, para que la muestra sea variada y representativa.

Te ruego que sigas los pasos que se indican a continuación:

1. **Aplica la prueba de Selectividad *Cavemen?*** (Sept.99-LOGSE) en una sesión normal de clase de 2º Bach.

2. **Aplica el C-test** en una sesión de clase. Puedes avisarles el día anterior de que van a hacer un tipo de examen nuevo y diferente, **deben completar un texto en el que falta la 2ª mitad de cada 2ª palabra**. El test tiene 100 omisiones o huecos. En los dos primeros textos, cada guión corresponde a una letra, en los dos siguientes no se les da esa "pista". Sólo se considera correcta una respuesta si se recupera la palabra exacta (les das previamente la hoja de instrucciones que explica el procedimiento). Es una prueba experimental, y se les comunicarán sus resultados.

   Los textos a partir de los cuales se construye el C-test pertenecen a distintas pruebas de Selectividad de otros años. Como verás, la dificultad de los ítems varía mucho, sabemos que algunos son muy difíciles de recuperar.

   Hay dos versiones del C-test, **A y B**, debes repartirlas de forma **aleatoria** (pares e impares) en la clase, para que la mitad de cada grupo haga un modelo, cada uno el que le toque.

3. Una vez aplicado el C-test en clase, debes administrar el **cuestionario** para que expresen de forma anónima su **opinión** sobre la prueba. No es necesario que pongan su nombre, sólo que contesten con sinceridad.

Sí tienen que poner el nombre en las otras pruebas, porque hay que analizar mediante un programa estadístico todos los datos de cada alumno concreto, en ningún caso para darles una calificación. Los nombres y apellidos sirven sólo para identificarlos durante la investigación. Después, en todo momento se mantendrá el anonimato y se les asignará un número identificativo.

4. Anota la **calificación** de cada alumno en Inglés en la **2ª evaluación** de este curso.

5. Te ruego que nos remitas las calificaciones que obtengan en **Selectividad en Inglés** en Junio los alumnos que se presenten.

Por tanto, necesitamos:

- los exámenes modelo de Selectividad *Cavemen?*
- los C-tests
- los cuestionarios
- las calificaciones de la 2ª Evaluación en Inglés
- la calificación de Inglés en las PAAU de junio (de los alumnos presentados)

MUCHAS GRACIAS POR TU COLABORACIÓN.

Apéndice 5. C-TESTS PERSPECTIVA EMPÍRICA TESIS

C-TEST A NAME SEX Male / Female
DATE
SCHOOL

First of all, read each text carefully trying to understand its meaning.
Remember that the **second half of every second word has been deleted**, beginning with word two in sentence two.
Then, complete the texts filling in the blanks with the appropiate letters
In the first two texts each dash corresponds to a single letter.

1. ROAD ACCIDENTS

Your chances of dying in a road accident double in France. For (1)ev_ _ _ million (2)vehi_ _ _ _, 300 (3)peo_ _ _ die (4)o_ the French (5)ro_ _ _ every (6)ye_ _, compared (7)t_ 140 (8)i_ Britain. (9)T_ _ carnage (10)h_ _ been (11)mu_ _ reduced (12)i_ the (13)pa_ _ two (14)dec_ _ _ _, but 8,000 (15)dea_ _ _ a (16)ye_ _ is (17)st_ _ _ an (18)aw_ _ _ lot (19)o_ grief (20)a_ _ suffering -(21)t_ _ equivalent (22)o_ fifty (23)la_ _ _ plane (24)cra_ _ _ _.

The French (25)gover_ _ _ _ _ has tough plans to halve the number of road deaths in five years. Their programme includes measures to discourage speeding, which is responsible for almost half the deaths of French roads.

ANSWERS:

1 ev_ _ _	6 ye_ _	11 mu_ _	16 ye_ _	21 t_ _
2 vehi_ _ _ _	7 t_	12 i_	17 st_ _ _	22 o_
3 peo_ _ _	8 i_	13 pa_ _	18 aw_ _ _	23 la_ _ _
4 o_	9 T_ _	14 dec_ _ _ _	19 o_	24 cra_ _ _ _
5 ro_ _ _	10 h_ _	15 dea_ _ _	20 a_ _	25 gover_ _ _ _ _

2. EVOLUTION

In the classification of animals there is an order called Primates. In (1)th_ _ _ appearance (2)t_ _ primates (3)rese_ _ _ _ the (4)hu_ _ _ being (5)mo_ _ than (6)a_ _ other (7)ani_ _ _ _ do. (8)I_ is (9)nat_ _ _ _ to (10)ded_ _ _ that (11)th_ _ are (12)mo_ _ closely (13)rel_ _ _ _ to (14)hu_ _ _ beings (15)th_ _ other (16)ani_ _ _ _ are. (17)I_ fact, (18)t_ _ human (19)be_ _ _ must (20)b_ included (21)a_ a (22)pri_ _ _ _, if (23)a_ _ sense (24)a_ all (25)i_ to be made of animal classification.

Once evolution is accepted, one must come to the inevitable conclusion that the various primates, including the human being, have developed from some single ancestral stem and that all are to varying degrees cousins, so to speak.

ANSWERS:

1 th_ _ _	6 a_ _	11 th_ _	16 ani_ _ _ _	21 a_
2 t_ _	7 ani_ _ _ _	12 mo_ _	17 I_	22 pri_ _ _ _
3 rese_ _ _ _	8 I_	13 rel_ _ _ _	18 t_ _	23 a_ _
4 hu_ _ _	9 nat_ _ _ _	14 hu_ _ _	19 be_ _ _	24 a_
5 mo_ _	10 ded_ _ _	15 th_ _	20 b_	25 i_

3. AMERICAN IMPERIALISM

Global leadership is both the price America pays and the benefit Americans derive from our wealth, our size and our strength. Worldwide (1)hun_____ and (2)pov_____, the (3)spr_____ of (4)nuc_____ weapons (5)a_____ the (6)prolif_____ of (7)vio_____ conflicts (8)ha_____ us (9)a_____. Economic (10)devel_____, better (11)educ_____, better (12)con_____ of (13)ar_____, cooperation (14)am_____ nations (15)a_____ the (16)peac_____ resolution (17)o_____ conflicts (18)he_____ Americans (19)en_____ greater (20)prosp_____ and (21)pe_____. Positive (22)a_____ active (23)engag_____ in (24)wo_____ affairs (25)i_____ the smart as well as the right thing for the United States to do. Yet among developed nations, America has become the least generous provider of either development aid or troops for peacemaking. Our virtue is fading.

ANSWERS:

1 hun	6 prolif	11 educ	16 peac	21 pe
2 pov	7 vio	12 con	17 o	22 a
3 spr	8 ha	13 ar	18 he	23 engag
4 nuc	9 a	14 am	19 en	24 wo
5 a	10 devel	15 a	20 prosp	25 i

4. WOMEN DOCTORS. ARE THEY DIFFERENT?

In 1974, 11% of the students graduating from medical school in the U.S. were female. In 1984 (1)th_____ proportion (2)w_____ close (3)t_____ 30%.(4)Wh_____ the (5)majo_____ of (6)doc_____ in (7)t_____ United States (8)a_____ still (9)ma_____, that (10)pat_____ is (11)chan_____ with (12)t_____ new (13)wa_____ of (14)med_____ students. (15)B_____ will (16)th_____ fact (17)ma_____ any (18)diffe_____ to (19)t____ medical (20)profe_____? Are (21)fem_____ doctors (22)rea_____ different (23)fr_____ male (24)doc_____?

Traditionally (25)nur_____ did the “female” jobs, such as taking care of patients’ basic physical needs and helping people and their families to face illness. Meanwhile, doctors did the aggressive part.

ANSWERS:

1 th	6 doc	11 chan	16 th	21 fem
2 w	7 t	12 t	17 ma	22 rea
3 t	8 a	13 wa	18 diffe	23 fr
4 Wh	9 ma	14 med	19 t	24 doc
5 majo	10 pat	15 B	20 profe	25 nur

C-TEST B NAME SEX Male / Female
DATE
SCHOOL

First of all, read each text carefully trying to understand its meaning. Then, complete the texts filling in the blanks with the appropiate letters
Remember that the **second half of every second word has been deleted**, beginning with word two in sentence two. In the first two texts each dash corresponds to a single letter.

1. AMERICAN IMPERIALISM

Global leadership is both the price America pays and the benefit Americans derive from our wealth, our size and our strength. Worldwide (1)hun_ _ _ and (2)pov_ _ _ _, the (3)spr_ _ _ of (4)nuc_ _ _ _ weapons (5)a_ _ the (6)prolif_ _ _ _ _ _ _ of (7)vio_ _ _ _ conflicts (8)ha_ _ us (9)a_ _. Economic (10)devel_ _ _ _ _ _, better (11)educ_ _ _ _ _, better (12)con_ _ _ _ of (13)ar_ _, cooperation (14)am_ _ _ nations (15)a_ _ the (16)peac_ _ _ _ resolution (17)o_ conflicts (18)he_ _ Americans (19)en_ _ _ greater (20)prosp_ _ _ _ _ and (21)pe_ _ _. Positive (22)a_ _ active (23)engag_ _ _ _ _ in (24)wo_ _ _ affairs (25)i_ the smart as well as the right thing for the United States to do. Yet among developed nations, America has become the least generous provider of either development aid or troops for peacemaking. Our virtue is fading.

ANSWERS:

1 hun_ _ _	6 prolif_ _ _ _ _ _ _	11 educ_ _ _ _ _	16 peac_ _ _ _	21 pe_ _ _
2 pov_ _ _ _	7 vio_ _ _ _	12 con_ _ _ _	17 o_	22 a_ _
3 spr_ _ _	8 ha_ _	13 ar_ _	18 he_ _	23 engag_ _ _ _
4 nuc_ _ _ _	9 a_ _	14 am_ _ _	19 en_ _ _	24 wo_ _ _
5 a_ _	10 devel_ _ _ _ _ _	15 a_ _	20 prosp_ _ _ _ _	25 i_

2. WOMEN DOCTORS. ARE THEY DIFFERENT?

In 1974, 11% of the students graduating from medical school in the U.S. were female. In 1984 (1)th_ _ proportion (2)w_ _ close (3)t_ 30%. (4)Wh_ _ _ the (5)majo_ _ _ _ of (6)doc_ _ _ _ in (7)t_ _ United States (8)a_ _ still (9)ma_ _, that (10)pat _ _ _ _ is (11)chan_ _ _ _ with (12)t_ _ new (13)wa_ _ of (14)med_ _ _ _ students. (15)B_ _ will (16)th_ _ fact (17)ma_ _ any (18)diffe_ _ _ _ _ to (19)t_ _ medical (20)profe_ _ _ _ _? Are (21)fem_ _ _ doctors (22)rea_ _ _ different (23)fr_ _ male (24)doc_ _ _ _ ? Traditionally (25)nur_ _ _ did the “female” jobs, such as taking care of patients’ basic physical needs and helping people and their families to face illness. Meanwhile, doctors did the aggressive part.

ANSWERS:

1 th_ _	6 doc_ _ _ _	11 chan_ _ _ _	16 th_ _	21 fem_ _ _
2 w_ _	7 t_ _	12 t_ _	17 ma_ _	22 rea_ _ _
3 t_	8 a_ _	13 wa_ _	18 diffe_ _ _ _ _	23 fr_ _
4 Wh_ _ _	9 ma_ _	14 med_ _ _ _	19 t_ _	24 doc_ _ _ _
5 majo_ _ _ _	10 pat_ _ _ _	15 B_ _	20 profe_ _ _ _ _	25 nur_ _ _

3. ROAD ACCIDENTS

Your chances of dying in a road accident double in France. For (1)ev_____ million (2)vehi_____, 300 (3)peo_____ die (4)o____ the French (5)ro_____ every (6)ye____, compared (7)t____ 140 (8)i___ Britain. (9)T____ carnage (10)h___ been (11)mu___ reduced (12)i___ the (13)pa___ two (14)dec_____, but 8,000 (15)dea____ a (16)ye____ is (17)st____ an (18)aw____ lot (19)o___ grief (20)a____ suffering -(21)t___ equivalent (22)o____ fifty (23)la_____ plane (24)cra_____.

The French (25)gover_____ has tough plans to halve the number of road deaths in five years. Their programme includes measures to discourage speeding, which is responsible for almost half the deaths of French roads.

ANSWERS:

1 ev	6 ye	11 mu	16 ye	21 t
2 vehi	7 t	12 i	17 st	22 o
3 peo	8 i	13 pa	18 aw	23 la
4 o	9 T	14 dec	19 o	24 cra
5 ro	10 h	15 dea	20 a	25 gover

4. EVOLUTION

In the classification of animals there is an order called Primates. In (1)th___ appearance (2)t____ primates (3)rese_____ the (4)hu_____ being (5)mo____ than (6)a____ other (7)ani_____ do. (8)I___ is (9)nat_____ to (10)ded____ that (11)th____ are (12)mo____ closely (13)rel_____ to (14)hu_____ beings (15)th____ other (16)ani_____ are. (17)I___ fact, (18)t____ human (19)be____ must (20)b___ included (21)a___ a (22)pri_____, if (23)a___ sense (24)a___ all (25)i___ to be made of animal classification.

Once evolution is accepted, one must come to the inevitable conclusion that the various primates, including the human being, have developed from some single ancestral stem and that all are to varying degrees cousins, so to speak.

ANSWERS:

1 th	6 a	11 th	16 ani	21 a
2 t	7 ani	12 mo	17 I	22 pri
3 rese	8 I	13 rel	18 t	23 a
4 hu	9 nat	14 hu	19 be	24 a
5 mo	10 ded	15 th	20 b	25 i

Apéndice 6. Textos sobre los que se diseñó el C-TEST

ROAD ACCIDENTS

Your chances of dying in a road accident double in France. For every million vehicles, 300 people die on the French roads every year, compared to 140 in Britain. The carnage has been much reduced in the past two decades, but 8,000 deaths a year is still an awful lot of grief and suffering –the equivalent of 50 large plane crashes.The French government has tough plans to halve the number of road deaths in five years. Their programme includes measures to discourage speeding, which is responsible for almost half the deaths of French roads.

EVOLUTION

In the classification of animals there is an order called Primates. In their appearance the primates resemble the human being more than any other animals do. It is natural to deduce that they are more closely related to human beings than other animals are. In fact, the human being must be included as a primate, if any sense at all is to be made of animal classification.

Once evolution is accepted, one must come to the inevitable conclusion that the various primates, including the human being, have developed from some single ancestral stem and that all are to varying degrees cousins, so to speak.

AMERICAN IMPERIALISM

Global leadership is both the price America pays and the benefit Americans derive from our wealth, our size and our strength. Worldwide hunger and poverty, the spread of nuclear weapons and the proliferation of violent conflicts harm us all. Economic development, better education, better control of arms, cooperation among nations and the peaceful resolution of conflicts help Americans enjoy greater prosperity and peace. Positive and active engagement in world affairs is the smart as well as the right thing for the United States to do. Yet among developed nations, America has become the least generous provider of either development aid or troops for peacemaking. Our virtue is fading.

WOMEN DOCTORS. ARE THEY DIFFERENT?

In 1974, 11% of the students graduating from medical school in the U.S. were female. In 1984 that proportion was close to 30%.While the majority of doctors in the United States are still male, that pattern is changing with the new wave of medical students.But will this fact make any difference to the medical profession? Are female doctors really different from male doctors?

Traditionally nurses did the “female” jobs, such as taking care of patients’ basic physical needs and helping people and their families to face illness. Meanwhile, doctors did the aggressive part.

Apéndice 7. Cuestionario retrospectivo

DATOS PERSONALES

Edad ..... Sexo V / M
IES.....
He aprendido Inglés:

- □ - en el colegio y después en el Instituto
- □ - además del colegio he asistido a clases en academias u otras instituciones de enseñanza de idiomas en alguna ocasión.
- □ - he asistido a cursos en países de habla inglesa

CUESTIONARIO SOBRE EL C-TEST

Queremos saber tu opinión sobre el test que acabas de realizar.
Es un nuevo tipo de examen que pretende medir tu competencia global en lengua inglesa.

1. ¿Has encontrado dificultades para realizarlo? ¿de qué tipo?
.....

2. Puntua del 1 al 5 el grado en que este examen mide los distintos aspectos de la lengua
Rodea con un círculo (1=mínimo, 5=máximo)

□ aspectos gramaticales	1	2	3	4	5
□ ortografía: spelling	1	2	3	4	5
□ conocimiento general de la lengua	1	2	3	4	5
□ fluidez	1	2	3	4	5
□ léxico: vocabulario	1	2	3	4	5

3. ¿Te parece un examen adecuado? 1 2 3 4 5
(1=nada adecuado, 5=muy adecuado)

4. ¿Te parece un examen completo? 1 2 3 4 5
(1=nada completo, 5=muy completo)

3. ¿Crees que reflejará bien tus conocimientos de Inglés?
(1=mal, 5= muy bien) 1 2 3 4 5

6. ¿Te gustaría que tu acceso a la Universidad dependiera de un test como éste?
Sí
No

7. ¿Y que formara parte de la prueba de Inglés de Selectividad?
Sí
No

Muchas gracias por tu colaboración.

Apéndice 8. *Cavemen?*

NAME
DATE SCHOOL

Lea **todo el texto** cuidadosamente.
Lea atentamente **todas las preguntas** de la prueba.
Proceda a responder en **lengua inglesa** a las preguntas.
Calificación: La puntuación máxima de la prueba es de 10 puntos.

CAVEMEN?

The head of the Philippine agency for national minorities, Manuel Elizalde, announced to the world in 1971 that a tribe of Stone Age people, never exposed to the modern civilization, had been discovered in the jungle.

Wearing loincloths made of orchid leaves, the Tasaday, as they were called, lived in caves, subsisting on insects, small aquatic life and wild fruits and vegetables. They did not farm and had no method for keeping time. They used no weapons and had no word for war.

The news excited scientists and journalists. A platform was built in the rain forest to help helicopters ferry observers in and out. The cavemen became media darlings. *National Geographic* devoted a cover story to the Tasaday, and NBC television offered Elizalde $50,000 to let them produce a documentary on the cave men.

It wasn't until 1986, when the Marcos regime fell, that a Swiss journalist revisited the mysterious people. He was shocked to find the cave dwellers living in huts, dressed in T-shirts and shorts. The journalist reported that they told him they had been instructed by Elizalde to pretend to be cavemen.

QUESTIONS

1. Are the following sentences TRUE or FALSE? Copy the evidence from the text.
No marks are given for only TRUE or FALSE:

a. The Tasaday's diet was supposedly based only on meat.
b. Everything about such cavemen seemed to have been an invention.

(Puntuación máxima: 2 puntos)

2. In your own words and based on the ideas from the text, answer the following questions:

a. How did the world react when the existence of these cavemen was made known?
b. What were the main characteristics of the Tasadays' way of life?.

(Puntuación máxima: 2 puntos)

3. Find the words or phrases in the text that mean:

a) not cultivated (paragraph 2)
b) information (paragraph 3)
c) surprised (paragraph 4)
d) to simulate (paragraph 4)

(Puntuación máxima: 1 punto)

4. Complete the following sentences. Use the appropriate form of the word in brackets when given.

a. Savages are people----------way of life is completely different--------that in industrialized societies.
b. The Tasaday---------to wear---------particular type of clothing.
c. When the Europeans first arrived----------the New World, it was a continent---------- (inhabit) by numerous tribes.
d. The Philippines was a Spanish colony---------1898 when it------(hand) over to the USA.

(Puntuación máxima: 2 puntos)

5. Write about 80 to 100 words on one of the following topics:

a. Have you ever read a book or seen a film about primitive societies?. Write about it.
b. Main contrasts between life in civilized and primitive societies.

(Puntuación máxima: 3 puntos)

Apéndice 9. Normas para la corrección de la prueba de Inglés de la PAAU

	UNIVERSIDAD COMPLUTENSE DE MADRID PRUEBAS DE ACCESO A LOS ESTUDIOS UNIVERSITARIOS DE LOS ALUMNOS DE BACHILLERATO LOGSE AÑO 2001 **MATERIA: INGLÉS**	EXAMENES

CRITERIOS ESPECÍFICOS DE CORRECCIÓN

La prueba consistirá en "el análisis de un texto de un idioma extranjero (el inglés en este caso), del lenguaje común, no especializado. A partir del texto propuesto el estudiante realizará un comentario personal y responderá a cuestiones relacionadas con el texto, que serán planteadas y respondidas por escrito en el mismo idioma, sin ayuda de diccionario ni de ningún otro manual didáctico"(BOE no.257). El **texto** contendrá **alrededor de 250 palabras** y su comprensión no exigirá conocimientos especializados ajenos a la materia de la prueba. La dificultad del texto estará controlada, a fin de permitir al alumno que realice la misma en el tiempo previsto. La **puntuación total** del examen será de **10 puntos.** Al comienzo de la prueba se incluirán unas instrucciones generales para la realización de la misma, en lengua castellana. El resto de la prueba estará totalmente redactada en inglés, y **el alumno usará exclusivamente la lengua inglesa en sus respuestas.**

Valoración y objetivos de cada una de las preguntas:

Pregunta 1: Hasta 2 puntos. Se trata de medir exclusivamente **la comprensión lectora**. El alumno deberá decidir si dos frases que se le presentan son verdaderas o falsas, copiando a continuación únicamente el fragmento del texto que justifica su elección. Se otorgará 1 punto por cada apartado. Se calificará con 0 puntos la opción elegida que no vaya justificada.

Pregunta 2: Hasta 2 puntos. Se pretende comprobar dos destrezas: **la comprensión lectora y la expresión escrita,** mediante la formulación de dos preguntas abiertas, que el alumno deberá contestar basándose en la información del texto, pero utilizando sus propias palabras en la respuesta. Cada una de las preguntas valdrá un punto, asignándose 0,5 puntos a la comprensión de la pregunta y 0,5 puntos a la corrección gramatical de la respuesta.

Pregunta 3: Hasta 1 punto. Esta pregunta trata de medir **el dominio del vocabulario** en el aspecto de la comprensión. El alumno demostrará esta capacidad, localizando en el párrafo/s que se le indica un sinónimo, adecuado al contexto, de cuatro palabras o definiciones. Se adjudicará 0,25 por cada apartado.

Pregunta 4: Hasta 2 puntos. Con esta pregunta se pretende comprobar **los conocimientos gramaticales** del alumno, en sus aspectos morfológicos y/o sintácticos. Se presentarán oraciones con huecos que el alumno deberá completar/rellenar. También podrán presentarse oraciones para ser transformadas u otro tipo de ítem. Se adjudicará 0,25 a cada "hueco en blanco", y en el caso de las transformaciones o ítems de otro tipo se concederá 0,5 con carácter unitario.

Pregunta 5: Hasta 3 puntos. Se trata de una composición -de 100 a 150 palabras- en la que el alumno podrá demostrar su **capacidad para expresarse libremente en la lengua extranjera**. Se propondrán **dos opciones**, entre las que el alumno elegirá **sólo una.** Se otorgarán 1,5 puntos por el buen dominio de la lengua -léxico, estructura sintáctica, etc.- y 1,5 por la madurez en la expresión de las ideas -organización, coherencia y creatividad.