UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA

ANÁLISIS DE LA FIABILIAD EN LAS PUNTUACIONES HOLÍSTICAS EN ITEMS ABIERTOS

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Marian Amengual Pizarro

Bajo la dirección del doctor

Honesto Herrera

Madrid, 2004

ISBN: 84-669-1947-3

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA

Departamento de Filología Inglesa I

ANÁLISIS DE LA FIABILIDAD EN LAS PUNTUACIONES HOLÍSTICAS DE ÍTEMS ABIERTOS

TESIS DOCTORAL

Marian Amengual Pizarro

Director: Dr. Honesto Herrera Soler

Madrid, 2003

ÍNDICE

CAPÍTULO 5. PRINCIPIOS BÁSICOS DE LAS PRUEBAS DE EXPRESIÓN ESCRITA

CAPÍTULO 6. LOS MÉTODOS DE EVALUACIÓN DE LAS PRUEBAS DE EXPRESIÓN ESCRITA

CAPÍTULO 7. PRINCIPALES VARIABLES DE LA EXPRESIÓN ESCRITA

II PARTE.
PERSPECTIVA EMPÍRICA

CAPÍTULO 8. DESCRIPCIÓN Y ANÁLISIS DE LA MUESTRA

CAPÍTULO 9. ANÁLISIS DE MATERIALES (I): EL CUESTIONARIO

CAPÍTULO 10. ANÁLISIS DE MATERIALES (II): LOS ENSAYOS

CONCLUSIONES Y DISCUSIÓN DE RESULTADOS

RECONOCIMIENTOS

Esta tesis es el resultado de un trabajo de varios años en los que he podido contar con muchas ayudas y apoyos.

En primer lugar, quiero mostrar mi agradecimiento a mi director, el Dr. Honesto Herrera Soler, por su asistencia en la elaboración y el diseño de esta tesis. A él le debo los primeros conocimientos sobre el tema principal que aquí se trata y la inspiración de muchos de los pasos que posteriormente he ido dando. Sus conocimientos estadísticos me han instruido y han hecho posible que abordara este trabajo con seriedad y rigor empírico. Su apoyo científico y moral ha sido constante por lo que mi agradecimiento es inconmensurable.

En segundo lugar, quiero dar las gracias al profesor B. O'Sullivan por la ayuda prestada en el diseño de este trabajo y por la facilitación de mucha de la bibliografía aquí utilizada. Asimismo, he de agradecer las orientaciones científicas de la profesora C. Martínez-Arias que han inspirado gran parte del desarrollo empírico de este trabajo.

En tercer lugar, no quiero dejar de mencionar a los profesores que han participado en este estudio sin los cuales esta tesis nunca habría existido. Soy consciente del esfuerzo y la incomodidad que ha supuesto

llevar a cabo la tarea encomendada por lo que agradezco especialmente su colaboración.

Finalmente quiero dar las gracias a mi familia y a mis amigos por sus ánimos, comprensión y el cariño constante que me han brindado todos estos años.

INTRODUCCIÓN

1.1. Enunciado del tema de tesis, su motivación y especificación de los objetivos.

En esta tesis se realiza un análisis de la fiabilidad de las puntuaciones holísticas de los correctores en la evaluación de los ensayos de la prueba de Inglés. Esta prueba junto con otras pruebas de distintas disciplinas (por ejemplo, Lengua Española, Filosofía, etc.) forma parte de las Pruebas de Acceso a la Universidad (PAAU)[1] o Selectividad.

El origen de este trabajo surgió de la curiosidad que despertó en la autora, como consecuencia de su implicación en la corrección de las PAAU, la observación de la variabilidad de puntuaciones que los correctores asignan a la pregunta abierta del ensayo. Este ejercicio se incluye, por regla general, en la prueba de Inglés de Selectividad de la mayoría de las universidades españolas.

La aparente falta de fiabilidad de las puntuaciones holísticas de los ensayos hacía preciso por una parte, el estudio del marco teórico del estado de la cuestión en la literatura de la evaluación y, por otra, una investigación

[1] A partir de ahora utilizaremos las siglas PAAU para referirnos a las Pruebas de aptitud para el acceso a la universidad.

estadísticamente rigurosa que planteamos como objetivo prioritario de este trabajo.

El marco teórico que nutre la investigación llevada a cabo en este trabajo se comentará en el primer capítulo de forma general y a lo largo de los capítulos sucesivos de esta tesis de forma más concreta. Cabe decir que el campo de la evaluación de la lengua o *Language Testing* (LT) es tan extenso y dinámico que resulta prácticamente imposible plasmar en un trabajo la totalidad de las investigaciones realizadas hasta el día de hoy:

> The field (LT) has become so large and so active that it is virtually imposible to do justice to it, (...), and it is changing so rapidly that any prediction of trends is likely to be outdated before it is printed. (Alderson y Banerjee, 2001)

Sin embargo, la mayoría de los estudios llevados a cabo sobre la evaluación de los ensayos reconocen que existen fluctuaciones en las valoraciones del rendimiento de los candidatos que emiten los correctores. Estas diferencias se observan tanto entre los distintos correctores (i.e. fiabilidad inter-corrector) como entre un mismo corrector en distintas ocasiones (i.e. fiabilidad intra-corrector).

El estudio de la fiabilidad de los correctores resulta especialmente relevante en el contexto de las PAAU dada la importancia que los resultados de estas pruebas tienen en la elección de la futura carrera universitaria de los candidatos que concursan en ellas.

En este trabajo, pretendemos analizar el fenómeno de la fiabilidad desde una aproximación cuantitativa rigurosa. El objetivo que se pretende es:

a) averiguar si las puntuaciones holísticas de los ensayos son fiables,
b) en caso contrario, estudiar, en su momento, la adopción de ciertas medidas destinadas a minimizar los factores fortuitos y las inconsistencias que se observan en las puntuaciones de los correctores.

Algunos estudios atribuyen la falta de fiabilidad de las puntuaciones a variables de los correctores como son: el género, la edad, la experiencia profesional, el lugar de trabajo, etc. (ver Hamp-Lyons 1990f; Brown 1991; Vann, Lorenz y Meyer 1991; Herrera 2000a). Estos estudios van a llevarnos a examinar en este trabajo la influencia que ejercen las variables género y situación laboral en el comportamiento de los correctores en la evaluación de los ensayos.

Al mismo tiempo, entendemos que definir los perfiles docentes, actitudinales y evaluadores de los correctores que participan en este estudio contribuirá al análisis de los factores género y situación laboral. Nuestro objetivo es sistematizar los diversos estilos de enseñanza y evaluación observados y estudiar su posible efecto en las valoraciones de los correctores. El análisis se planteará siguiendo la línea del trabajo llevada a

cabo por Oxford *et al.* (1991) y Dreyer (1998) tanto desde una perspectiva cualitativa como cuantitativa.

Por último, dentro del estudio de la fiabilidad, que constituye el objetivo primordial de este trabajo, pretendemos investigar los elementos de los ensayos que los correctores destacan en sus evaluaciones. De acuerdo con los resultados de varios estudios, los correctores responden a distintos elementos del ensayo o *facets of writing* que determinan la puntuación final que se les asigna (Diederich *et al.* 1961; Grobe 1981; Vaughan 1991; Sweedler-Brown 1994).

En nuestro intento por identificar los elementos que los correctores evalúan diseñamos una escala analítica basada en estudios anteriores (Diederich *et al* 1961.; Jacobs *et al.* 1981) que contenía siete categorías: *contenido*, *organización*, *gramática*, *vocabulario*, *registro*, *mecánica* y *presentación*. Esta escala iba destinada a captar el valor o peso que los correctores asignan a los distintos elementos del ensayo.

Entendimos que las correlaciones entre las evaluaciones holísticas y los distintos elementos analíticos nos podrían aportar información relevante en cuanto a la importancia que los correctores asignan a estos elementos y la influencia que ejercen en la valoración global del ensayo. Siguiendo los estudios de Brown (1991), Connor-Linton (1995) y Milanovic *et al.* (1996) recogimos información cualitativa y cuantitativa sobre las elementos positivos y negativos que los correctores destacaban en cada uno de los ensayos a medida que los iban evaluando. La obtención de esta información

nos iba a permitir delimitar con mayor exactitud los resultados obtenidos anteriormente.

Por último, dentro de los aspectos que parecen afectar en mayor medida las evaluaciones holísticas de los correctores, algunos estudios destacan la influencia del error (McDaniel 1985; Santos 1988;Vann, Meyer y Lorenz 1991). La investigación de este último aspecto se plantea a través del estudio de la evaluación de distintos errores relacionados con las siete categorías analíticas anteriormente citadas e incluidas en frases individuales y descontextualizadas.

No obstante, y debido a las críticas suscitadas en contra de la metodología de estudio basada en el análisis de los errores descontextualizados (Davies 1983; Ludwig 1982), decidimos compatibilizar este análisis con la evaluación de los aspectos analíticos contenidos en el discurso, es decir, en los ensayos. La combinación de ambas metodologías nos va a permitir obtener una visión más real de la reacción de los correctores ante el error y del papel que desempeña el contexto en el que se incluyen.

La investigación de estos últimos aspectos se abordó principalmente desde una perspectiva pedagógica dado que, como elementos activos dentro del proceso de enseñanza-aprendizaje, deseábamos descubrir la reacción de los correctores ante los distintos errores y averiguar los elementos relacionados con la *forma* y el *contenido* de los ensayos que determinan su evaluación.

De igual modo, creemos que la obtención de dicha información nos ayudará a proveer a nuestros estudiantes con un sistema de retroalimentación (*feed-back*) útil y adecuado que les posibilite acomodar sus expectativas a la de los correctores de manera que se les garantice el éxito académico.

Así fue como quedaron determinados los elementos centrales de nuestra investigación:

1) el análisis de la fiabilidad de las puntuaciones holísticas de los correctores en la evaluación de los ensayos,
2) el diseño de los perfiles docentes, actitudinales y evaluadores de los correctores de nuestro estudio
3) la influencia que ejercen las variables género y situación laboral en el comportamiento de los correctores y
4) los aspectos de la producción escrita que los correctores destacan en sus evaluaciones, entre ellos la incidencia del factor error.

1.2. Hipótesis

Nuestro trabajo es un intento de responder a seis hipótesis de investigación a través de un amplio recorrido por el campo de LT. Estas hipótesis se corresponden con las preocupaciones en torno el estudio de la

fiabilidad en general y de forma concreta con el estudio de la fiabilidad inter-corrector e intra-corrector:

HIPÓTESIS 1. No hay diferencias significativas entre los diversos correctores en la evaluación holística de los ensayos (i.e. fiabilidad inter-corrector).

HIPÓTESIS 2. Las puntuaciones de un mismo corrector en las distintas ocasiones son consistentes (i.e. fiabilidad intra-corrector).

HIPÓTESIS 3. No hay diferencias significativas entre las evaluaciones holísticas y las evaluaciones analíticas.

HIPÓTESIS 4. No hay diferencias significativas en las valoraciones de los correctores de acuerdo con las variables género y situación laboral.

HIPÓTESIS 5. Todos los componentes del ensayo ejercen el mismo peso en la evaluación holística final.

HIPÓTESIS 6. No hay diferencias significativas entre las evaluaciones de los errores contenidos en frases descontextualizadas y los errores contenidos en el discurso (i.e. los ensayos).

Con el objeto de ratificar las hipótesis antes formuladas, nos planteamos las siguientes preguntas a las que daremos respuesta en función de los datos obtenidos en nuestra aportación empírica (II Parte de la tesis):

1. ¿Cuál es el grado de fiabilidad de las puntuaciones holísticas y analíticas de los ensayos de Selectividad entre los diversos correctores (i.e. fiabilidad inter-corrector)?

2. ¿Hasta qué punto son fiables las puntuaciones de un mismo corrector en la evaluación de los ensayos de Selectividad en diferentes ocasiones (i.e. fiabilidad intra-corrector)?

3. ¿Hay correlación entre las evaluaciones holísticas y las evaluaciones analíticas de los ensayos?

4. ¿Qué efecto ejerce el dominio lingüístico de los ensayos en las puntuaciones holísticas finales?

5. ¿Cuáles son los perfiles docentes, personales y evaluadores de los correctores que participan en nuestro estudio?

6. ¿Qué influencia ejercen las variables género y lugar de trabajo de los correctores en las puntuaciones holísticas de los ensayos de Selectividad?

7. ¿Qué principales aspectos problemáticos valoran los profesores en los ensayos de los estudiantes?

8. ¿Cuáles son las categorías analíticas que determinan las evaluaciones holísticas de los ensayos?

9. ¿Qué aspectos positivos y negativos destacan los correctores en la evaluación de los ensayos de Selectividad?

10. ¿Qué errores cometidos por los estudiantes se penalizan de forma más estricta?

1.3. Organización y contenido de la tesis

Por último, cabe mencionar la organización y el contenido de esta tesis que se divide en dos partes fundamentales. La primera parte establece los fundamentos teóricos sobre la evaluación de la lengua y la evaluación de la expresión escrita. En la segunda parte, se ofrece la aportación empírica de este trabajo. Asimismo, se incluye esta introducción y un apéndice al final de este volumen en el que se recoge el material que se ha utilizado para llevar a cabo este estudio.

La primera parte de la tesis comprende los siete primeros capítulos (véase índice). En el capítulo 1, se hace una revisión crítica de los principales enfoques de la evaluación de la lengua (LT). El capítulo 2 contiene la definición de este concepto y sus rasgos más característicos. En este capítulo también se analizan los modelos de dominio de la lengua inglesa más destacados que inspiran el diseño de las pruebas evaluadoras en general y de las pruebas de expresión escrita en particular. El capítulo 3 define las pruebas de expresión escrita y sus rasgos característicos dentro del contexto académico. En el capítulo 4, se aborda el estudio del ensayo desde una perspectiva histórica que nos ayuda a entender el enfoque de la enseñanza y de la evaluación actual del mismo. A continuación, en el capítulo 5, se estudian los principios básicos que rigen las pruebas de expresión escrita y se presta especial atención a la determinación del constructo de expresión escrita. El capítulo 6, analiza los dos principales métodos de evaluación en las pruebas de expresión escrita: el método holístico y el analítico y se estudian las ventajas e inconvenientes que presentan cada uno de ellos. Esta primera parte se cierra con el capítulo 7 que nos resume las variables de la evaluación de la expresión escrita más relevantes en nuestro trabajo: la figura del corrector, las variables relativas a la tarea y las variables relativas a los candidatos. En la investigación de la figura del corrector, objeto principal de estudio de este trabajo, se examinan los factores personales y elementos de la producción escrita más destacados que influyen en los correctores. Asimismo, se hace una breve introducción a la técnica de formación de los correctores en técnicas de

evaluación destinada a garantizar la fiabilidad de las puntuaciones entre los correctores. A continuación, se analizan las variables relativas a la tarea: los efectos producidos por el método de evaluación, especialmente la elección del tema del ensayo, y la influencia que ejerce el factor tiempo en la calidad del producto final. Finalmente, se aborda el estudio de las variables relativas a los candidatos como escritores de los textos.

La segunda parte de la tesis desarrolla la parte empírica de nuestro trabajo. El capítulo 8 introduce el contexto de nuestro estudio y la metodología llevada a cabo en nuestra investigación. Los resultados del análisis de materiales de nuestro trabajo se presentan en los dos capítulos siguientes. El capítulo 9, analiza los resultados del cuestionario suministrado a los correctores. En el capítulo 10, se comentan los resultados obtenidos del estudio empírico de los ensayos.

Por último se recogen las conclusiones derivadas de esta investigación y sus principales aplicaciones pedagógicas.

PRIMERA PARTE

CAPÍTULO 1. PERSPECTIVA HISTÓRICA DE LOS PRINCIPALES ENFOQUES SOBRE LA EVALUACIÓN DE LA LENGUA

1.1. Introducción

El objetivo inicial de la evaluación del aprendizaje de una lengua o *Language Testing* (LT)[1] de acuerdo, tanto con la tradición británica como con la del positivismo norteamericano, fue satisfacer una creciente demanda pedagógica en la enseñanza de lenguas. Hoy en día, LT persigue evaluar los conocimientos adquiridos por los estudiantes.

Este nuevo planteamiento surgió a raíz de los cursos intensivos de enseñanzas de lenguas extranjeras a adultos durante la Segunda Guerra Mundial. En este contexto se hizo explícita la competencia comunicativa y ello motivó su evaluación (*communicative testing*, Fulcher 1998). Dicho enfoque se aplicó a campos paralelos como la elaboración de material didáctico para la enseñanza de lenguas y su metodología (Davies, 1990).

En este capítulo haremos un breve repaso de los más recientes y principales enfoques de LT. Empezaremos por un breve introducción de la influencia que tuvo la visión estructuralista del lenguaje y las pruebas de

[1]A partir de ahora, utilizaremos las siglas LT para referirnos a la evaluación de la lengua.

elementos discretos. En segundo lugar, se hablará del resurgimiento de las pruebas globalizadas o integrales. Dada la gran influencia que ejerció el movimiento comunicativo en la enseñanza, el aprendizaje y la evaluación de segundas lenguas, nos detendremos en él especialmente. Analizaremos de forma extensa las investigaciones realizadas y las principales características que definen la tendencia actual que ha generado el movimiento comunicativo y que son relevantes en nuestro trabajo. En último lugar, comentaremos los avances actuales más significativos llevados a cabo en el campo de LT.

1.2. La etapa Estructuralista. Las pruebas objetivas o de elementos discretos

Durante la década de los años 50 y a principios de los años 60 la visión estructuralista conductista del lenguaje ejerció una notable influencia en los métodos didácticos y evaluativos. La publicación de la obra de Lado *Language Testing* (1961, 1964) va a suponer el ejemplo más claro de la visión estructuralista del aprendizaje y la evaluación de la lengua.

La teoría de Lado (1964) sobre LT asume que: “language is a system of habits in communication” (p. 27) y defiende la utilización de pruebas de elementos discretos de carácter objetivo como son las pruebas de elección múltiple. Dichas pruebas están compuestas por ítems de acuerdo con una aproximación taxonómica o clasificatoria de las destrezas o habilidades lingüísticas.

De este modo, desde mediados de los años 60 y durante los años 70 se puede afirmar que el campo de LT se nutrió de la visión teórica de la

habilidad lingüística basada en el desarrollo secuencial de cuatro destrezas fundamentales (i.e. comprensión oral, expresión oral, comprensión lectora y producción escrita) y de sus componentes (por ejemplo: gramática, vocabulario y pronunciación).

Cabe señalar que, Lado (1964) le concede una especial importancia al análisis contrastivo que serviría para elaborar un inventario de errores previsibles que pueden cometer los estudiantes de lenguas extranjeras. La insistencia de Lado en el análisis contrastivo no cumplió con las expectativas previstas. No obstante, la aproximación taxonómica de LT de orden analítico defendida por Lado (1961, 1964), máximo exponente del estructuralismo en este campo, tuvo sus seguidores. Junto a él, diversos autores defendieron las pruebas de elementos discretos y compartieron su preocupación por la fiabilidad psicométrica (Valette 1967; Carroll 1968; Harris 1968; Heaton 1975) y la interpretación de los resultados de las pruebas en relación a normas. Este último enfoque se denominó evaluación normativa (*norm-referencing*): "(norm-referencing)...grades an individual's performance in relation to that of his / her peers, that is in terms of relative performance rather than their absolute performance." (Gipps, 1994)

De forma sintética se puede decir que los estructuralistas conciben el lenguaje como un fenómeno que debe de analizarse utilizando un enfoque científico ("scientific approach", Lado 1964) a base de componentes (i.e. fonológico, morfosintáctico y léxico-semántico) y unidades y subunidades (por ejemplo, el fonema, el morfema, el lexema, etc.). De acuerdo con este planteamiento el lenguaje puede dividirse en diversos componentes o en

distintas partes sujetas a análisis. Al ser una aproximación científica, los resultados que se obtengan deben ser precisos y significativos. De ahí, su énfasis en la construcción de pruebas objetivas que disponen de una mayor fiabilidad estadística.

1.3. Las Pruebas Integrales

El resurgimiento de las pruebas integrales o globalizadas aparece como réplica a la excesiva atomización que planteaba la visión estructuralista del lenguaje. Este movimiento coincide, además, con el interés que suscita la gramática transformacional-generativa cuya visión sintética del lenguaje se antepone a su visión analítica. La gramática transformacional-generativa no centra su atención en las taxonomías sino en los procesos y en la generación del lenguaje entendido como un todo (Alcaraz y Ramón, 1980).

Como se podrá comprobar, a medida que entramos en la era integradora-sociolingüística (*integrative-sociolinguistic*, Skehan 1987)[2] se enfatizará la importancia que se le concede a los aspectos de la comunicación y el contexto.

La mayoría de las críticas en contra de la visión estructuralista de la lengua provienen de Oller (1976, 1979) quien defendió la visión de la competencia de la lengua entendida como un conjunto de habilidades unificadas y que interactúan de tal modo que no pueden separarse y

[2]Skehan (1987) adapta los tres períodos de la historia de la evaluación de la lengua que distinguió Spolsky (1975) (i.e. *the pre-scientific*, *the psychometric-structuralist* y *the psycholinguistic-sociolinguistic*) y distingue tres etapas: *the pre-scientific*, *the psychometric-structuralist* y *the integrative-sociolinguistic*

analizarse de forma individual. La *competencia comunicativa* se define como una competencia global y, por lo tanto, requiere una integración de los elementos (de ahí, el término *integrative testing*) que no puede ser captada por la suma de las distintas partes que componen una prueba de elementos discretos. Oller (1979) nos explica:

> The concept of an integrative test was born in contrast with the definition of a discrete point test. If discrete items take language skill apart, integrative tests put it back together. Whereas discrete items attempt to test knowledge of language one bit at a time, integrative tests attempt to assess a learner's capacity to use many bits all at the same time. (p. 37)

Algunas de las pruebas integrales más conocidas y citadas como ejemplos típicos son: la redacción, las pruebas cloze (*cloze test*), las pruebas C (*C*-Tests) o los dictados. Diversos autores sostienen que las pruebas integrales son más válidas que las pruebas de elementos discretos dado que éstas engloban la totalidad del fenómeno comunicativo (Oller y Richards 1973; Palmer y Spolsky 1975; Savignon 1982). De hecho, en la actualidad, ha surgido un renovado interés por la utilización de ciertos formatos de las pruebas integrales como son las pruebas Cloze (Chapelle y Abraham 1990; Stevens 1991; Brown 1993; Herrera 2002) o las pruebas C (Chapelle, 1994).

La controversia creada en torno a la posible definición de un *constructo*[3] a través de una prueba integral de dominio de la lengua (i.e. *proficiency test*) ha continuado hasta nuestros días. El debate se ha centrado

en la posibilidad de dividir la habilidad de la lengua en distintos componentes susceptibles de ser analizados individualmente.

Oller (1979) sostiene que el dominio de la lengua o *General Language Proficiency* (GLP) está formado por una competencia única que subyace al resto de las destrezas lingüísticas. Su concepto de dominio global (*overall proficiency*) junto con la idea de la *indivisibilidad* de la lengua desembocará en la formulación de su teoría de la competencia unitaria (*unitary trait hypothesis*). Esta teoría ha sido muy cuestionada recientemente debido a la falta de evidencia empírica que la apoye y al creciente número de investigaciones que defienden, por el contrario, la teoría de la divisibilidad de competencias o *divisibility hypothesis* (Farhady 1983; Porter 1983). Se ha de decir que este nuevo planteamiento va ganando cierto reconocimiento entre los investigadores, en parte, debido a que la propia experiencia pedagógica nos demuestra continuamente que los estudiantes exhiben diferentes dominios de la lengua en las diversas tareas académicas que acometen.

1.4. Las pruebas analíticas versus las pruebas integrales

Davies (1982) afirma que a mediados de los años 70, las distintas concepciones del lenguaje dibujarían un continuo. Según esta visión teórica, las pruebas de elementos discretos representarían el extremo de un polo que se extendería hacia el polo opuesto y cuyo extremo vendría representado por las pruebas integrales y de carácter globalizador.

---

[3]El constructo es un término que utiliza la psicología para definir “underlying skills or attributes”. Gipps (1994) lo define del siguiente modo: “a construct is an explanatory device, so-called because it is a theoretical construction about the nature of human behaviour.”

Davies (1982) señala que existe una *tensión* entre los dos polos del continuo, es decir, entre la visión analítica de la lengua y la visión integral o globalizadora. Este autor interpreta el progreso en el terreno de LT como el paso de un extremo del continuo a otro, o lo que es lo mismo, de una concepción de la lengua a otra[4].

A pesar de ello, Davies (1982) considera que el modo más fructífero de resolver la *tensión* que hay entre la concepción analítica y la integral es a través de: "a combination of these two views, the analytical and the integrative" (p.131). Asimismo, sostiene que es muy probable que no exista una prueba totalmente analítica o totalmente integral y relaciona estos dos conceptos con los aspectos de fiabilidad (*reliability*) y validez (*validity*):

> The two poles of analysis and integration are similar to (and may be closely related to) the concepts of reliability and validity. Test reliability is increased by adding to the stock of discrete items in a test: the smaller the bits and the more of these there are, the higher the potential reliability. Validity, however, is increased by making the test truer to life, in this case more like language in use. (Davies 1982: 131)

La mayor ventaja que se atribuye a las pruebas de elementos discretos es su fiabilidad. Recordemos que para ello se utilizan pruebas objetivas fácilmente cuantificables. Sin embargo, una de las principales desventajas que presentan estas pruebas reside en la dificultad que plantea la definición del constructo que se pretende medir. Algunos autores

[4] Esta *tensión* denunciada por Davies (1982) coincide con las dos últimas etapas en la evolución de LT establecidas por Spolsky (1975): la *psychometric-structuralist* o analítica (etapa 2) y la etapa *psycholinguistic-sociolinguistic* o integral (etapa 3). La etapa 1

cuestionan la posibilidad de medir el constructo de dominio la lengua (*construct proficiency*) analizando simplemente los diversos aspectos que lo componen. Oller (1979) acusa las deficiencias de este último planteamiento en la siguiente cita:

> Discrete-point analysis necessarily breaks the elements of language apart and tries to teach them (or test them) separately with little or no attention to the way those elements interact in a larger context of communication. What makes it ineffective as a basis for teaching or testing languages is that crucial properties of language are lost when its elements are separated. The fact is that in any system where the parts interact to produce properties and qualities that do not exist in the part separately, the whole is greater than the sum of its parts...organisational constraints themselves become crucial properties of the system which simply cannot be found in parts separately. (p.212)

Otra de las criticas a las que el paradigma psicométrico tendrá que hacer frente es la idea de la *unidimensionalidad*. Esta idea hace referencia a la conceptualización de los constructos y a las técnicas utilizadas para el análisis de los ítems en las pruebas (Goldstein 1993; Gipps 1994). Según la teoría psicométrica, los diversos ítems de una prueba deberían medir un único atributo o destreza subyacente. Sin embargo, esta estructura unidimensional resulta ilógica ya que muchos de los atributos o destrezas que se miden en las pruebas son más bien multidimensionales. Ello explica que ciertos investigadores critiquen y rechacen las técnicas actuales que se basan en estos primeros supuestos (Goldstein, 1992).

---

denominada *pre-scientic* representa el período anterior a los años 50 en el que no se realizó ninguna investigación seria sobre el proceso evaluativo.

1.5. La evaluación de la competencia comunicativa

En el primer coloquio sobre la investigación en el campo de LT, *Language Testing Research Colloquium* (LTRC) en 1979, el modelo analítico basado en la división de destrezas y componentes y la visión de una competencia única e indivisible del dominio de la lengua (*unitary competence hipótesis* (UCH), Oller 1979) fueron duramente criticados.

Un año antes, en 1978 Widdowson sacaba a la luz su libro *Teaching Language as communication*. Surgieron nuevas ideas en torno a la habilidad comunicativa, de modo que, a finales de los años 70 y a principios de los 80 podemos hablar de la llegada de un nuevo enfoque: la evaluación de la competencia comunicativa (*communicative language testing*, (CLT)[5]) que Morrow (1979) bautizaría con el nombre de *The Promised Land*[6].

El movimiento de CLT, primordialmente considerado como un fenómeno británico, surgiría como una rebelión contra las pruebas de elección múltiple y, sobre todo, contra el concepto de fiabilidad establecido en el trabajo de Lado (1961). Como se recordará, la fiabilidad de las pruebas evaluadoras tan sólo se podía incrementar utilizando pruebas objetivas como son las pruebas de elección múltiple. Este hecho supone tener que hacer frente a la *tensión* que existe entre los conceptos de fiabilidad y validez denunciada por Davies (1978). Dicha *tensión* se explica porque el concepto de validez conlleva el diseño de tareas paralelas a las actividades o tareas

[5] A partir de ahora utilizaremos la abreviatura CLT para referirnos a la evaluación comunicativa de la lengua.

[6] Morrow (1979) tradujo los tres períodos de la historia de LT que distinguió Spolsky (1975) (*the pre-scientific*, *the psychometric-structuralist* y *the psycholinguistic-sociolinguistic*) en los períodos llamados: *The Garden of Eden*, *the Vale of Tears* y *The Promised Land*. Este

reales. Sin embargo: “there is no real-life situation in which we go around asking or answering multiple-choice questions” (Underhill 1982: 18). Por el contrario, la fiabilidad únicamente se incrementa si se utilizan medidas objetivas. De ahí que, cuanto más válida sea una prueba menos fiable es esta última.

En este primer subepígrafe 1.5.1. realizaremos una breve introducción a las principales teorías que inspiraron el movimiento comunicativo. Posteriormente, en el subepígrafe 1.5.2. entraremos en profundidad en el análisis de las pruebas comunicativas y comentaremos sus rasgos más distintivos.

1.5.1. Fuentes teóricas

El nuevo enfoque de la habilidad lingüística expuesta por los proponentes de la llamada *competencia comunicativa* (Hymes, 1972; Canale and Swain, 1980) ejerció una enorme influencia en el ámbito de LT durante la década posterior. Ciertos aspectos de la lengua como el aspecto discursivo o el sociolingüístico fueron considerados esenciales en el nuevo modelo de competencia comunicativa y el contexto pasó a jugar un papel decisivo en la interpretación del discurso. Hymes (1972) subrayó el hecho de que la competencia comunicativa implicaba la habilidad de *usar* la lengua.

A partir del modelo bidimensional de Hymes (1972) que comprendía una dimensión lingüística y otra sociolingüística surgieron otros modelos

---

último período señalaba la llegada de la evaluación comunicativa de la lengua a finales de los años 70 y a principios de los años 80.

como el de Canale y Swain (1980) y, finalmente, el de Canale (1983). Este último modelo estaba compuesto por cuatro dimensiones o aspectos: el aspecto lingüístico, el sociolingüístico, el discursivo y, por último, la competencia estratégica. Sin embargo, y a pesar de la importancia que tuvieron estos modelos en la compresión y definición del constructo de CLT, la mayoría de estos modelos no han sido validados científicamente.

Morrow (1979) y Canale y Swain (1980) defendieron la necesidad de que CLT incluyera tanto el aspecto de *competence* (conocimiento de la forma y uso de la lengua) como el de *performance* o activación de la competencia en una situación comunicativa concreta. De este modo, CLT se concibe como la evaluación del aspecto de *performance* relacionado con la producción efectiva de ideas en contextos específicos (Morrow, 1979).

Entre los modelos actuales de competencia comunicativa destaca el modelo de Bachman (1990) denominado *communicative language ability* (CLA). Este modelo retiene básicamente los mismos componentes de los modelos comunicativos de Canale y Swain (1980) y Canale (1983) pero expande el papel que juega la competencia estratégica. La descripción de las funciones de la competencia estrategia explican como los diversos componentes de la competencia de la lengua (i.e. gramatical, textual, pragmática y sociolingüística) interactúan los unos con los otros y con los aspectos de la situación o contexto del uso de la lengua. Como veremos en el capítulo 2 (subepígrafe 2.4.1.), el modelo de Bachman (1990) ha sido especialmente productivo en el ámbito de LT, especialmente para el diseño de pruebas evaluadoras que tengan en cuenta las características del método

y de las tareas. Este modelo es el que inspira el desarrollo del ejercicio del ensayo en la prueba de Inglés de Selectividad.

Una reciente expansión del modelo de Bachman (1990) es el modelo de Bachman y Palmer (1996) que expande el role de la competencia estratégica como estrategias metacognitivas (i.e. *goal setting*, *assessment*, y *planning*) e incluye una discusión de la función que realizan el *topical knowledge* o k*nowledge schemata* y el *affective schemata* en el uso de la lengua.

1.5.2. Las pruebas comunicativas

1.5.2.1. Concepto y rasgos característicos

El trabajo de Morrow (1979) y su provocativa aproximación comunicativa de LT desató un extenso debate acerca de lo que se entendía por pruebas comunicativas y auténticas que se prolonga hasta el día de hoy (Alderson 1981; Canale 1984; Spolsky 1985; Fulcher 2000).

Morrow (1979) denunció que las pruebas convencionales no medían ciertos aspectos importantes del uso de la lengua en una situación comunicativa, esto es: la interacción, el contexto, el propósito, la activación de la competencia, la autenticidad, etc. A partir de ahí, empezaron a surgir determinados conceptos que se relacionaban directamente con CLT. Entre ellos destacan:

- Tareas o actividades reales basadas en comportamientos reales;
- Autenticidad. Ello implica la implementación de tareas con un propósito definido y que tengan lugar en un contexto específico y significativo;
- Validez aparente (*face validity*). Es decir, que las pruebas parezcan medir situaciones reales de comunicación;
- *Performance*. Este concepto sugiere que se tengan en cuenta aspectos de la comunicación como el hecho de que la lengua es impredecible o la necesidad de que la comunicación se base en la interacción o negociación de significados.

Uno de los mayores inconvenientes que cita Rea (1991) dentro del paradigma comunicativo es el escaso número de publicaciones que hay en torno a los aspectos comunicativos de la evaluación. Sobre todo, si dichas publicaciones se comparan con el gran número de ellas editadas en torno a los aspectos comunicativos de la enseñanza.

A pesar de ello, algunos autores definen los rasgos o características esenciales que toda prueba comunicativa debe poseer. Así, por ejemplo, Swain (1984) nos describe una serie de criterios y premisa básicas que cabe tener presente en la elaboración de las pruebas comunicativas: "1) *start from somewhere*-; 2) *concentrate on content*; 3) *bias for best* y 4) *work for washback.*" En otras palabras, se ha de disponer de un modelo teórico que sirva de base y punto de referencia; el contenido ha de ser motivador e interesante; se ha de estimular una buena producción escrita y, por último,

se ha de utilizar la información obtenida en las pruebas para favorecer el proceso didáctico. (ver Swain, 1984)

Con respecto a este último punto, Davies (1990) sostiene que:

> Communicative tests were not introduced because it was thought important for the teaching to be communicative. They were introduced because it was claimed that that is what language learning should be like – whether right or wrong (p. 147)

Este autor continúa diciendo que las pruebas ejercerán siempre un cierto impacto (i.e. *washback*) en el proceso didáctico y evaluativo. De ahí, la importancia de que tanto las pruebas de lengua como los programas y los métodos de enseñanza y evaluación estén bien coordinados.

Rea (1991) nos resume también algunos de los criterios de las pruebas comunicativas que han logrado cierta *aceptabilidad* dentro del paradigma comunicativo. Entre ellos se incluyen:

> - the assessment of "... skills ... certain realistic things using language"
> - task-based: realistic language tasks
> - realistic tasks using authentic materials
> - the integration of language skills
> - the integration of languages
> - information-getting and information-sharing activities
> - appropriacy of tasks to target students
> - appropriacy and clarity of rubrics
> - context
> - learner participation
>
> (Rea 1990: 107).

A pesar de la consideración que se merecen los criterios anteriores, algunos autores cuestionan la posibilidad real de construir pruebas que

midan la competencia comunicativa. Ello se debe a que todavía, hoy en día, resulta difícil entender el concepto de prueba comunicativa. Harrison (1991) se pregunta:

> Does it mean assessing the linguistic behaviour of the student in an imitation of a real life setting, or assessing the skills which are considered essential for this kind of behaviour? (p. 97)

Weir (1983) más que centrarse en la definición explícita de este concepto prefiere aclararnos que la única diferencia entre los procesos de enseñanza y evaluación (*teaching* / *testing*) dentro del paradigma comunicativo radica en la cantidad de ayuda que se le ofrece al candidato por parte del profesor o compañeros:

> The help that is normally available in the teaching situation e.g. prompts, reformulation of questions, encouragement, correction and the opportunity to try again, is removed in a test for reasons of reliability of measurement. (Weir 1983: 14)

En cualquier caso, la necesidad de definir y especificar el constructo, es decir: “what it is that is to be tested” (Weir 1983: 20) resulta un hecho innegable dentro del paradigma comunicativo. Weir (1983) expone que el creciente interés demostrado en el ámbito del Inglés para Fines Específicos o *English for Specific Purposes* (ESP) refleja precisamente esta preocupación. De ahí, que el autor proponga que este acrónimo signifique: “English for **Specified** Purposes” (Weir 1983: 20). Dicha propuesta pretende

enfatizar la necesidad de que los procesos de enseñanza y de evaluación se entiendan como situaciones específicas o particulares y no como situaciones generales. Alcaraz (2000) comparte esta misma opinión y aboga por la utilización del nombre “Inglés Profesional y Académico” (IPA) porque resulta más concreto y específico que el de Inglés para Fines Específicos (IFE). Esta nueva aproximación implica, además, que el desarrollo de CLT requiere la elaboración de una gran variedad de pruebas que den respuesta a diferentes fines o propósitos. Con ello se consigue que: “no-one now will fruitlessly pursue the ‘one best method’.” (Rea 1990: 108).

A pesar de las críticas en contra de CLT, ha habido avances significativos en el terreno de las pruebas comunicativas tanto en la práctica como en la investigación. Como resultado: “communicative testing *is* here to stay.” (Rea 1990: 110).

1.5.2.2. La extrapolación de los resultados

Probablemente, la mayor objeción que se levanta en contra de CLT lo constituye la dificultad de extrapolar los resultados más allá de las pruebas comunicativas concretas.

Weir (1983) nos sintetiza el problema a continuación:

> A communicative test implies the specification of performance tasks closely related to the learner’s practical activities, that is, to the communicative contexts in which he would find himself; therefore creating a problem of generalisability of tasks selected. (p. 14)

Para resolver este dilema, Davies (1982) considera establecer un compromiso entre la evaluación de la competencia lingüística y la evaluación de la competencia comunicativa. Según este autor, estas dos competencias se encuentran fuertemente ligadas a los aspectos de: fiabilidad / validez, evaluación de elementos discretos / evaluación integral. Davies (1982) manifiesta:

> 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. (p. 149)

El argumento anterior le permite a Davies (1982) justificar la incorporación de los aspectos gramaticales en las pruebas comunicativas:

> What remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that grammar is at the core of language learning...Grammar is far more powerful in terms of generalisability than any other language feature. Therefore grammar may still be the most salient feature to teach, and to test. (p. 151)

No obstante, Kelly (1978) rebate dicho planteamiento. El autor sostiene que no se sabe con certeza qué peso tienen los aspectos formales de la lengua como son, por ejemplo, los aspectos morfológicos o el uso

indebido de la condicional del tipo II en Inglés sobre la comunicación en lengua inglesa.

Cabe destacar que Morrow (1977) admite y reconoce la dificultad que plantean las pruebas comunicativas a la hora de extrapolar los resultados:

> The very essence of a communicative approach is to establish particular situations with particular features of context, etc., in order to test the candidate's ability to use language appropriate in terms of a particular specification. While it is hoped that the procedure discussed will indeed be revealing in those terms, they cannot strictly speaking reveal anything of the candidate's ability to produce language which is appropriate to a situation different in even one respect from that established. (p. 53)

Skehan (1988) también se muestra algo escéptico en cuanto a la posibilidad de llevar a cabo la CLT de forma satisfactoria. El autor manifiesta:

> Language testing research requires large test batteries and subject sizes, as well as difficult operationalisations in test form of the constructs embodied in the models of communicative competence. Such research may encounter difficulty in finding appropriate settings which enable effective research design. It may also have difficulty in developing measures to adequately test the constructs concerned. To a certain extent this may reflect the testability of applied linguistic theories in general, but it may well be the case that the determining factor in future validation studies is the resourcefulness of investigators to devise adequate research studies to address the issues they have identified. (p. 5)

Por el contrario, Alderson *et al.* (1981) parecen quitarle importancia a este asunto. Según estos autores, la extrapolación de los resultados no es un aspecto primordial a tener en cuenta en CLT:

> ...if we cannot generalise from performance in one situation to performance in a variety of situations, if we can say something about performance in <u>one</u> situation, then we have made progress (Alderson *et al*.1981: 16)

No obstante, resulta difícil defender esta postura en la actualidad ya que lo que pretenden los investigadores es establecer inferencias a partir de una o varias tareas y extrapolarlas a un conjunto potencial de tareas llevadas a cabo en una situación real.

Así, los resultados obtenidos en las PAAU nos permitirán hacer ciertas inferencias sobre la habilidad lingüística de los candidatos en lengua inglesa.

1.6. El campo de la evaluación en las últimas décadas: estado de la cuestión.

Durante los años 80 se produjo una notable expansión en diversas áreas de conocimiento que, si bien no tuvieron tanta relevancia como el movimiento comunicativo, iban a influir de forma considerable en el campo de LT.

Entre estas áreas de conocimiento, cabe destacar las investigaciones llevadas a cabo en el terreno de la adquisición de segundas lenguas o S*econd Language Acquisition* (SLA). Concretamente, el estudio y planteamiento de ciertas hipótesis destinadas a identificar diferentes estadios de adquisición del lenguaje propició que dichos estadios se incorporaran al ámbito de LT. Ello permitiría determinar diferentes niveles de dominio de la lengua (Pienemann *et al*, 1988).

Precisamente, la cooperación que se estableció entre estas dos áreas de investigación: adquisición de segundas lenguas y evaluación de la lengua confirma el hecho de que, a finales de los años 80, LT se consideraba una disciplina con entidad propia dentro del campo de la Lingüística Aplicada.

Los años 90 supondrán la continuación y consolidación de algunos cambios y avances que se iniciaron en la década anterior. A pesar de que autores como Skehan (1991) critican el escaso progreso que se había realizado hasta entonces en el terreno de LT, en los años 90, se van a producir logros notables en este campo. Sobre todo, en el ámbito concerniente a la activación de la competencia lingüística o *language performance*.

Según Bachman (2000), la década de los 90 se va a caracterizar por una serie de avances significativos en diferentes áreas que contribuirán a configurar la visión de LT a finales del siglo XX:

> *a) research methodology*
> *b) practical advances*
> *c) factors that affect performance on language tests*
> *d) authentic, or performance, assessments; and*
> *e) concerns with the ethics of language testing and professionalizing the field*
> (Bachman 2000: 4).

Siguiendo el esquema de Bachman (2000), en los siguientes subepígrafes, haremos un breve repaso de los principales elementos que

afectarán el estudio de LT. El esquema que vamos a desarrollar es el siguiente:

1.6.1. Métodos de investigación

1.6.1.1.Métodos cuantitativos

1.6.1.2. Métodos cualitativos

1.6.2. Nuevas Tecnologías

1.6.3. Investigación de los principales elementos que afectan el desarrollo de LT

1.6.4. *Performance assessment*

1.6.5. Aspectos éticos

A continuación, comentaremos de forma sintética cada uno de estos puntos:

1.6.1. Métodos de investigación

Dentro de los métodos de investigación distinguiremos los métodos cuantitativos y los cualitativos.

1.6.1.1. Métodos cuantitativos

Los clásicos coeficientes de fiabilidad basados en la norma y el análisis de factores van a ser sustituidos por metodologías cuantitativas superiores como son:

1.6.1.1.1 la evaluación criterial;

1.6.1.1.2. la teoría de la generalizabilidad;

1.6.1.1.3. la teoría de respuesta al ítem y

1.6.1.1.4. la *structural equation modelling*

1.6.1.1.1. La evaluación criterial.

Este tipo de evaluación surge como una alternativa a la evaluación normativa que, como explicaremos más adelante, es la que se lleva principalmente a cabo en las PAAU. La evaluación criterial implica interpretar los resultados de las pruebas a partir del propio individuo y no en relación al resto de los individuos o candidatos. Para ello, se toman como punto de referencia unos criterios o *standards* que definen el dominio de la habilidad que se desea medir.

El optimismo que muestran Bachman (1989) y Brindley (1989) por la incorporación de la evaluación criterial al campo de LT contrasta con el pesimismo de Skehan (1989) sobre este tema. Este último autor, a pesar de considerar deseable el progreso en el terreno de la evaluación criterial, cuestiona su factibilidad (Skehan, 1991). Entre otras aspectos, Skehan (1989) subraya el problema que representa el querer describir algo tan multi-dimensional como es *language performance* carente todavía de unos estadios de desarrollo claramente definidos.

Dentro del campo del *educational measurement* la evaluación criterial jugará un papel primordial ya que según afirma Gipps (1994) se trata de: “to use measurement constructively to identify strengths and weaknesses individuals might have so as to aid their educational progress” (p. 8). Recientemente, Bachman (2000) explica que las aplicaciones de la evaluación criterial se han adaptado y llevado al terreno de LT tanto en las pruebas de rendimiento (Hudson y Lynch, 1984) como en el desarrollo de las pruebas de admisión o *placement tests* (Brown, 1988). No ha ocurrido así,

en cambio, con las pruebas de dominio de la lengua como son las PAAU cuya evaluación es fundamentalmente normativa.

Cabe decir que, en contra de lo previsto, las investigaciones llevadas a cabo en torno a la evaluación criterial no ha sido extensas. Probablemente, los avances más notables lo constituyen una importante base teórica ofrecida por Lynch y Davidson (1993) y algunas investigaciones recientes realizadas por Brown (1993) y Kunnan (1992).

1.6.1.1.2. La teoría de la generalizabilidad

Esta teoría permite estimar los efectos de los múltiples aspectos que constituyen una fuente potencial de error en la medición de los resultados (Martínez-Arias, 1995). Si se logra estimar la contribución de cada uno de estos aspectos en la puntuación de la prueba, se obtiene un coeficiente de la generalizabilidad que permite extrapolar los resultados de la prueba a otras áreas de producción. Entre las diversas variables que se estudian destacan: la fiabilidad inter-corrector o *rater consistency* en la corrección de ensayos (Lehmann, 1983) y la fiabilidad de los diversos ítems en una prueba (Brown, 1999).

1.6.1.1.3. Teoría de respuesta al ítem (TRI).

Este modelo permite calcular las propiedades estadísticas de los ítems y la de las habilidades de los sujetos. Con ello se consigue minimizar la posible fuente de error que un grupo de sujetos o la forma específica de una prueba determinada pudiera producir. De hecho, la teoría de la respuesta al ítem se basa en el supuesto expresado por Bachman (1990) de

que: “an individual’s expected performance on a particular test question, or item, is a function of both the level of difficulty of the item and the individual’s level of ability” (p. 202).

Uno de los mayores inconvenientes que presenta este modelo para el análisis de los datos de las pruebas es el supuesto de la *unidimensionalidad*. Es decir, la idea de que todos los ítems de la prueba evalúan la misma habilidad. Dicho supuesto resulta problemático porque, como dijimos, hoy en día, el dominio de la lengua se entiende mayoritariamente como una habilidad multidimensional (Bachman 1990; Buck 1994).

La TRI se ha aplicado principalmente a pruebas de dominio de la lengua estandarizadas y se ha utilizado en evaluaciones nacionales e internacionales. El modelo Rash de facetas múltiples es el más utilizado a la hora de investigar los efectos de la multitud de *facetas* que intervienen en la medida de las puntuaciones de las pruebas. Entre estas facetas destacan las referidas a los *correctores* o *raters* y las referidas a las *tareas* (ver McNamara 1990; Boldt 1992; Bachman *et al.* 1995; Lumley y McNamara 1995; Weigle 1998).

La implementación de esta técnica en nuestro trabajo no ha sido posible debido al reducido tamaño de la muestra.

1.6.1.1.4. Structural equation modelling (SEM).

Este modelo investiga la estructura de los factores de las medidas que utilizamos y la relación entre dichos factores y las posibles variables ocultas. Bachman (2000), que aplicó por primera vez esta técnica a LT (Bachman y

Palmer, 1981), nos explica que el SEM también se puede utilizar para investigar relaciones direccionales entre diferentes conjuntos de variables ocultas tanto dependientes como independientes.

1.6.1.2. Métodos de investigación cualitativos

Los métodos de investigación cualitativos van gozando cada vez de mayor prestigio y se utilizan para investigar situaciones diversas como, por ejemplo, el efecto que ejerce la personalidad del sujeto en los resultados de una prueba o el estudio de las estrategias que utilizan los candidatos en sus respuestas.

Entre las diferentes técnicas cualitativas destacan: las observaciones, el uso de entrevistas y cuestionarios y, sobre todo, las informaciones obtenidas a través de métodos como la introspección. La implementación de esta última técnica aporta resultados bastante reveladores (Cohen 1984; Alderson 1988; Storey 1997; Green 1998). Skehan (1991) opina que las técnicas de introspección pueden llegar a convertirse en herramientas de gran utilidad en la validación de los resultados de las pruebas.

Actualmente, se ha superado la visión simplista de la supuesta incompatibilidad entre las técnicas cuantitativas y las cualitativas. Ambas metodologías se conciben como complementarias como nos lo demuestran estudios recientes (Weigle 1994; Banerjee y Luoma 1997; Sasaki 1996).

Este último planteamiento se recoge en nuestro trabajo. En él las técnicas esencialmente cuantitativas se intentan combinar con la utilización

de algunas técnicas cualitativas como es el cuestionario. Ello nos permitirá enriquecer, con toda seguridad, los datos de nuestro estudio.

1.6.2. Nuevas Tecnologías

En este punto, cabe destacar los siguientes avances:

1.6.2.1. Inglés para Fines Específicos

1.6.2.2. La evaluación del vocabulario

1.6.2.3. La evaluación computacional

1.6.2.1. Inglés para Fines Específicos

El planteamiento comunicativo de Munby (1978), especialmente en sus comienzos, ejerció una gran influencia sobre esta disciplina. Ello se tradujo en una mayor especificación en los programas de lengua para la enseñanza.

El área de investigación del Inglés para Fines Específicos se ha extendido a otros campos paralelos en la última década y ha cobrado un renovado interés en el terreno de LT para fines ocupacionales y académicos (McNamara 1990; Read 1990; Brown 1993; Clapham 1993; Alcaraz 2000; Douglas 2000; Davies 2001).

1.6.2.2. La evaluación del vocabulario

A partir de los años 90 se ha despertado un creciente interés por la evaluación del vocabulario tanto en el campo de la investigación como en el desarrollo de pruebas de vocabulario (Read, 2000).

1.6.2.3. La evaluación computacional

Se ha producido un incremento notable en la utilización de la tecnología computacional en el desarrollo de las pruebas evaluadoras (Alderson y Windeatt 1991; Laurier 1991; Gruba y Corbel 1997). Entre las diversas aplicaciones, cabe mencionar el gran interés mostrado por el uso del ordenador en la corrección de pruebas de ítems abiertos (Bennett *et al*. 1990; Henning *et al*. 1993). Asimismo, el *Educational Testig Service* (ETS) ha desarrollado un sistema automatizado para la evaluación de las habilidades productivas de la lengua denominado *e-raters* que permite simular la conducta humana en la evaluación de los ensayos que son objeto de estudio de este trabajo (Burstein y Leacock 2001).

Los aspectos más problemáticos relativos a la implementación de esta tecnología se hallan en la definición precisa de los constructos que se miden (Burstein *et al*. 1996) y en la dificultad de extrapolar los resultados obtenidos.

1.6.3. Investigación de los principales elementos que afectan el desarrollo de LT

Entre las diversas variables que se destacan (ver Bachman, 2000) cabe citar:

1.6.3.1. Las características que enmarcan el desarrollo de las pruebas

1.6.3.2. El proceso de evaluación

1.6.3.3. Las características de los candidatos

1.6.3.1. Las características que enmarcan el desarrollo de la pruebas

Se investiga la influencia de variables como, por ejemplo, la diversidad de ítems, o las distintas tareas y formatos de las pruebas evaluadoras. Cobra especial interés el estudio del corrector. En torno a él se investiga la experiencia laboral (objeto de estudio en este trabajo) y su formación en los procesos evaluativos (Weigle 1994; Lukmani 1996; Weigle 1998). Los resultados de los estudios, en general, indican: “a recognition of the fallibility of testing” (Skehan 1991: 10).

1.6.3.2. El proceso de evaluación

Aquí, se aborda el estudio de las estrategias y los procesos utilizados por los sujetos en la realización de las pruebas. Los métodos principales de investigación son los métodos introspectivos (Buck 1991; Wu 1998) y los cuestionarios.

1.6.3.3. Las características de los candidatos

Las variables más estudiadas son: el nivel académico de los sujetos, la lengua nativa, el género y la influencia de la personalidad del sujeto en el desarrollo de las pruebas orales (Berry, 1993).

1.6.4. *Performance assessment*

Este nuevo movimiento, que surge dentro del campo del *educational measurement*, se encuentra muy ligado a los conceptos de *comunicación* y *autenticidad.* Se presenta como una alternativa a otros métodos de evaluación como las generalizadas y denostadas pruebas de elección múltiple. Dentro del *performance assessment* se incluyen diversos métodos alternativos de evaluación como son: los diarios, las conferencias, los *portfolios* (especialmente, en la producción escrita), etc.

La falta de fiabilidad, validez y el impacto negativo que producen algunas de estas propuestas han sido duramente criticados (Hamayan 1995; Messick 1995; Hamp-Lyons 1996; Brown y Hudson 1998; Clapham 2000). Las denominadas *evaluaciones alternativas* se consideran: “too general and shortsighted” (Brown y Hudson, 1998).

1.6.5. Aspecos éticos

La preocupación por los problemas éticos relativos al uso indebido de los resultados de las pruebas se empieza a constatar a partir de los años 80 y cobra un especial interés en los años 90 (McNamara 1998; Hamp-Lyons 1998). Entre otros aspectos cabe destacar las investigaciones llevadas a cabo en torno al denominado efecto rebote o *washback* que producen las pruebas evaluadoras sobre la enseñanza que le precede; la ética en el uso de las pruebas (Elder 1997; Lynch 1997; Norton y Starfield 1997; Rea-Dickins 1997; Shohamy 2001a, 2001b) y el abuso o uso indebido de las

pruebas por razones sociales y políticas (Spolsky 1981; Hawthorne 1997; Shohamy 1997; Spolsky 1997; Brindley 2001).

La solución a estas cuestiones se pretende encontrar a través del la elaboración del código de la ética (ALTE[7], 1998) y del profesionalismo (Stansfield 1993; Davies 1997).

El estudio de la fiabilidad de las pruebas evaluadoras que se realiza en este trabajo constituye un primer paso para garantizar la aplicación ética de los resultados que se obtienen en las PAAU.

[7]Las siglas ALTE denominan la asociación de evaluadores de la lengua en Europa: *Association of Language Testers in Europe*.

CAPÍTULO 2. LA EVALUACIÓN DE LA LENGUA: EL CONCEPTO DE DOMINIO DE LA LENGUA EXTRANJERA Y SU APLICACIÓN EN EL CONTEXTO ACADÉMICO

2.1. Introducción

En este capítulo se recogen, en primer lugar, las distintas aproximaciones al concepto de LT y las tendencias principales que informan este ámbito. A continuación, se revisa el concepto de dominio de la lengua y se analizan los principales modelos que configuran la nueva visión de dominio de lenguas extranjeras vigente en la actualidad. Dada la importancia que tiene la definición del constructo de expresión escrita en este trabajo, introduciremos este concepto brevemente para pasarlo a examinar con mayor detenimiento en el capítulo 5 (subepígrafe 5.2.1.1.).

2.2. Definición de la evaluación de la lengua o LT

A pesar de que existe una abundante y extensa bibliografía en torno a LT, no resulta fácil encontrar una definición precisa de este concepto. Ni en la obra clásica de Lado *Language Testing*, 1961 ni en las obras de autores posteriores como Valette (1967), Davies (1968, 1982) o Heaton (1982) se

nos ofrece una definición explícita de LT, si bien se da por supuesto lo que dicho concepto implica.

De hecho, y siguiendo lo que parece ser ya una tradición, la gran mayoría de libros recientes sobre LT omiten la explicación del término y, generalmente, tras una breve introducción sobre el estado de la cuestión se mencionan de forma sintética los propósitos que LT persigue.

Probablemente, este último punto constituya un intento por delimitar el campo sobre el que se va a hablar (ver Hughes 1989; Heaton 1990, Bachman y Palmer 1996).

A pesar de ello, algunos autores sugieren ciertos elementos que configuran e informan el ámbito de LT. De este modo, podemos observar que LT se asocia frecuentemente a los conceptos de medida y de establecimiento de un orden dentro de unos parámetros determinados.

Davies (1987), por ejemplo, manifiesta que: “a test is a procedure which establishes a rank order” (p. 140). En su obra posterior, Davies (1990) vuelve a enfatizar la base psicométrica de las pruebas evaluadoras: “A test is a measurement of a characteristic of an individual” (p. 144). Según Bachman y Palmer (1996), la función cuantitativa de las pruebas relega a un segundo plano su función pedagógica: “...the primary purpose of tests is to measure. Tests can serve pedagogical purposes, to be sure, but this is not their *primary* function” (p.19).

Siguiendo esta misma línea, Brown (1988) define una prueba del modo siguiente: “A test, in plain, ordinary words, is a method of measuring a person’s ability or knowledge in a given area.” Posteriormente, este autor

nos resume los componentes esenciales que dicha definición pretende capturar.:

> A test is first a *method*...Next, a test has the purpose of *measuring*... A test measures a person's *ability* or *knowledge* that is, competence... Finally, a test measures *a given area.* (Brown 1988: 219)

Alderson (1990), por su parte, subraya la emisión de juicios y valoraciones durante el desarrollo de LT. Según este autor, las valoraciones son inherentes al propio proceso evaluativo:

> Indeed, language testing is an area of applied linguistics that requires judgements at every level of activity and every stage in test development and validation. Testers have to judge whether test specifications are fit for their purpose, whether test content reflects the test's specifications, whether the test method is appropriate for the test's purpose, whether scoring criteria are appropriate, and whether candidates' performance meet those criteria. Judgements abound in language testing and are inevitable...(p. 46)

La carencia de una definición precisa de LT no impide que esta área de conocimiento disfrute de una entidad propia. Ya en los años 80, Alderson (1983) afirmaba: "We no longer stand before an "abyss of ignorance"'(p.90). Algunos años después, Alderson considera que: "Language testing has "come of Age"' (Alderson, 1995). Los últimos avances en la investigación y estudio de la evaluación van a consolidar esta disciplina y convertirla en una rama independiente de la Lingüística Aplicada.

Alderson (1990) nos propone la siguiente definición:

> Language testing is an area of applied linguistics that combines the exercise of professional judgement about language, learning, and the nature of achievement of language learning with empirical data about students' performances and, by inference, their abilities. (p. 46)

Uno de principales argumentos que justifican la falta de una definición de LT comúnmente aceptada en la literatura de la evaluación es la diversidad de intereses y puntos de vista que defienden los distintos autores. Estos últimos intentan plasmar en LT sus principales preocupaciones dado que: "The test is an operationalisation of one's theory of language, language use and language testing" (Alderson 1981: p. 54). De acuerdo con esta afirmación, la puesta en práctica de la evaluación hace explícita las diversas creencias y teorías sobre la lengua y la enseñanza. De ahí, la falta de acuerdo entre los autores en el establecimiento de una definición común y única de LT que pudiera herir susceptibilidades y conducir a enfrentamientos.

Puede que esta última postura sea la acertada ya que la carencia de una definición de LT le añade a este concepto una mayor flexibilidad y dinamismo haciendo a la vez innecesaria su defensa a ultranza. Siguiendo este razonamiento, se ha optado en este trabajo por la práctica común de utilizar el término sin definirlo. No obstante, los elementos esenciales sobre los que podríamos basar nuestra definición de LT contemplarían: 1) medida

de la habilidad del candidato, 2) emisión de un juicio de valor y 3) responsabilidad y utilización ética de los resultados de las pruebas.

2.3. El dominio de la lengua: su definición

En este trabajo, explicar el concepto de dominio de la lengua inglesa resultará fundamental dado que este concepto se relaciona directamente con la visión que se tenga de la naturaleza y las exigencias de LT.

A medida que el campo de LT ha ido avanzado han surgido varios modelos de dominio de la lengua sobre los que poder asentar y diseñar las diversas pruebas evaluadoras o *proficiency tests*. Aún así, la mayoría de los modelos de dominio de segundas lenguas revelan grandes discrepancias entre ellos. Esto dificulta la llegada a un consenso en la adopción de un modelo único o "the best model" (Alderson 1991: 8) que permita diseñar la prueba evaluadora que defina dicha capacidad del modo más apropiado.

Farhady (1983) asegura que el concepto de dominio de la lengua dentro del ámbito de LT es uno de los más pobremente definidos. En esta misma línea, Chastain (1988) mantiene que: "to date, the profession has no acceptable definition of proficiency." (p.49). Davies (1981), por su parte, adopta una postura más extrema y concluye: "My position, then, on the issue of General Language Proficiency (GLP) is that it is essentially a non-issue theoretically." (p. 182)

La definición de dominio de la lengua ha experimentado cambios notables en sus connotaciones reflejando en cada caso las diversas

actitudes y enfoques que se han mantenido en la enseñanza y la evaluación de una lengua extranjera.

Tradicionalmente, este concepto se ha ligado al de actuación o activación de la competencia (i.e. *performance*) en las pruebas que pretenden medir la competencia lingüística. Esto es, la habilidad para producir frases gramaticalmente correctas debido al conocimiento de las reglas lingüísticas.

El concepto de actuación o *performance* se relaciona directamente con la dicotomía propuesta originalmente por Chomsky (1965) sobre la competencia versus la activación y puesta en funcionamiento de dicha competencia (*competence* vs. *performance*). Esta distinción será recogida en aproximaciones posteriores que se hagan sobre LT. Sin embargo, a partir de la llegada del enfoque comunicativo a finales de los años 70 y principios de los 80, se va a hablar de una doble competencia en el dominio de la lengua: la competencia comunicativa y la competencia lingüística. Según este nuevo enfoque, el uso efectivo de la lengua en situaciones específicas requiere un dominio tanto de la actuación (i.e. *use*) como de la competencia (i.e. *usage*) (Widdowson, 1978).

Siguiendo este último planteamiento, Canale y Swain (1980) distinguen en su estudio dos clases de competencia a las que denominan: competencia comunicativa y actuación comunicativa. Esta nueva visión del concepto de dominio de la lengua se enmarca siempre dentro de un contexto específico y se relaciona con unos propósitos y necesidades determinadas. En palabras de Kelly (1978): “...a test constructed to measure

a candidate's ability to use the language of interest (the 'target' language) in certain specified communication situations."

El nuevo enfoque sociolingüístico-comunicativo que aborda el estudio de la habilidad del uso de la lengua dentro de un contexto específico y válido es el que precisamente va a permitir la determinación del constructo de expresión escrita en los distintos contextos académicos que se valoran en cada momento y entre los cuales se enmarca este trabajo.

2.4. Modelos de dominio de la lengua

En este epígrafe introduciremos los modelos de dominio de la lengua que más honda influencia han ejercido en las investigaciones llevadas a cabo en estas dos últimas décadas. En los subepígrafes 2.4.1. y 2.4.2. examinaremos en profundidad los dos tipos de modelos más relevantes en nuestro estudio.

Chalhoub-Deville (1997) nos ofrece en su trabajo una interesante síntesis sobre los principales modelos de dominio de la lengua de estos últimos tiempos. Siguiendo la división establecida por Stern (1983), Chalhoub-Deville (1997) distingue entre los modelos de componentes (*componential models*) y los modelos de niveles de dominio lingüístico (*levels of proficiency*). Esta última autora afirma que el propósito fundamental que persiguen los modelos de componentes es la descripción de los distintos elementos que componen la habilidad de dominio de la lengua. Por el contrario, el interés fundamental de los modelos de niveles de

dominio lingüístico reside en la representación progresiva de la capacidad o habilidad de dominio de la lengua en diversas etapas sucesivas.

2.4.1. Modelos de componentes de dominio de la lengua

La mayoría de las teorías de dominio de la lengua, contrariamente a las diversas visiones no teóricas, se basan en el concepto de competencia. Es decir, en el conocimiento o conjunto de conocimientos subyacente en los individuos y que se manifiesta en el comportamiento verbal a un nivel de actuación. Dentro de los modelos de componentes de dominio de la lengua nos interesa principalmente el modelo de competencia unitaria (Oller, 1976), que avanzamos brevemente en el capítulo 1, y el modelo de competencia comunicativa que define las principales directrices que informan la prueba de Inglés de las PAAU.

Como ya dijimos, Oller (1979) establece que la estructura del dominio de la lengua podría interpretarse a través de un único factor general o *competencia unitaria* (*Unitary Competence Hipótesis*, (UCH)). Esta última determina que:

> ...there will be reliable variance shared by all of the tests and essentially no unique variance shared by tests that purport to measure a particular skill, component, or aspect of language proficiency. (Oller 1979: 425)

No obstante, el modelo UCH fue duramente criticado en su momento tanto desde el punto de vista teórico como desde el punto de vista

metodológico. Algunas autores sugieren que los resultados obtenidos por Oller (1979) se pueden atribuir a la propia naturaleza de los datos sobre los que experimentó. Hughes (1981), por ejemplo, considera que varias de las pruebas que se usaron en el estudio de Oller (I1979) no estaban midiendo realmente el constructo de dominio de la lengua que se suponía debían medir. Asimismo, Chalhoub-Deville (1997) rechaza la técnica del análisis de factores de los principales componentes del estudio utilizada por Oller (1979). Según nos explica Chalhoub-Deville (1997), esta última técnica no divide los diferentes tipos de varianza, esto es, la varianza común, la varianza específica de la prueba y la varianza del error. Debido a ello, la significación del primer factor derivado se acaba, normalmente, sobreestimando.

En estudios posteriores, el mismo Oller (1983) ha rectificado su postura y ha admitido que: “the strongest form of the unitary hypotheis was wrong” (p. 352). En línea con el pensamiento actual, este autor también reconoce que el factor general que representa el dominio de la lengua en segundas lenguas puede ser dividido en diferentes componentes analíticos (Cummins 1979; Bachman y Palmer 1982; Farhady 1983).

Con respecto a este último punto de vista, cabe decir que, hoy en día existe un cierto consenso en cuanto a la aceptación de la idea propuesta por Farhady (1983) que sugiere que, probablemente, la habilidad de dominio de la lengua está compuesta por un factor general y por otro grupo de factores más específicos. No obstante, todavía queda mucho trabajo pendiente en este terreno si bien la idea de enfocar la enseñanza y la evaluación

basándose en la habilidad de un dominio de la lengua parcialmente divisible resulta defendible. Esta última postura, implica el tener que validar por separado cada una de las partes que se plantean en las diferentes situaciones de enseñanza y de evaluación concretas. Así, el ejercicio del ensayo que se incluye en la prueba de Inglés de Selectividad tiene como principal objetivo medir la habilidad de producción escrita del candidato en lengua inglesa.

Pero sin duda, el modelo de dominio de la lengua basado en el estudio de sus diferentes componentes que mayor influencia ha ejercido en el campo de la enseñanza y de la evaluación de segundas lenguas es el modelo de competencia comunicativa. Como ya avanzamos (ver capítulo 1), este modelo, propuesto por Canale y Swain (1980), distingue tres tipos de competencia: la competencia gramatical, la competencia sociolingüística y la competencia estratégica. Posteriormente, Canale (1983) incluyó una nueva categoría al distinguir entre la competencia sociolingüística y la competencia discursiva. A pesar de que dicho modelo no ha sido validado empíricamente, su influencia ha sido decisiva y ha servido de inspiración a otros modelos posteriores. Entre ellos destaca el modelo propuesto por Bachman (1990) denominado c*ommunicative Language Ability* (CLA) sobre el que se fundamenta el diseño de la prueba de Inglés de Selectividad.

El modelo propuesto por Bachman (1990), a diferencia de los modelos anteriores, se encuentra avalado por estudios teóricos y empíricos. Bachman (1990) incluye tres componentes que interactúan en su modelo: competencia lingüística, competencia estratégica y mecanismos

psicofisiológicos. El modelo CLA es uno de los más elaborados y ha sido calificado como “the best that we have at present” (Skehan 1991: 15).

El diseño de la prueba de Inglés de las PAAU se basa en uno de los componentes del modelo CLA propuesto por Bachman (1990). Este componente es el de la competencia lingüística. El principal componente de dicha competencia es el de la competencia organizativa, que incluye la competencia gramatical y la competencia textual, las cuales se subdividen en gramática, léxico, comprensión lectora y composición para dar una descripción más detallada del constructo. Dentro de este marco operativo, nos interesa centrarnos en el proceso de composición.

2.4.2. Niveles de dominio de la lengua

Algunos autores favorecen la utilización de escalas de valoración para representar los diversos niveles de dominio de la lengua de forma gradual y secuencial en el desarrollo de las pruebas. La escala de valoración más utilizada y extendida dentro del contexto académico son las *ACTFL*[1] *Guidelines*. Esta escala intenta describir los diferentes niveles de dominio de las cuatro destrezas (i.e. comprensión escrita, expresión escrita, expresión oral y comprensión oral) distribuyéndolos en nueve bandas y a lo largo de una única escala que parte del nivel *Novice-Low* y alcanza el nivel *Superior*. Alderson (1991) nos explica que la utilización de escalas de valoración como las *ACTFL Guidelines* sirve tres propósitos fundamentales:

> There are thus three distinguishable purposes of scales, which I have called **user-oriented**, **assessor-oriented** and **constructor-oriented**. The functions they serve are, respectively, reporting results, guiding the rating process, and controlling test construction. (p. 74)

A pesar de que el uso de estas escalas puede resultar atractivo, Chalhoub-Deville (1997) nos recuerda que estas últimas y, en concreto, las *ACTFL guidelines*, han sido objeto de duras críticas. Bachman (1988) y North (1993), por ejemplo, cuestionan su validez y critican la falta de estudios empíricos que respalden las diferentes bandas y niveles de dominio lingüístico que las conforman. Hill (1991), por su parte, argumenta que la totalidad de las bandas que se incluyen en las escalas y que definen la destreza general o dominio de la lengua no puede contemplarse en una única prueba: "for the simple reason that the test would have to be far too long" (Hill 1991: 239).

Una de las soluciones propuestas para resolver este dilema es la que opta por la construcción de pruebas que se centran en un punto concreto de la escala imaginaria que representa el dominio de la lengua. Esta aproximación enlaza con la idea que defiende Vollmer (1981) y que contempla el dominio de la lengua como un proceso dinámico. Según Vollmer (1981): "Testing language proficiency means making a cut at a given point in time in order to form a more or less rough idea of a person's state of advancement" (p. 164).

---

[1] Las siglas ACTFL se utilizan para referirse a: *The American council for the teaching of*

La gran diversidad de teorías y modelos de dominio de la lengua extranjera que se han desarrollado evidencian la dificultad que supone obtener una definición del constructo de dominio de la lengua que se acepte de forma unánime. Hoy en día, es prácticamente impensable concebir un modelo que se adecue perfectamente a todos los contextos. Chalhoub-Deville (1997), prudentemente, nos aconseja mantener los diferentes papeles que juegan, por una parte, los modelos teóricos, que describirían la naturaleza del dominio de una segunda lengua a un nivel general, y por otra parte, los modelos funcionales, que nos describirían el constructo enmarcándolo en un contexto específico. Para ello, sería necesario establecer unos marcos o puntos de referencia (*proficiency assessment frameworks*) que se desarrollarían en contextos apropiados y de acuerdo con el modelo teórico de dominio de la lengua que mejor representara el estado de la cuestión en el campo de investigación pertinente. La función de estos marcos evaluativos es la siguiente:

> ...(assessment frameworks) do not provide all-inclusive representations of the proficiency of the construct. Instead these frameworks emphasize specific aspects of that more general theoretical construct to reflect the characteristics of their specific contexts. (Chalhoub-Deville 1997: 15)

De acuerdo con este planteamiento, el establecimiento del constructo de expresión escrita que nos ocupa en este trabajo debería reflejar las características específicas del contexto académico en el que nos hallamos

---

foreign languages.

(i.e. acceso a los estudios universitarios). Asimismo, cabe subrayar que las profundas implicaciones sociales de las pruebas de dominio de la lengua como son las PAAU, objeto de estudio en este trabajo, recomiendan la necesidad de definir constructos de dominio de la lengua que sean válidos. En una interesante cita, Vollmer (1981) expone al respecto:

> If there is any doubt or any considerable degree of uncertainty as to what proficiency (in theoretical terms) really is or what proficiency tests really measure it would be irresponsible, in my opinion, to continue to use the construct (as if it were well defined) or to administer any proficiency measure (as if were sure about its validity) and use the test scores to make more or less irreversible decisions (p. 168)

Este autor nos advierte de la responsabilidad que supone tomar decisiones a partir de los resultados de pruebas como las citadas anteriormente. En nuestro caso concreto, los resultados de las PAAU van a afectar seriamente la vida de los candidatos ya que condicionarán la elección de su futura carrera universitaria:

> I would seriously question, however, many a case in which proficiency tests are being exploited for placement purposes or, even worse, career decisions to be based upon them. We should not take it for granted that proficiency testing is done worldwide: each single situation in which such cutting decisions on the basis of proficiency scores are said to be necessary should be questioned and the procedures applied should be publicly called for justification over and over again. We as a society on the whole simply cannot afford to classify people and divide them up (allotting educational and professional chances of different kinds) as long as the question of construct validity of the instruments used are not clarified somewhat further. This is especially true with the concept and measures of GLP (General Language Proficiency). (Vollmer 1981: 168-169)

CAPÍTULO 3. LAS PRUEBAS DE EXPRESIÓN ESCRITA

3.1. Introducción

En este capítulo se intentará establecer, en primer lugar, una definición de las pruebas de expresión escrita. A continuación, se examinará con profundidad el contexto académico que nos ocupa y que delimita las características básicas del ejercicio del ensayo que se incluye en las prueba de Inglés de Selectividad.

3.2. Definición de las pruebas de expresión escrita

En este trabajo hemos decidido adoptar la definición de una prueba de expresión escrita defendida por Hamp-Lyons (1991b). Así pues, al hablar de una prueba de expresión escrita o *writing test*: “we mean only tests that test writing through the production of writing”. Este tipo de prueba directa se puede caracterizar por, al menos, cinco aspectos básicos:

- First, each individual taking the assessment must actually, physically write at least one piece of continuous text of 100 words or longer (100 words being widely regarded as a minimum sample) and may write several pieces and/or considerably longer pieces.
- Second, while the writer is provided with a set of instructions and a text, picture, or other “prompt” material, she or he is given considerable room within which to create a response to the prompt.
- Third, every text written by a candidate is read at least one, usually two or more, human reader-judges who has been through some form of preparation or training for the essay evaluation process.

- Fourth, the judgements made by readers are tied in some way, tightly or loosely, to some common yardstick, such as a set of sample essays, a description of expected performance at certain levels or one or several rating scales.
- Fifth, the readers' responses to the writing are expressed as a number or numbers of some kind, instead of or in addition to written or verbal comments; scores on the test are recorded and can be retrieved for review by higher or external authority as needed. (ver Hamp-Lyons 1991b: 5-6)

Dichas características enmarcan casi perfectamente el tipo de habilidades y el constructo de expresión escrita que se pretende medir en el ejercicio del ensayo que se incluye en la prueba de Inglés de Selectividad. El único criterio que señala Hamp-Lyons (1991b) en su definición de pruebas de expresión escrita que, desafortunadamente, no es aplicable a la evaluación formal del ensayo en el contexto de las pruebas de Selectividad, es el criterio número tres concerniente a la formación de los correctores. Como se verá el capítulo 7 (subepígrafe 7.1.4.5.), esta última técnica utilizada con el propósito de asegurar y garantizar la fiabilidad de las puntuaciones asignadas por los correctores no se contempla en el desarrollo de las PAAU.

3.3. La producción escrita en el contexto académico

En este epígrafe se realizará una breve introducción al contexto académico. Posteriormente, en los subepígrafes siguientes entraremos en profundidad en el estudio de las principales características de la producción escrita. Destacaremos su relación con el aprendizaje y el papel que desempeña el profesor. A continuación, se establecerán las tendencias que

definen la escritura académica. Por último, se analizarán algunas de las taxonomías que nos permiten enmarcar el ejercicio del ensayo dentro de la producción escrita académica.

3.3.1. Introducción

Redactar un texto constituye una tarea ardua y compleja tanto para los hablantes nativos como para los hablantes no nativos. La gran multiplicidad de destrezas que intervienen en la producción escrita dificulta el dominio de dicha habilidad y representa un reto para los hablantes de cualquier lengua. Autores como Lázaro (1996) o Alcaraz (2000) señalan que la producción escrita es, a menudo, la habilidad que mayores dificultades de adquisición plantea.

Sin embargo, se ha de reconocer que los estudiantes de segundas lenguas afrontan un doble reto a la hora de producir un texto escrito. Kroll (1990b) argumenta:

> As teachers, we must realize that for those engaged in learning to write in a second language, the complexity of mastering writing skills is compounded both by the difficulties inherent in learning a second language and by the way in which first language literacy skills may transfer to or detract from the acquisition of second language skills. (p. 2)

3.3.2. Características básicas de la producción escrita

Rosen (1981) nos sintetiza algunos de los aspectos básicos que definen la producción escrita en la siguiente cita:

> The writer is a lonely figure cut off from the stimulus and corrective of listeners. He must be a predictor of reactions and act on his predictions. He writes with one hand tied behind his back, being robbed of gesture. He is robbed too of the tone of his voice and the aid of clues the environment provides. He is condemned to monologue; there is no one to help out, to fill the silences, put words in his mouth, or make encouraging noises (Rosen 1981)

Como sabemos, las diferencias entre la lengua oral y la lengua escrita son numerosas. La producción escrita carece, por ejemplo, de los recursos paralingüísticos y suprasegmentales presentes en cualquier intercambio oral como son: los gestos, el movimiento corporal, el tono, el acento, el ritmo, etc. El escritor al no contar con la presencia física del receptor, carece de sistema de retroalimentación o *feedback* y, por consiguiente, no puede clarificar las ideas o revisarlas a lo largo del proceso.

Entre las muchas cualidades distintivas que caracterizan el discurso escrito *versus* el discurso oral, podemos destacar:

- una mayor organización en el desarrollo de los argumentos y exposición de ideas;
- una mayor precisión en la construcción de las frases con el fin de evitar cualquier tipo de ambigüedad;
- el uso de estructuras gramaticales más complejas;
- la selección de un vocabulario específico de mayor densidad léxica;
- y la creación de un estilo que se adecue a las exigencias del contexto en el que se enmarca.

Este último elemento, es decir, el contexto será, precisamente, el que determinará la utilización de unas convenciones retóricas específicas y características de los diversos tipos de texto de producción escrita (i.e. la descripción, la narración, etc.).

En este sentido, se ha de decir que el ejercicio del ensayo de la prueba de Inglés de Selectividad se basa en las funciones comunicativas y retóricas necesarias para la escritura académica (i.e. no especializada) a nivel general. La selección del contenido de los temas es de carácter neutral o semi-técnico. No por ello, la tarea resulta poco compleja dado que, tanto el nivel de formalidad exigido en los textos como el desarrollo de las estructuras retóricas, gramaticales y discursivas que se utilizan en la presentación de la información escrita son en ocasiones muy distintos en ambas lenguas. Esto es, en la lengua nativa y en la segunda lengua.

3.3.3. La relación entre la producción escrita y el aprendizaje: el papel del profesor

Los defensores de la inclusión de las tareas de producción escrita dentro de los *curricula* de enseñanza señalan la bien documentada y estrecha relación que existe entre el proceso de producción escrita y el aprendizaje. En opinion de Zamel (1982), la producción escrita implica: “exploring one’s thoughts and learning from the act of writing itself what these thoughts are.” (p.197).

El propio proceso lineal que conlleva la producción final de una pieza escrita requiere que el escritor organice y ordene sus pensamientos. En este sentido, la escritura desempeña un papel fundamental en el esclarecimiento y expansión del pensamiento. Hamp-Lyons (1990) vincula esta idea al resurgimiento de las pruebas de producción escrita directas en los colegios y universidades de Norte América en la década de los años 90. Esta autora afirma que las consecuencias negativas que se derivaron de la exclusión de las pruebas de producción escrita directas motivó posteriormente y, como contrapartida, el cambio y la desaparición de las pruebas de expresión escrita indirectas (i.e. las prueba de elección múltiple, la prueba C, etc):

> This change is the result of social pressure from schools, colleges, and parents, who argued that failure to learn and practice writing reasonable lengths of text in school was leading to declining literacy levels and to a college-entry population that could not think critically about intellectual ideas and academic material. (Hamp-Lyons 1990: 69)

Por otra parte, el hecho de que la enseñanza de la producción escrita se conciba como un elemento que propicia el aprendizaje no debe confundirse y utilizarse como pretexto para conseguir únicamente unos objetivos gramaticales concretos a nivel superficial o para consolidar el vocabulario aprendido. Hedge (1988) declara al respecto:

> ...but successful writing depends on more than the ability to produce clear and correct sentences. I am interested in tasks which help students to write whole pieces of communication, to link and develop information, ideas, or arguments for a particular reader or group of readers. (p. 8)

De hecho, al papel que ha de desempeñar el profesor en el desarrollo de la producción de la escritura académica así como la determinación del tipo de discurso que se ha de estimular en las clases de composición ha suscitado polémicas diversas. En general, se considera que los profesores deben mostrar a sus estudiantes los diversos formatos de escritura (i.e. ensayos, cartas, etc.) que se van a utilizar en el aula. Asimismo, se han de estudiar sus principales funciones (i.e. descripción, narración, etc.) y las características y estructuras organizativas que facilitan el proceso comunicativo y definen los distintos tipos de texto.

Kroll (1990c) resalta la dificultad que representa esta tarea para los estudiantes de segundas lenguas:

> For English as a second language (ESL) students, it seems fair to say that writing academic papers is particularly difficult. ESL students must learn to create written products that demonstrate mastery over contextually appropriate formats for the rhetorical presentation of ideas as well as mastery in all areas of language, a Herculean task given the possibilities for error (p. 140)

3.3.4. Escritura académica *versus* escritura personal: la composición o ensayo

Dentro de la comunidad universitaria y, durante estos últimos años, se ha hecho extensiva la diferencia entre dos posturas que abordan el planteamiento de los distintos tipos de tareas de producción escrita que se han de requerir a los estudiantes. Por una parte, se encuentran aquellos

profesores partidarios de animar al alumnado para que se involucre personalmente en las tareas de expresión escrita. Por otro lado, se hallan aquellos profesores que defienden que los estudiantes han de producir un discurso escrito de carácter académico, expositivo, neutral y objetivo.

En opinión de Spack (1988), la definición de escritura académica no se ha llegado a determinar de forma satisfactoria. No obstante, este tipo de escritura se enmarca dentro de un contexto específico en el que existen unas convenciones establecidas que deben respetarse y unas expectativas que se han de cumplir. Hamp-Lyons (1991e) considera que casi toda la producción escrita académica en el ámbito universitario puede entenderse como:

> ... a discourse exchange (Coulthard & Ashby, 1975), in which the *initiation* comes from the instructor, the student makes a *response*, and the instructor provides a *follow-up.* (p. 128)

Algunos autores como Brooks (1980) establecen una distinción entre escritura académica y escritura personal. Este autor compara el ensayo de tipo creativo y personal frente al ensayo académico. Según Brooks (1980), el ensayo académico puede definirse como:

> …a public, impersonal essay which is expected to be largely factual and instrumental...This (…) essay has a much more clearly defined content and a more formal structure. It is usually designed to test the candidate's ability to present the appropriate information clearly and concisely. (p.4)

No obstante, la distinción entre escritura académica y escritura personal puede parecer algo simplista. Así, Mlynarcyk (1992) argumenta que los aspectos personales pueden mostrarse e incluirse en los ensayos académicos tradicionales. De hecho, no tener en cuenta la experiencia personal de los estudiantes puede incentivar una producción escrita demasiado académica y formulista e incluso propiciar lo que Neel (1988) define, humorísticamente, como el *antiwriting*:

> I am not writing. I hold no position. I have nothing at all to do with discovery, communication, or persuasion. I care nothing about the truth. What I *am* is an essay. I announce my beginning, my parts, my ending, and the links between them. (Neel 1988: 85)

En el ejercicio del ensayo de la prueba de Inglés de Selectividad, a pesar de que el tema se encuentra normalmente vinculado a un texto de lectura, se involucra claramente al estudiante apelando a experiencias personales vividas, opiniones subjetivas, etc.

Cuando se plantea la necesidad de evaluar la producción escrita académica se asume, normalmente, que existe un cuerpo de conocimientos común de escritura académica. Este último permite, a través de una única prueba o tarea, predecir la habilidad del estudiante para acometer dicha tarea en el contexto universitario (ver Johnson, 1981). No obstante, algunos autores cuestionan este supuesto y ponen en duda la validez predictiva de las evaluaciones de producción escrita que sirven cualquier propósito (*all-purpose writing assessments*) (Hamp-Lyons 1991e; Horowitz 1991). Por el

contrario, estos autores defienden la idea de categorizar las diferentes tareas de producción escrita en sus respectivos géneros lingüísticos (Horowitz, 1991). Asimismo, se propone la construcción de pruebas que incluyan un contenido académico específico ya que la habilidad de producción escrita se halla, generalmente, muy vinculada a la tarea (Hamp-Lyons 1991e; Horowitz 1991).

Esta nueva visión de la producción escrita, más disciplinaria y específica, ha suscitado cierta polémica. McEldowney (1976), por ejemplo, considera que los temas de las pruebas de producción escrita académica han de ser de naturaleza semi-técnica dirigidos al estudiante corriente y no al especialista:

> It is not the purpose of the English test to assess the candidate's knowledge of any subject area but to see how he can handle straight-forward ideas in appropriate English. (McEldowney 1976: 13)

El debate en torno a la utilización de pruebas de tema libre frente a las de contenido específico pone en evidencia los distintos puntos de vista que se establecen entre dos tipos concretos de prueba: las pruebas de rendimiento y las pruebas de dominio lingüístico. Las pruebas de rendimiento o *achievement tests* evalúan el dominio de la materia estudiada. Por su parte, las pruebas de dominio lingüístico (*proficiency tests*) evalúan el dominio real de la habilidad de producción escrita.

Este último tipo de pruebas suelen ser de carácter normativo ya que, al contrario de lo que sucede con las pruebas criteriales, se intenta relacionar la actuación de un candidato con la del resto de los candidatos de modo que ambas actuaciones puedan compararse (Brown, 1988). Su objetivo es conseguir altos niveles de discriminación entre los candidatos y repartir las puntuaciones a lo largo de una curva de distribución normal (ver Gipps 1994; Herrera 1999). Asimismo, cabe añadir que únicamente las pruebas de dominio de la lengua de carácter normativo permiten la estimación de la fiabilidad a través de los métodos de la teoría clásica de la evaluación. Ello explica, en parte, su uso mayoritario en las pruebas nacionales e internacionales de expresión escrita (i.e. *Test of Written English*, (TWE), etc.) entre las que se incluyen las PAAU.

3.3.5. Taxonomías de producción escrita

La producción de taxonomías de escritura sobre las que fundamentar el diseño de los distintos cursos académicos se recoge extensamente en la literatura. Así, Johnson (1981) propone tener en cuenta dos tipos de parámetros en el análisis del discurso de la producción escrita. En primer lugar, se han de considerar las funciones comunicativas citadas por Munby (1978) y basadas en Wilkins (1973, 1976) como son: causa y efecto, clasificación, etc. En segundo lugar, se han de analizar las funciones retóricas que evidencian la relación que se establece entre las diferentes frases dentro del discurso (i.e. organización, introducción, etc). A través de

este esquema, se tratarán de identificar qué funciones comunicativas son características de cierto tipo de discurso.

En este sentido, la publicación de la obra *Communicate in Writing* (Jonson, 1981) representa un intento por relacionar las funciones comunicativas con las funciones retóricas que necesitarán los estudiantes en sus distintas disciplinas. Jonson (1981) sintetiza estas funciones en tres procesos básicos:

1. Descripción de cosas e ideas
2. Descripción de procesos y sucesos
3. Desarrollo de un argumento

Weir (1993), basándose en el esquema anterior, investiga las distintas habilidades o estrategias que han de desarrollar los estudiantes en las tareas de producción escrita académicas. Este autor nos resume las principales operaciones que se llevan a cabo en el marco de la escritura académica en la siguiente tabla (Tabla 3.1.):

Tabla 3.1. Resumen de las operaciones que se realizan en la escritura académica (Weir, 1993)

1. **Describing Phenomena and Ideas** which might involve:
Definition
Classification
Identification
Comparison and Contrast
Exemplification
Summary

2. **Describing process** which might involve:
Purpose
Describing means, results, process, change of state
Sequential description
Instructions
Summary

3. **Argumentation** which might involve:
Stating a proposition
Stating assumptions
Induction
Deduction
Substantiation
Concession
Summary
Generalisation
Speculation/comment/evaluation

Como puede apreciarse, la gran diversidad de operaciones y estrategias de producción escrita que se estudian dentro del mundo académico imposibilita el diseño de pruebas como las de Selectividad que las incluyan en su totalidad. Esta preocupación fue la que motivó el análisis de necesidades llevado a cabo por Bridgeman y Carlson (1984). Su estudio permitió obtener dos tipos de tareas representativas de la producción escrita académica de los estudiantes. Dichas tareas, que fueron las que finalmente

se incluyeron en el *Test of Written English* (TWE), son: *comparar, contrastar y defender una postura* y *describir e interpretar un gráfico o escala.*

Aparte de las distintas operaciones y estrategias, cabe destacar la gran variedad de tipos de texto y formatos que se utilizan en la evaluación formal de la producción escrita en el contexto académico: producción de informes, resúmenes, redacción de notas y ensayos e incluso tesis y proyectos en el caso de los estudiantes de postgrado (ver Bernárdez 1995). De entre todas estas técnicas, el ejercicio del ensayo de respuesta libre en condiciones de tiempo limitadas es, quizás, la técnica más utilizada en el mundo académico tanto en pruebas nacionales como en las internacionales.

CAPÍTULO 4. ENSAYOS EN SEGUNDAS LENGUAS: EVOLUCIÓN Y DESARROLLO

4.1. Introducción

En este capítulo se recoge la evolución histórica de las principales investigaciones realizadas sobre la enseñanza de la producción escrita. Se analizan, especialmente, los estudios llevados a cabo en torno al ensayo en el contexto de segundas lenguas. El capítulo se cierra con una síntesis de los principales aspectos y orientaciones que definen el enfoque actual sobre el que se basa la enseñanza y la evaluación del ensayo.

Desde 1945, la evolución histórica de los ensayos en segundas lenguas se encuentra estrechamente vinculada a los enfoques y orientaciones existentes sobre la enseñanza de la producción escrita en segundas lenguas. No obstante, los especialistas de este campo van a dejarse guiar por la investigación llevada a cabo en los ensayos de primeras lenguas. Estos últimos cuentan con una tradición que data desde principios del siglo anterior. Como se verá, muchas de las investigaciones realizadas sobre los ensayos de segundas lenguas se han basado en estudios sobre la producción escrita de hablantes nativos.

Dicho planteamiento puede invalidar los resultados de aquellos estudios que no contemplen las diferencias contextuales de ambas lenguas: L1 y L2[1]. En este sentido, no cabe olvidar que el contexto único en el que se desarrollan los procesos de producción escrita en segundas lenguas determinarán enfoques y aproximaciones distintas al de los estudios en primeras lenguas.

Silva (1990) establece cuatro aproximaciones cronológicas a partir de 1945 que nos permiten trazar la historia evolutiva de la enseñanza e investigación de la producción escrita en segundas lenguas. Estas aproximaciones se basan en una serie de enfoques sobre la enseñanza de la producción escrita que nos ayudan a configurar la visión actual del ensayo desde el punto de vista académico. Los enfoques que se destacan son los siguientes: el ensayo guiado, la retórica tradicional, el enfoque basado en el proceso y el inglés para fines académicos.

4.2. El ensayo guiado

El ensayo guiado o *controlled composition*[2] se encuentra muy ligado a la figura de Charles Fries[3] precursor del método audiolingual para la enseñanza de segundas lenguas. Este método primaba el desarrollo del habla o lengua oral y relegaba la producción escrita a un segundo plano, de

[1] Las abreviaturas L1 y L2 forman parte de la terminología común utilizada para referirse a la lengua nativa (L1) y a la segunda lengua (L2).
[2] Otra posible denominación en lengua inglesa es la de *guided composition*
[3] Según Fries, célebre lingüista estructuralista, los materiales didácticos más efectivos para la enseñanza de los idiomas son aquellos construidos sobre la base de una descripción científica de la lengua objeto que se ha de aprender, la cual se compara con una descripción científica paralela de la lengua materna del estudiante.

modo que, dicha destreza se veía como un simple instrumento de refuerzo en la adquisición de hábitos orales.

De acuerdo con este planteamiento, la enseñanza de la producción escrita se limitaba a los ejercicios de combinación de frases (O'Hare, 1973), que permitían al estudiante explorar las distintas opciones sintácticas, y a la utilización de pasajes de discurso conectado que ofrecían al estudiante la posibilidad de manipular las formas lingüísticas dentro del propio texto (Kunz 1972; Paulston y Dykstra 1973). Algunos ejemplos de estos ejercicios son: los ejercicios mecánicos (i.e. *drills*, *fill-in*), las transformaciones o las sustituciones.

La filosofía sobre la que se fundamentaba este enfoque era que la producción escrita reforzaba y evaluaba la aplicación correcta de reglas gramaticales. Algunos autores como Erazmus (1960) o Brière (1966) defendieron que los ejercicios de escritura podían convertirse en ensayos de producción libre destinados a promover una mayor fluidez en la habilidad de la expresión escrita. Sin embargo, esta idea fue desestimada por la mayoría de autores (ver Silva, 1990).

De forma sintética, se puede afirmar que el objetivo principal de las tareas del ensayo guiado era el logro de la corrección formal de la lengua, cuyo aprendizaje se entiende como un proceso de formación de hábitos destinados a evitar los errores principalmente producidos por la interferencia de la primera lengua.

4.3. La retórica tradicional

El creciente interés por la producción de textos escritos más extensos que den respuesta a las necesidades de los estudiantes de segundas lenguas va a desembocar en la primera versión de la denominada retórica tradicional. Este enfoque combina los principios básicos del paradigma tradicional aplicado a la enseñanza del ensayo en primeras lenguas con la teoría de Kaplan sobre retórica contrastiva.

El artículo de Kaplan (1967) introdujo el concepto de retórica contrastiva con el propósito de explicar la estructura de pensamiento en lengua inglesa. Dicha lengua presentaba un desarrollo predominantemente linear frente a las estructuras formales de los párrafos que exhibían otras lenguas y culturas. Kaplan (1967) definía la retórica como: “the method of organizing syntactic units into larger patterns” (p. 15). Debido a ello, su principal preocupación se centraba en la construcción lógica de las formas discursivas. De ahí, también, sus dos focos de interés primordiales: el párrafo y el desarrollo de la composición o ensayo. Cabe mencionar, que dentro de los modelos de organización y desarrollo de los párrafos (i.e. narración, descripción, exposición y argumentación), el modelo de la exposición era el que se consideraba más apropiado para utilizar dentro del contexto universitario con los estudiantes de segundas lenguas (ver Silva, 1990).

Como se puede apreciar, en este nuevo enfoque, la forma retórica vuelve a dominar y centrar el interés de la enseñanza de la expresión escrita. Esto va a suponer la utilización de técnicas que enfatizan la imitación

de párrafos o de modelos de ensayo, la producción escrita a partir de esquemas y la ordenación de párrafos en secuencias lógicas (Kaplan y Shaw, 1983).

Asimismo, las investigaciones sobre los aspectos formales de la lengua van a sentar la base de estudios posteriores sobre la producción escrita en segundas lenguas. A modo de ejemplo, cabe citar los trabajos de Reid (1990) y su análisis contrastivo sobre el número de pasivas y el número de pronombres en diferentes lenguas, diversos estudios sobre la forma que presentan los ensayos en varias lenguas (Eggington, 1987; Hinds, 1987) o varios estudios sobre la cohesión y la coherencia (ver Connor, 1984; Johns, 1984) entre otros.

4.4. El enfoque basado en el proceso

La oposición manifiesta de los profesores e investigadores a la puesta en práctica de una enseñanza basada en el enfoque exclusivamente formal de la lengua que negaba el pensamiento crítico y el desarrollo de las ideas originó un nuevo movimiento de enseñanza de la producción escrita. Dicho movimiento se inspiró, al igual que sucediera en ocasiones anteriores, por las investigaciones llevadas a cabo en el desarrollo del ensayo en primeras lenguas. De este modo, se introdujeron nuevos conceptos como son: *proceso*, *significación*, *invención* y *múltiples borradores* (Raimes, 1991).

El enfoque basado en el proceso va a centrar su atención en la figura del escritor como creador de ideas. El proceso de elaboración y producción del ensayo se entenderá como: “a non-linear, exploratory, and generative

process whereby writers discover and reformulate their ideas as they attempt to approximate meaning." (Zamel 1983: 165). Desde esta perspectiva, la forma queda supeditada al contenido y a la expresión de ideas.

Dentro del enfoque basado en el proceso, Faigley (1986) distingue dos corrientes: el expresivismo y el cognitivismo. El expresivismo se originó en las primeras décadas del siglo XX y alcanzó su cumbre a finales de los años 60 y a principios de los 70. Según nos explica Johns (1990), la expresión del pensamiento personal cobró relevancia y la producción escrita enfatizó el proceso más que el producto en una experiencia que se entendía como: "a discovery of the true self" (Berlin, 1982). Las tareas en el aula iban destinadas a promover el desarrollo personal y se utilizaban los ensayos personales y el uso de periódicos para la consecución de dicho fin.

Sin embargo, fue el grupo de los cognitivistas el que ejerció una mayor influencia en las investigaciones llevadas a cabo en la enseñanza de segundas lenguas. La aproximación de la enseñanza como proceso entiende la producción escrita como un proceso complejo, creativo y recursivo que es semejante en los estudiantes de primeras y segundas lenguas (Jacobs 1982; Hayes y Flower 1983; Zamel 1983; Raimes 1987, etc.). Según nos explican Flower y Hayes (1981) el desarrollo de la producción escrita de los estudiantes nativos ingleses se manifiesta a través de una serie de procesos continuos a lo largo de una tarea, de modo, que estos no empiezan y terminan en un solo borrador. Friendlander (1990) señala al respecto:

> Traditional approaches to writing, such as modes of discourse or grammar-based approaches, falter because they do not help students to see writing as an evolving process. (p. 110)

Por el contrario, esta nueva orientación concibe la composición como un proceso que aporta numerosos beneficios a los estudiantes de segundas lenguas. Diaz (1986) destaca algunas de estas ventajas cuando afirma que:

> …that not only are process strategies and techniques strongly indicated and recommended for ESL students, but also when used in secure, student-centered contexts, the benefits to these students can go beyond their development as writers. (p. 41)

La respuesta de los profesores a este último enfoque se tradujo en la adopción de medidas diversas dentro del aula. Entre ellas, cabe mencionar la concesión de un período de tiempo más extenso a los estudiantes para que elaboren los textos escritos. La ampliación de dicho periodo se utilizaría para planificar, seleccionar temas, generar ideas, escribir múltiples borradores y retrasar la provisión de *feedback* hasta la fase última de edición.

Autores como Zamel (1982) recomendaron tratar la producción escrita como un proceso en el que la corrección de errores lingüísticos y gramaticales adquiría una importancia decreciente. Zamel (1982) opinaba que el desarrollo de la competencia en el proceso del ensayo era prioritaria al desarrollo de la competencia lingüística ya que era esta primera competencia la que posibilitaba al estudiante la adquisición eficaz de la

habilidad de escritura en lengua inglesa. La autora sostenía que cuando los estudiantes experimentaban la composición como un proceso, la producción escrita final, o sea, el producto mejoraba como consecuencia (Zamel 1982).

El creciente énfasis otorgado al contenido y a la comunicación y la consecuente disminución del interés por la corrección de errores lingüísticos va a suponer un duro reto para algunos profesionales. Tanto es así, que autores como Leki (1990) llegan a calificar la función del profesorado como de esquizofrénica dado que, según manifiesta la autora, se han de conjugar tres facetas distintas al mismo tiempo: la faceta de lector real, la faceta de formador y, por último, la faceta de corrector o evaluador de la producción escrita. Leki (1990) sostiene que los profesores que apoyan el enfoque basado en el proceso se ven obligados a vivir en una contradicción constante, incluso esquizofrénica, al pretender ser ambas cosas, es decir, *colaboradores* y *jueces* a la vez. La autora considera que esta esquizofrenia ejerce un enorme impacto en nuestros estudiantes ya que, entre otras cosas, a dichos estudiantes se les intenta convencer de que han de desarrollar un sentido de la *audiencia* a quien poder dirigir el mensaje escrito. No obstante, los estudiantes son conscientes de que, en realidad, la audiencia es siempre la misma:

> As long as a teacher is evaluating a student's performance, it does not matter how much we try to persuade ourselves and our students that the audience is their classmates or the university community; the students know very well that whoever gives the grade is the audience, an audience who will decide something very important about their futures. (Leki 1990: 60)

La despreocupación por los aspectos formales de la lengua se llevó a posturas un tanto extremas. Tanto es así, que algunos autores decidieron omitir cualquier tipo de referencia gramatical en los textos escritos (Bennesch y Rorschach 1989; Crammer 1985, etc.). En opinión de estos autores, el lector del ensayo debía interesarse por la negociación de significados que se entendía básicamente como comunicación y expresión de ideas. La expresión formal del texto no debía ser, por consiguiente, innecesariamente atendida.

Como era lógico prever, algunas de las críticas más duras en contra del enfoque basado en el proceso surgirán de la comunidad académica. Así, Horowitz (1986a) sostiene que este enfoque no prepara al estudiante de forma adecuada para las tareas de producción escrita (por ejemplo, el ensayo) que deberá afrontar durante su académica. El énfasis que otorga este planteamiento a la figura del escritor y al desarrollo de su pensamiento se convierte en “almost total obsession”, lo cual resulta inapropiado para las expectativas del mundo académico (ver Horowitz, 1986a). Con ello se consigue, por el contrario, dar una impresión errónea a los estudiantes sobre el modo en el que van a ser evaluados:

> The process approach overemphasizes the individual’s psychological functioning and neglects the sociocultural context, that is, the realities of academia – that, in effect, the process approach operates in a sociocultural vacuum (Horowitz 1990: 17)

Así pues, y a pesar de la extensiva aceptación del enfoque de la producción escrita basado en el proceso con el consiguiente énfasis en el

contenido y, sobre todo, en la comunicación, dentro de las aulas, algunos profesores optarán por la adopción de paradigmas más tradicionales. La evaluación del ejercicio del ensayo en las PAAU es uno de los muchos ejemplos ilustrativos en donde se juzga el texto final o producto *versus* el proceso. Ello, como veremos en el capítulo 5 (epígrafe 5.5), responde al criterio de la factibilidad (i.e. inversión de tiempo, coste económico, recursos humanos, etc.) que dificulta la incorporación de este último enfoque en las pruebas de carácter nacional.

Sin embargo, el proceso que siguen los estudiantes en la elaboración del ensayo sigue estando presente en las aulas. Puede que esto se deba a que la adopción del enfoque basado en el proceso desempeña un papel reconciliador y unificador: "The process approach more than any other seems to be providing unifying theoretical and methodological principles." (Raimes 1991: 422).

4.5. Conclusiones generales de los estudios de investigación de la producción escrita en segundas lenguas

Durante los años 80, las investigaciones llevadas a cabo en torno a los procesos de producción escrita en segundas lenguas se extendieron rápidamente para apoyar las nuevas tendencias que se iban introduciendo en la enseñanza. Krapels (1990)[4] clasifica estos estudios y subraya siete argumentos principales que destacamos a continuación:

[4] Ver Krapels (1990) para un resumen exhaustivo sobre las principales investigaciones y estudios llevados a cabo en el proceso de producción escrita en segundas lenguas.

1. A lack of competence in writing in English results more from the lack of composing competence than from the lack of linguistic competence (e.g. Jones 1982; Zamel 1982; Raimes 1985a).
2. ...differences between L1 and L2 writers relate to composing proficiency rather than to their first languages (e.g. Zamel, 1983).
3. ...one's first language writing process transfers to, or is reflected in, one's second language writing process (e.g. Edelsky 1982; Gaskill 1986; Jones and Tetroe 1987).
4. The composing processes of L2 writers are somewhat different from the composing processes of L1 writers, a finding that contradicts item (2) (e.g. Raimes 1985 a, b, 1987; Arndt 1987).
5. First language use when writing in a second language, a fairly common strategy among L2 writers, varies (e.g., Martin-Betancourt 1986; Cumming 1987; Friedlander, 1990).
6. Using L1 when writing in L2 frequently concerns vocabulary and enables the L2 writer to sustain the composing process (e.g., Raimes 1985a; Martin-Betancourt 1986; Arndt 1987).
7. Certain writing tasks, apparently those relate to culture-bound topics, elicit more first language use when writing in a second language than other tasks do (Lay 1982: Burtoff 1983; Johnson 1985).

(Krapels 1990: 49-50).

Como se observa, y a pesar de que muchos de los estudios anteriores señalan prometedoras líneas futuras de investigación, la falta de una metodología común y el escaso número de sujetos utilizados como muestra en el análisis de los datos impiden hacer generalizaciones. Ello dificulta la obtención de un cuerpo de conocimiento válido sobre el que poder establecer inferencias. Según Krapels (1990):

> Studies cannot contribute significantly to the body of knowledge if very different research designs are used across studies. Lack of comparability affects generalizability, which in turn decreases the significance of a study's findings. (p. 51)

Conscientes de la importancia que supone la obtención de una muestra representativa de sujetos sobre la que poder inferir y generalizar los resultados, en este trabajo nos aseguramos de que la muestra, consistente en treinta y dos sujetos, fuera significativa desde un punto de vista estadístico y, por lo tanto, válida. Esto nos va a permitir, no sin la debida cautela, extrapolar y hacer extensivos las conclusiones obtenidas más allá de nuestro trabajo.

4.6. El inglés para fines específicos

Nuevas perspectivas surgirán en torno a los ensayos escritos en segundas lenguas. Entre estas últimas, cabe destacar el nuevo enfoque defendido por los proponentes del inglés para fines específicos quienes van a cuestionar el papel que desempeña el enfoque de la producción escrita basado en el proceso en el contexto académico.

Los estudios que apoyan dicha corriente buscan adecuarse a las tareas de producción escrita universitaria. Para conseguir este objetivo, se procede al análisis de contenidos específicos y de tareas que el estudiante deberá acometer durante su vida académica (Bridgeman y Carlson 1983; Horowitz 1986b). Desde esta nueva perspectiva, el contenido específico de las asignaturas será el que determine ahora las tareas de producción escrita que se realicen, el estudio de los diferentes formatos, la organización retórica (Selinker, Todd-Trimble y Trimble 1978) y la selección de los materiales de estudio más apropiados y que mejor se adecuen al aprendizaje de la producción escrita en lengua inglesa.

Este planteamiento también suscitará algunas críticas. Así, Raimes (1991) argumenta que el Inglés para fines específicos (IFE) y, por ende, el Inglés para fines académicos (IFA), corre el peligro de traducirse en el desarrollo de mini cursos elaborados sobre temas concretos o basados en instrucciones específicas en torno a campos conceptuales que dejan poco margen de autonomía al profesorado:

> With an autonomous ESL class, a teacher can – and indeed often does – move back and forth among approaches. With ESL attached in the curriculum to a content course, such flexibility is less likely. There is always the danger that institutional changes in course structure will lock us into an approach that we want to modify or abandon. (Raimes 1991: 411)

Además de la preocupación por el contenido académico, dentro del campo del IFE, se va a mostrar un especial interés por la figura del lector representante de la comunidad académica. En este sentido, el lector se convierte en el principal foco de atención de la producción escrita. El escritor, por su parte, deberá responder satisfactoriamente a las expectativas y demandas del lector para poder asegurar su acceso a la comunidad académica.

La filosofía que subyace esta nueva aproximación defiende que el estudiante a través de la producción escrita llega a convertirse en un miembro más de la comunidad académica e intenta aproximarse a ella (Horowitz 1986b; Silva 1990). La producción escrita se convierte, entonces, en un acto social en el que el lector es un lector experto y todopoderoso (*all-powerful*) (ver Johns 1990: 31). El enfoque social construccionista de la

teoría del ensayo que supone esta nueva orientación le otorga al lector el poder de aceptar o rechazar la producción escrita del estudiante según su grado de adecuación a las normas y convenciones establecidas por la comunidad académica. Johns (1990) nos aclara al respecto:

> In an academic context, the faculty audience is particularly omniscient, for they set the entire classroom agenda and have the final word on paper grading. (p. 31)

De acuerdo con esta aproximación, responder a las expectativas de los representantes de la comunidad académica será esencial para poder garantizar el éxito académico.

4.6.1. La comunidad académica

Algunos autores manifiestan que la adquisición de las convenciones de la comunidad académica puede resultar traumático para los estudiantes que proceden de otras comunidades sociales y culturales alejadas de la académica y con las cuales no se identifican. Hamp-Lyons (1991c) asegura que esta situación puede llegar a generar graves conflictos personales:

> When the writer perceives herself as a member of a community which is in disharmony with the wider community, the writer may experience stress, seeing the values of one community as in conflict with those of the other(s) (p. 56)

La autora continúa diciendo que, en una prueba de producción escrita, cuando se les pide a los estudiantes que escriban para una

comunidad específica se les está pidiendo que demuestren su predisposición para convertirse en miembros de una comunidad que puede resultarles “wholly unfamiliar” (Hamp-Lyons 1991c: 57). Asimismo, Johns (1990) afirma que:

> ESL students often run into major difficulties attempting to use the language of a discourse community when they do not fully understand the context for language use or the audience addressed (p.33)

De hecho, algunos autores califican la comunidad académica como de “powerful and controlling” y llegan a identificarla con la clase sociales dominantes (Ver Raimes, 1991).

Desde una perspectiva pedagógica, nos parece interesante la observación que apunta Hamp-Lyons (1991c):

> While we, as teachers and judges of the writing of nonnative users of English, do not yet possess sufficient knowledge of culturally determined writing behaviors to be able to teach students what to change in their writing to come close to “mainstream” expectations, we could at least take the first step of ensuring that we ourselves are clear about what those expectations are, and doing our utmost to interpret those expectations for ESL writers. (p. 61)

La información básica que Hamp-Lyons (1991c) sugiere hacer explícita a los estudiantes con el fin de ayudarles a asimilar las expectativas de la producción escrita académica, se puede resumir en los siguientes puntos:

> . test purpose (barrier testing? placement? diagnosis?)
> . format
> . test length
> . number of questions and score weighting if any
> . kind(s) of writing to be valued
> . criteria
> . scoring method
> . score reliability
> . qualifications of judges
> (Hamp-Lyons 1991c: 61)

El ejercicio de la prueba de Inglés de Selectividad facilita a los candidatos la mayor parte de la información citada anteriormente por Hamp-Lyons (1991c). Así, los candidatos tienen conocimiento verbal y escrito del objetivo de la prueba, del formato y de la extensión de la misma. También se les informa sobre el número de preguntas y la puntuación que recibe cada una de ellas según los criterios de evaluación establecidos para este propósito. Los dos únicos criterios que no se hacen públicos son el de la aptitud o calificación de los correctores, si bien se requiere que los correctores sean profesores titulados de enseñanza secundaria y de universidad, y el criterio de la fiabilidad inter-corrector. Como veremos en el capítulo siguiente, este tipo de fiabilidad estima el grado de consistencia de las puntuaciones obtenidas entre los diversos correctores que corrigen la prueba y su estudio constituye uno de los objetivos principales de este trabajo.

4.7. El enfoque actual

La gran variedad de teorías y enfoques que existen en torno a la enseñanza y la evaluación de la producción escrita en segundas lenguas, en general, y en torno al ensayo, en particular, ilustran la complejidad del proceso.

En principio, la aceptación de un enfoque único e incuestionable que resuelva todos los inconvenientes que se plantean en la elaboración y desarrollo del ensayo se ha desestimado. Tal y como afirma Johns (1990): "...no single, comprehensive theory of ESL composition can be developed on which all can agree." (p: 33). Por consiguiente, será responsabilidad de cada profesor abordar el estudio del contexto específico en el cual se va a trabajar y llevar a cabo un análisis de las necesidades que presentan los estudiantes. A partir de ahí, se adoptará la metodología que mejor se adecue a las circunstancias concretas de cada situación.

Según Raimes (1991), e independientemente de la metodología que se escoja, conviene conseguir el equilibrio entre cuatro elementos básicos que debieran estar presentes en toda teoría. Estos elementos son: la forma, el escritor, el contenido y el lector. Asimismo, hoy en día, se considera fundamental el hecho de: "not to seek for universal prescriptions" (Raimes 1991: 422).

CAPÍTULO 5. PRINCIPIOS BÁSICOS DE LAS PRUEBAS DE EXPRESIÓN ESCRITA

5.1. Introducción

El diseño de cualquier prueba debe reunir unos requisitos básicos y esenciales que justifiquen su uso en circunstancias concretas. Sin embargo, las consideraciones primordiales a tener en cuenta en el diseño de las pruebas pueden sintetizarse en dos aspectos básicos: validez (*validity*) y fiabilidad (*reliability*). Según nos explican Bachman y Palmer (1996): “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.” (p. 19).

En este capítulo se estudiarán detenidamente las principales aproximaciones al concepto de validez. Se prestará especial atención a la validez de constructo y a la determinación del constructo de expresión escrita que se define en las pruebas de expresión escrita. Seguidamente, se analizará el concepto de fiabilidad que constituye el objetivo fundamental de este trabajo. Posteriormente, comentaremos la cualidad del efecto rebote o

washback y la de la factibilidad (*practicality*). Por último, examinaremos la relación que se establece entre los conceptos de fiabilidad y validez.

5.2. La validez

En este epígrafe se definirá el concepto de validez y se analizarán sus principales características. En los siguientes subepígrafes, se estudiarán otras aproximaciones al concepto de validez como son: la validez de contenido, la validez aparente, el efecto rebote (i.e. *washback validity*) y la validez criterial.

Establecer el concepto de validez dentro del campo de LT resulta algo complejo debido a la dificultad que plantea su propia definición y a la dificultad que representa la consecución de su principal objetivo. Esto es: ¿mide la prueba lo que realmente se propone medir o mide, por el contrario, otras habilidades no deseadas?.

Hughes (1989) afirma que una prueba se considera válida: "...if it measures accurately what it is intended to measure" (p. 22). No obstante, la validez de una prueba puede establecerse desde perspectivas diversas y, de hecho, en la literatura observamos diversas concepciones de validez. Así, Cumming[1] (1996) cita 16 tipos distintos de validez descritos por Angoff (1988) en su recopilación histórica del concepto. Como veremos, en los últimos tiempos, las numerosas concepciones de validez propuestas

[1] Los tipos de validez que Angoff (1988) distingue son: *concurrent validity*, *construct validity*, *content validity*, *convergent validity*, *criterion-related validity*, *discriminant validity*, *ecological validity*, *face validity*, *factorial validity*, *intrinsic validity*, *operational validity*, *population validity*, *predictive validity*, *task validity*, *temporal validity* y *validity generalization*.

parecen englobarse bajo un único concepto que es la validez de constructo (Messick, 1989).

Dentro de la evaluación de la expresión escrita, algunos autores defienden con vehemencia la idea de que las pruebas directas como el ensayo, que evalúan la habilidad de la expresión escrita de los candidatos a través de la misma producción escrita, pueden predecir mejor que las pruebas indirectas (i.e. pruebas que miden la habilidad de producción escrita a través de otros medios), el comportamiento lingüístico deseado (Jacobs *et al.* 1981; Hamp-Lyons 1991a).

Según Jacobs *et al.* (1981), el mero hecho de utilizar pruebas de expresión escrita directas garantiza su validez dado que: “a direct test of writing is an unarguably valid measure of writing proficiency” (p.3). De acuerdo con este planteamiento, el principal inconveniente que presentan las pruebas de producción escrita indirectas (por ejemplo, las pruebas de elección múltiple, las pruebas C, etc.) es su aparente *invalidez.* Hamp-Lyons (1991b) defiende de forma enérgica este argumento:

> ...to test writing through some means other than writing, when it has itself been a focus of instruction, or when it is itself a criterion ability for the context the test taker hopes to enter, is entirely unacceptable (p.b)

Así, se puede afirmar que el uso extensivo de las pruebas de expresión escrita directas en la mayoría de instituciones reconocidas responde al deseo de incrementar su validez.

5.2.1. La validez de constructo

En la actualidad, el concepto de validez de constructo parece ejercer el papel de hiperónimo. Como tal, la validez de constructo se entiende como un principio que engloba las restantes formas de validez. Anastasi (1988) comenta al respecto:

> ... content, criterion-related and construct validation do not correspond to distinct or logically co-ordinate categories. On the contrary, construct validity is a comprehensive concept which includes the other types. (p.153)

Así pues, la validez se considera como: "a unitary concept requiring multiple types of evidence to support specific inferences made from test scores" (Moss 1992: 230). Esta idea es la que defienden otros autores como Cronbach (1988) y, especialmente, Messick (1989) principal responsable de la nueva concepción de validez.

Según Messick (1989), la validez representa un concepto unitario que justifica la interpretación de los resultados de las pruebas (*test interpretation*) y el uso (*test use*) que se haga de ellas. Esta nueva aproximación va a prestar especial atención a las *consecuencias* que se deriven de la implementación de las pruebas en contextos sociales específicos dado que: "...social values cannot be ignored in considerations of validity" (Messick 1989: 19). Así pues, las consecuencias sociales, la validez de constructo y el

significado de los resultados de las pruebas son aspectos que se hallan fuertemente interrelacionados.

Messick (1989) sintetiza las distintas facetas del concepto unitario de validez en la siguiente tabla:

Tabla 5.1. Facetas de la Validez (Messick 1989:20)

	Test Interpretation	**Test Use**
Evidential Basis	Construct validity	Construct validity + Relevance / Utility
Consequential Basis	Value implications	Social consequences

La tabla anterior nos muestra que la validez de constructo (*construct validity*) constituye la base (*Evidential Basis*) tanto de la interpretación de la prueba (*Test interpretation*) como de su uso (*Test use*). Este último aspecto se asocia a su vez con la relevancia de la prueba (*Relevance*), en cuanto al propósito que persigue, y con la utilidad (*Utility*) de la misma en el contexto en el que se aplica. Las consecuencias (*Consequential Basis*) de la interpretación de la prueba vienen determinadas por la evaluación o estimación de las implicaciones (*Value implications*) del constructo y por la utilización de la prueba y sus consecuencias sociales (*Social consequences*) tanto actuales como potenciales. Estas últimas se suman a su vez a las demás formas: validez de constructo (*Construct validity*), relevancia / utilidad (*Relevance / Utility*) e implicaciones (*Value implications*).

De este modo, Messick (1989) establece que la validez de constructo es el concepto que une la validez del uso de la prueba con la validez de la interpretación de la prueba. Así, juzgar si una prueba cumple o no su función requiere evaluar las consecuencias sociales intencionadas e incidentales de su uso. Este enfoque, que expande el concepto de la validez para incluir consideraciones sociales y éticas, va a formar la base de lo que se ha denominado *consequential validity* (Gipps, 1994).

5.2.1.1. El constructo de expresión escrita

Si queremos justificar la interpretación de los resultados de cualquier prueba debemos demostrar empíricamente que la prueba mide realmente las áreas o habilidades de la lengua que nos hemos propuesto medir. Para ello, debemos definir, en primer lugar, el constructo que deseamos medir. El constructo supone, en este sentido, establecer hipotéticamente las habilidades que se quieren medir. Bachman y Palmer (1996) describen el constructo como: “the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task” (p. 21).

A pesar de que Jacobs *et al*. (1981) defienden la validez de constructo de las pruebas de producción escrita directas y enfatizan su carácter comunicativo, en la actualidad, carecemos de unas bases teóricas claras y definitivas que definan el constructo o los constructos que subyacen la producción escrita. La necesidad de fundamentar el constructo sobre una base teórica es evidente:

> It is, after all, the theory on which all else rests: it is from there that the construct is set up and it is on the construct that validity, of the content and predictive kinds, is based. (Davies 1977: 63)

Para conseguir este objetivo, Carroll (1980) propone la utilización del esquema diseñado por Munby (1978) para la creación de diferentes perfiles de producción escrita que puedan adaptarse a las necesidades lingüísticas específicas. Sin embargo, el trabajo de Munby (1978) con el desarrollo exhaustivo de extensas listas de funciones y habilidades resulta poco operativo.

Como ya dijimos[2], la teoría de la *divisibilidad de competencias* va ganando cierto reconocimiento entre los investigadores, a pesar de la falta de evidencia empírica que la avale. De hecho, algunos de los estudios realizados con la finalidad de apoyar esta hipótesis como, por ejemplo, el de Bachman y Palmer (1996) se consideran innecesarios: “Common sense suggests that we do not **need** empirical research of the Bachman and Palmer or Oller kind to prove the obvious” (Alderson 1981: 188).

De acuerdo con esta teoría, diversos autores han intentado validar ciertos componentes o habilidades de la producción escrita en segundas lenguas. Para ello se ha utilizado el procedimiento estadístico de *multitrait-multimethod construct validation* (Campbell y Fiske 1959; Hamp-Lyons, Henning y DeMauro 1988). En el estudio de Campbell y Fiske (1959), la

habilidad comunicativa y la corrección lingüística muestran una mayor validez de constructo que el resto de las destrezas que se consideran: organización, argumento, interés, referencia y adecuación lingüística. Estos datos permiten inferir que, en principio, cualquier ensayo puede mostrar un nivel comunicativo distinto al de su corrección lingüística.

Dentro de esta misma línea, Cumming (1990, 1995) demuestra que el constructo de dominio lingüístico de una segunda lengua o *second language proficiency* es distinto al constructo de su producción escrita o *writing expertise.* A pesar de ello, en la evaluación de los ensayos ambos aspectos se evalúan conjuntamente.

Siguiendo el enfoque actual, en este trabajo vamos a asumir la existencia de un constructo de expresión escrita distinto al de otros constructos y, por lo tanto, representativo de uno de los muchos niveles de dominio de la lengua que se asumen.

Según Weir (1993) se requieren una serie de pasos para delimitar el constructo de expresión escrita. Como se observa en la Tabla 5.2. este autor distingue dos tipos fundamentales de validación: una validación anterior al diseño de la prueba (*a priori validation*) y otra validación posterior a la puesta en práctica de la misma (*a posteriori validation*):

[2] Ver capítulos 1 (epígrafe 1.3.) y capítulo 2 (epígrafe 2.4.)

Tabla 5.2. Metodología para la investigación del constructo de expresión escrita (Weir, 1993).

1. A PRIORI VALIDATION

1.1 SPECIFICATION of the CONSTRUCT

OPERATIONS

CONDITIONS under which these activities are performed

LEVEL OF PERFORMANCE: ASSESSMENT CRITERIA

These might be established through:

. Target situation analysis
. Theoretical literature
. Research literature
. Document analysis: course-books/tests

1.2 DEVELOPMENT OF PILOT TEST(S) TO OPERATIONALISE WRITING SPECIFICATION

. Format
. Rubric
. Timing
. Number of tasks
. Layout
. Invigilator's instructions
. Mini-trialling on self and on a few colleagues or students
. Produce first draft or mark scheme
. Moderate tasks and mark scheme in committee

2. A POSTERIORI VALIDATION

2.1 TRIAL ON REASONABLE SAMPLE

2.2 ESTABLISHING ESTIMATES OF RELIABILITY

. Internal and Intra marker reliability
. Internal consistency of analytic criteria

2.3 ESTABLISHING ESTIMATES OF EXTERNAL VALIDITY

. Qualitative expert judgement of items
. Feedback from test-takers (interview/questionnaire)
. Correlations against teacher's estimates

Weir (1993) sostiene que una validación de los resultados *a posteriori* es una condición necesaria pero no suficiente para definir el constructo. La valoración de las tareas a priori resulta imprescindible para que el contexto se incluya y se integre de forma adecuada en la prueba.

En este sentido, conviene recordar que, la validez del constructo de expresión escrita que deseamos medir se encuentra inexorablemente unida al contexto. Esto implica tener en cuenta dos aspectos fundamentales: a) la institución y b) los escritores en segundas lenguas (i.e. los estudiantes) (Hamp-Lyons, 1991b: 11). La especificación del contexto nos ofrece, además, la posibilidad de concretar el tipo de prueba que nos interesa (i.e. los ensayos). Así, las pruebas constituyen de algún modo definiciones prácticas de los diferentes constructos ya que intentan capturar y desarrollar la habilidad que se pretende medir.

En el contexto de las PAAU se siguen unas pautas muy limitadas. No se exige una validación a posteriori de la prueba si bien se intenta definir a priori el constructo del dominio de la lengua inglesa en general y, el constructo de la expresión escrita, en particular. El principal problema que nos encontramos es el que nos resume Kroll (1990c) a continuación:

> There is no single written standard that can be said to represent the "ideal" written product in English. Therefore, we cannot easily establish procedures for evaluating ESL[3] writing in terms of adherence to some model of native-speaker writing. (p.141)

[3] La abreviatura ESL se utiliza para refierse a *English Second Language* (i.e. Inglés como segunda lengua).

Debido a ello, la definición del constructo de expresión escrita se realiza a través de la especificación de las operaciones, condiciones bajo las cuales se realiza la tarea y, sobre todo, a través de los criterios evaluativos que se aplican en la corrección de los ensayos de la prueba de Inglés de Selectividad. Este planteamiento sugiere que cualquier interpretación justa que se haga de las puntuaciones obtenidas en esta última prueba ha de entenderse a partir de las especificaciones anteriores.

Por otra parte, consideramos que la estimación de la fiabilidad inter-corrector e intra-corrector que se realiza en este trabajo constituye un buen ejemplo de validación a posteriori que debiera estar presente en el diseño de las PAAU dada la trascendencia de sus resultados.

En este sentido, creemos que la pregunta formulada por Henning (1991): “How valid is valid?” (p.285) se responde atendiendo a la importancia que se atribuye a las decisiones que se toman a partir de las puntuaciones obtenidas en las pruebas. En nuestro caso concreto, las puntuaciones de las PAAU condicionarán la elección de la futura carrera universitaria de los candidatos. Por tanto, garantizar pruebas válidas y fiables debiera establecerse como un objetivo prioritario.

A pesar de que la visión unitaria del concepto de validez muestra el constructo como un eje unificador, existen otras aproximaciones al concepto que comentamos en los siguientes subepígrafes.

5.2.2. La validez de contenido

La validez de contenido se relaciona con los aspectos de relevancia y cobertura del contenido de la prueba. Según Hughes (1989):

> 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. (p. 22)

No obstante, la falta de una visión comúnmente aceptada de dominio de la lengua dificulta el que se garantice la relevancia de la muestra de las tareas incluidas en las distintas pruebas (Carroll, 1961). Moller (1982) argumenta:

> In the case of a proficiency test (...) the test constructors themselves decide the 'syllabus' and the universe of discourse to be sampled. The sampling becomes less satisfactory because of the extent and indeterminate nature of that universe. (p. 32)

Debido a ello, Henning (1987) clasifica la validez de contenido como de: "usually a non-empirical expert judgement" (p. 190).

Algunos autores prefieren incluir la validez de contenido dentro de lo que se denominan pruebas de rendimiento (*achievement tests*). Estas pruebas, al evaluar lo que se ha aprendido en un programa de enseñanza concreto, facilitan la inserción de tareas representativas de la muestra que se desea medir. A pesar de ello, Weir (1993) nos recomienda observar la

relación entre las especificaciones de las pruebas de dominio de la lengua ofrecidas por los expertos.

En la evaluación de los ensayos, la validez de contenido trata de garantizar que el contenido de los ensayos sea relevante y responda a las necesidades e intereses de los estudiantes. Henning (1991) considera necesario asegurarse de que el peso o valor que se atribuye a los distintos aspectos del contenido del ensayo es el apropiado. Para ello, el autor recomienda tomar como punto de referencia algunos estudios como el de Henning y Davidson (1987) que utilizan el método estadístico de regresión múltiple para generar coeficientes beta estandarizados. Estos coeficientes sugieren diferentes valores que podrían aplicarse a los diversos elementos del ensayo como son: el contenido, la organización, la expresión, la estructura y la mecánica

Hamp-Lyons (1990, 1991b), sin embargo, argumenta que el problema de la validez de contenido se resuelve mejor atendiendo a la validez de constructo. Según la autora, la validez de contenido no puede reducirse a la mera demostración del dominio de un cuerpo de conocimientos específico sugerido por el tema a tratar en el ensayo. Más bien se trata de ver lo que el sujeto puede hacer con las ideas esenciales y las destrezas que se entiende ha de dominar:

> …those who have worked with writers know that good writing is not about mastery of the components: It is about how all these things are put together for meaningful purpose. The question of content validity on a writing test, then quickly becomes one of construct validity. (Hamp-Lyons 1991b: 11)

5.2.3. Validez criterial

La validez criterial hace referencia a la validez establecida al correlacionar los resultados obtenidos en una prueba y aquellos obtenidos en otra prueba distinta u otras medidas basadas en un criterio externo e independiente que se considera que mide la misma capacidad. La prueba se valida, entonces, por comparación con el criterio externo previamente establecido.

La validez criterial se divide esencialmente en dos clases o tipos: validez concurrente y validez predictiva. La validez concurrente se establece al comparar los resultados de un grupo de candidatos en una prueba (por ejemplo, una prueba de expresión escrita) con los resultados del mismo grupo de candidatos en otra prueba distinta o medida similar, administrados más o menos al mismo tiempo (Davies 1983; Hughes 1989). Lado (1961) llegó a afirmar que la validez de una prueba podía determinarse únicamente si se basaba en ejemplos de validez concurrente que mostraran evidencia relacionada con criterios externos:

> If the two sets of scores correlate highly, that is, if the students who make high scores on the valid criterion test also score high on the experimental test and if those who score low on one also score low on the other, we say that the test is valid. (p. 30)

Davies (1983) comparte esta misma opinión y sostiene que:

> The external criterion, however hard to find and however difficult to operationalise and quantify remains the best evidence of a test's validity. All other evidence, including reliability and the internal validities is essentially circular. (Davies 1983: 1)

No obstante, en el ámbito de la evaluación de la producción escrita, algunos autores nos advierten de que la validez concurrente entre las pruebas de producción escrita directas y las indirectas que se establece puede resultar engañosa (Underhill 1982; Shohamy y Reeves 1985). Estos autores explican que las altas correlaciones que se observan en muchos estudios llevados a cabo en los años 70 entre las pruebas directas y las indirectas no significa que dichas pruebas estén midiendo la misma habilidad o destreza. Los resultados basados en las correlaciones únicamente demuestran que existe una relación directa entre las dos medidas que se utilizan. Hamp-Lyons (1990) defiende este argumento con determinación:

> Few ESL professionals these days are prepared to accept that we can test writing by any means other than writing: They believe that correlations of .80 with other tests, leaving as they do 36% of the variance in test score unaccounted for (correlation = *r*; variance = *r2*), leave unrevealed the most interesting part of the writer's repertoire. (p. 72)

Por otra parte, la validez predictiva hace referencia a la certeza con la que los resultados de una prueba de producción escrita pueden, por ejemplo, predecir una producción *futura* (Hughes, 1989). Sin embargo,

conseguir una buena validez predictiva representa una tarea compleja. Entre los principales inconvenientes destaca el margen de tiempo entre que transcurre la prueba inicial y la obtención de la evidencia posterior con la que correlacionar los resultados. Numerosos factores externos e incontrolables pueden, además, haber afectado al candidato modificando, por consiguiente, las expectativas previstas. Así pues, la validez predictiva se ha de aplicar e interpretar con cautela.

Algunos estudios como el de Breland *et al.* (1987) intentaron determinar hasta qué punto las pruebas de producción escrita directas e indirectas como son el ensayo o las pruebas de elección múltiple respectivamente, conjuntamente o tomadas por separado, podían predecir las puntuaciones de los estudiantes al inicio de sus estudios en lengua inglesa. Este estudio, a pesar de constituir un buen ejemplo de validez predictiva para la evaluación de la producción escrita, fue duramente criticado (Cherry 1989; Greenberg 1988). En general, se puede decir que los estudios realizados en torno a la validez predictiva han sido ineficaces (ver Cripper y Davies, 1988).

5.2.4. Validez aparente

La validez aparente es un concepto se une estrechamente a la validez de contenido. Este tipo de validez trata de averiguar hasta qué punto la prueba *parece* medir lo que realmente desea medir. En otras palabras: “does the test, on the “face” of it, *appear* to test what it is designed to test?” (Brown 1987: 222).

Algunos autores consideran que este tipo de validez no constituye una forma auténtica de validez dado que la validez aparente se basa en una impresión u opinión subjetiva y no en un análisis objetivo de la prueba. Además, el que se *parezca* que se está midiendo lo que se desea medir no garantiza que, en efecto, se esté midiendo aquello que realmente se pretende. Anastasi (1982) expone al respecto:

> (face validity)...is not validity in the technical sense; it refers, not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test 'looks valid' to the examinees who take it, the administrative personnel who decide on its use, and another technically untrained observers. Fundamentally, the question of face validity concerns rapport and public relations (Anastasi 1982: 136)

A pesar de que la validez aparente es un concepto escasamente científico y poco técnico, su importancia no se puede obviar. Tanto es así que, dentro del paradigma comunicativo, Morrow (1979) llegó a supeditar de forma muy polémica el concepto de fiabilidad al de validez aparente: "Reliability, while clearly important, will be subordinate to face validity. Spurious objectivity will no longer be a prime consideration..." (p. 151).

Desde la perspectiva de los candidatos, se ha de reconocer que la validez aparente de la prueba, también denominada *atractivo de la prueba*, puede ejercer un papel crucial en el éxito de su implementación. Para garantizar la motivación y el interés de los candidatos en la prueba, estos últimos debe de estar convencidos de que la prueba mide la capacidad que se desea medir, es decir, de que la prueba es válida. Una prueba carente de

validez aparente puede acabar siendo rechazada por los estudiantes o autoridades educativas.

Las pruebas objetivas directas han gozado siempre de gran validez aparente, incluso durante el apogeo de la era psicométrica-estructuralista. La inclusión del *Test of Written English* (TWE) como prueba opcional en la sección 2 del *Test of English as a Foreign Language* (TOEFL) en 1986 evidencia la gran validez aparente que se les concede.

De acuerdo con esta premisa, los candidatos atribuyen una gran validez aparente al ejercicio del ensayo de la prueba de Inglés de Selectividad. Este ejercicio se percibe como válido pese a la gran desventaja que supone la *subjetividad* de su corrección. La validez aparente del ejercicio del ensayo le añade un gran atractivo y predispone a los candidatos positivamente hacia la tarea, la cual consideran justa y creíble.

Bailey (1996) subraya la importancia de esta cualidad en la siguiente cita:

> If a test lacks face validity, it is unlikely that its results, no matter how clearly presented, will promote effective learning, improved teaching or positive curricular reforms (p.275)

De esta forma, Bailey (1996) une el concepto de validez aparente al efecto rebote o *washback validity* que comentamos en el siguiente epígrafe.

No obstante, y debido a que la validez aparente no deja de ser una *impresión* de validez carente de base empírica es necesario ir más allá de este concepto y buscar formas adicionales de estimar la validez de las

pruebas, en general, y de las pruebas de expresión escrita en particular (ver Hamp-Lyons 1991f; Henning 1991).

5.3. El efecto rebote

De nuevo, este es un concepto que muchos autores describirían fuera del ámbito de la validez. El efecto rebote, conocido más popularmente como *wasback validity* o *backwash*, se define como el efecto beneficioso (i.e. positivo) o perjudicial (i.e. negativo) que una prueba ejerce sobre la enseñanza o el aprendizaje que le precede. Buck (1988) nos explica la naturaleza de este fenómeno:

> There is a natural tendency for both teachers and students to tailor their classroom activities to the demands of the test, especially when the test is very important to the future of the students, and pass rates are used as a measure of teacher success. This influence of the test on the classroom (referred to as *washback* by language testers) is, of course very important; this washback effect can be either beneficial or harmful. (Buck 1988:17)

La influencia que ejercen las pruebas evaluadoras sobre la enseñanza no debe subestimarse. No obstante, algunos autores consideran que este argumento no guarda relación alguna con el concepto de validez o con el hecho de que la prueba mida realmente las capacidades que se pretenden medir (Hamp-Lyons 1996; Watanabe 1996). Estos autores critican la falta de evidencia empírica que avale el efecto rebote o *washback* y concluyen que muchas de las afirmaciones que defienden su existencia son:

"too simplistic" (Alderson y Hamp-Lyons 1996: 295) o se basan en suposiciones "too biased and naive" (Watanabe, 1996).

El estudio del efecto rebote resulta problemático debido a la imposibilidad de separar este fenómeno de las otras múltiples variables que influyen en el proceso de enseñanza-aprendizaje (Bailey 1996; Messick 1994). De hecho, varios estudios apuntan la posibilidad de que el efecto rebote no sea atribuible a una sola variable. Factores como el estilo individual de los profesores (Alderson y Hamp-Lyons, 1996) o la observación de que la introducción de una prueba afecta únicamente ciertos aspectos concretos de la enseñanza (Wall y Alderson, 1993) parecen confirmar que el efecto rebote no se produce de forma sistemática. En este sentido, Messick (1996) señala:

> Technically speaking, evidence of teaching and learning effects should be interpreted as washback (...) only if that evidence can be linked to the introduction and use of the test (p. 243)

A pesar de las críticas, la existencia del efecto rebote ha sido mayoritariamente aceptada y ha recibido distintas denominaciones según los autores: *test impact* (Baker, 1991), *systemic validity* (Fredericksen y Collins, 1989) o *consequential validity* (Messick, 1989).

Las concepciones que se tienen en torno a este fenómeno son también diversas. Messick (1996), por ejemplo, considera que el efecto rebote forma parte de la visón global y unitaria de la validez de constructo en la *interpretación* y en el *uso* de las pruebas. De este modo, Messick (1996)

asocia el efecto rebote a la necesidad de construir pruebas que sean ante todo válidas:

> ...negative washback *per se* should be associated with the introduction and use of less valid tests and positive washback with the introduction and use of more valid tests because construct under-representation and construct-irrelevant variance in the test could precipitate bad educational practices while minimizing these threats to validity should facilitate good educational practices. (p. 247)

Sin embargo, la visión más completa y exhaustiva del efecto rebote nos la ofrecen Alderson y Wall (1993) quienes llegan a establecer quince hipótesis (*Washback Hipótesis*) en torno a este concepto (ver, Alderson y Wall, 1993)[4].

Dentro del paradigma comunicativo, la necesidad de promover un efecto rebote positivo sobre la enseñanza se planteará como un objetivo prioritario (Hughes 1989[5]; Morrow 1991; Swain 1984). El enfoque comunicativo relaciona el efecto rebote negativo con las pruebas estandarizadas y poco comunicativas. El efecto rebote positivo se vincula, por el contrario, a las pruebas *auténticas* y *directas*. Según este planteamiento: “Ideally, the move from learning exercises to test exercises should be seamless.” (Messick 1996: 241).

El impacto positivo y beneficioso que ejercen las pruebas de producción escrita directas (i.e. el ensayo) sobre la enseñanza que les

[4] Ver Alderson y Wall (1993: 120-211) para el desarrollo y comentario de las 15 *Washback Hypothesis* que establecen estos autores.

precede constituye uno de sus principales atractivos. La idea que se defiende es que si evaluamos directamente las habilidades de expresión escrita que pretendemos desarrollar éstas últimas tenderán a potenciarse y a practicarse en el aula. Este pensamiento es el que Hughes (1989) expone: "If we want people to learn to write compositions, we should get them to write compositions in the test." (p. 45).

Las actividades conocidas como *teaching to the test* o *preparation for the examinations* realizadas en los Estados Unidos y en el Reino Unido respectivamente ilustran este último planteamiento potenciando y promoviendo el desarrollo de ejercicios de producción escrita directos como el ensayo.

Jacobs (1982) destaca, además, que el desarrollo de la producción escrita permite estructurar el pensamiento y promueve el aprendizaje. Respecto a este último punto, Hamp-Lyons (1991c) señala que la práctica de la producción escrita permite a los sujetos participar plenamente en la sociedad. La autora considera que la utilización de las pruebas de producción escrita indirectas es desalentador e inaceptable: "for those of us who see the act of writing as central to full participation in society." (p. 9).

En este sentido, algunos autores subrayan el impacto que ejercen las pruebas tanto a un macronivel (i.e. sociedad, sistema educativo, etc.) como a un micronivel (i.e. los individuos) (Madaus 1988; Gipps 1994; Bachman y

[5] Hughes (1989) dedica todo un capítulo a como conseguir un washback positivo (Ver capítulo 6, p.44-47).

Palmer 1996). Se enfatiza, sobre todo, el gran poder que pueden llegar a tener dichas pruebas (Madaus 1990; Alderson y Wall 1993; Shohamy 1994).

Precisamente, el impacto que suponen las PAAU tanto dentro del sistema educativo, a nivel general, como para los propios sujetos, a nivel individual, nos lleva a examinar la prueba de Inglés y, en concreto el ejercicio del ensayo, con detenimiento. Garantizar la construcción y evaluación de pruebas válidas y fiables supone una gran responsabilidad que conviene no olvidar.

5.4. La fiabilidad

La fiabilidad es otro de los requisitos fundamentales que se han de considerar en el diseño de las pruebas de lengua. La fiabilidad de una prueba se define como la consistencia o estabilidad de las medidas o resultados. Las preguntas fundamentales que giran en torno a este concepto son: ¿Es posible obtener resultados idénticos o similares en una misma prueba realizada en ocasiones distintas? o ¿son similares las valoraciones que establecen los distintos correctores en una misma prueba? Así pues, la fiabilidad se relaciona directamente con la consistencia de los resultados que la prueba produce y con la consistencia o nivel de acuerdo entre los diversos correctores en sus valoraciones del rendimiento del candidato.

La fiabilidad es un concepto estadístico que se puede expresar a través de un coeficiente de fiabilidad utilizando una escala que va del -1 al 1. El -1 indicaría una fiabilidad negativa o, lo que es lo mismo, una completa falta de fiabilidad. El 1 indicaría, por el contrario, una fiabilidad positiva o una

consistencia absoluta y, por lo tanto, perfecta. El nivel mínimo nivel de fiabilidad aceptable varía según las pruebas. Generalmente, se mueve en torno al 0,7 aunque un nivel $>$ 0,8 sería más deseable.

Sin duda, la mayoría de las críticas en contra de las pruebas de expresión escrita directas se atribuyen a la falta de consistencia o fiabilidad de las puntuaciones que se obtienen. Este fenómeno se asocia, principalmente, a una clase específica de fiabilidad denominada fiabilidad inter-corrector (*inter-rater reliability*). Este tipo de fiabilidad estima la consistencia de las puntuaciones de los correctores en la evaluación de una misma prueba. Por su parte, la fiabilidad intra-corrector (*intra-rater reliability*) investiga la consistencia del mismo corrector en ocasiones diferentes. El estudio de este último tipo de fiabilidad no ha suscitado, sin embargo, tanto recelo como el primero.

En este trabajo, ambos tipos de fiabilidad (i.e. fiabilidad inter-corrector y fiabilidad intra-corrector) se establecen como objetivos prioritarios. Como hemos visto, estos enfoque estiman que la fuente del error en la evaluación de la habilidad de la producción escrita se halla o bien en los diversos correctores (i.e fiabilidad inter-corrector) o bien en el mismo corrector (i.e. fiabilidad intra-corrector).

Indudablemente, la figura del corrector constituye una fuente manifiesta de error en la estimación de la fiabilidad de las pruebas de expresión escrita. Ello se explica porque: “...human evaluators are used and the possibility of human error exists.” (Hamp-Lyons ,1990).

Diversos estudios empíricos demuestran que aspectos potencialmente irrelevantes al constructo de la habilidad de producción escrita que se pretende medir como son: las diferencias culturales y el género de los candidatos (Newcomb 1977; Goddard-Spear 1983), la experiencia educativa de los correctores (Hamp-Lyons, 1989), los errores superficiales del ensayo, etc. pueden afectar seriamente las valoraciones del rendimiento de los candidatos. Esto explica, en parte, las grandes discrepancias que se observan en las puntuaciones que se asignan a los ensayos.

Así pues, garantizar la fiabilidad de los resultados, reducir su inconsistencia y asegurar su justa aplicación son algunos de los objetivos fundamentales que persiguen los expertos. Para ello se han propuesto diversas alternativas. Jacobs *et al*. (1981), por ejemplo, establecen siete pasos destinados a conseguir una lectura fiable:

> 1) adopt a holistic evaluation approach;
> 2) establish criteria to focus reader's attention on significant aspects of the compositions;
> 3) set a common standard for judging the quality of the writing;
> 4) select readers from the same background;
> 5) train readers until they can achieve close agreement in their assessments of the same papers,
> 6) obtain at least two independent readings of each composition
> 7) monitor the readers periodically to check their consistency in applying the standards and criteria of the evaluation
> (Jacobs *et al.* 1981: 28)

En las pruebas de Selectividad se respetan los cuatro primeros criterios. Es decir, el ensayo se evalúa siguiendo una orientación holística si

bien y, de acuerdo con el segundo punto señalado por Jacobs *et al.* (1981), se establecen ciertos criterios evaluativos para orientar la atención de los correctores hacia ciertos aspectos puntuales del ensayo. Existen, por tanto, unos criterios evaluativos comunes. Siguiendo con el esquema de Jacobs *et al.* (1981), los correctores que participan en la evaluación de las PAAU son profesores cualificados procedentes de enseñanzas medias y universidad. No obstante, los tres últimos criterios que establecen estos autores relacionados con la formación de los correctores en las diversas técnicas de evaluación (i.e. *rater training*) no se contemplan en el desarrollo de las PAAU.

Este último proceso de orientación se plantea como principal objetivo formar a los correctores en la aplicación de detallados niveles descriptivos y otros métodos destinados a reducir las discrepancias en las puntuaciones que se asignan a los distintos ensayos. Numerosos estudios demuestran las ventajas de la formación de los correctores y de sus efectos positivos (Freedman, 1979; Jacobs *et al*., 1981; Carlson et al. 1985). No obstante, en la actualidad, algunos autores cuestionan la efectividad de este procedimiento (Vaughan, 1991; Henning 1996; Weigle 1998).

Una de las técnicas más populares que se utilizan para garantizar la fiabilidad de las puntuaciones en la evaluación de la producción escrita es la denominada evaluación múltiple (*multiple scoring*). Esta técnica se vincula frecuentemente a la aplicación de escalas holísticas. Muchos programas de

evaluación de la producción escrita internacionales de prestigio[6] utilizan la figura de más de un corrector en la evaluación de los ensayos. Normalmente se utilizan dos correctores con la intervención de un tercero en caso de desacuerdo extremo.

Jacobs *et al.* (1981) afirman que el uso de tres correctores aumenta sustancialmente la fiabilidad de las puntuaciones. Dichos autores llegan a obtener correlaciones de 0,89 y 0,94 en sus investigaciones. Otros autores, por el contrario, aseguran que se requieren de cuatro a cinco correctores preparados e independientes para conseguir un nivel de fiabilidad óptimo (Breland *et al.*, 1987).

Henning (1991) nos explica la relación matemática que existe entre el número de correctores que se necesitan y el nivel de fiabilidad que se desea conseguir a través de la siguiente fórmula:

Fig. 5.1. Nivel de fiabilidad (Hennig 1991:289)

$$N = \frac{rd\,(1-ri)}{ri\,(1-rd)}$$

Donde, *N* = el número de veces que se debe incrementar el número de correctores para conseguir el nivel o la valoración deseada.

ri= el valor de fiabilidad inicial

rd= el nivel de fiabilidad deseado

La técnica de la evaluación múltiple se fundamenta en la hipótesis de que la participación de un mayor número de correctores en la evaluación de

[6] Algunos ejemplos de la intervención de dos o tres correctores se encuentran en la aplicación del *ESL Composition Profile* de Jacobs *et al.* (1981) y en el TWE (Test of Written English).

la producción escrita de los candidatos reducirá el margen de error. Como consecuencia de ello, se aumentará la fiabilidad y se obtendrá una puntuación global más próxima a la valoración real de la habilidad de producción escrita del candidato que la que se obtiene mediante la utilización del criterio aplicado por un único corrector.

Una última alternativa que se propone para aumentar la consistencia de las puntuaciones de la producción escrita es utilizar, no una, sino varias muestras de producción escrita y evaluarlas. Basar la puntuación final en distintos ejercicios escritos nos permitiría reducir la variabilidad que pudiera producirse entre las diferentes producciones. Jacobs *et al.* (1981) se muestran partidarios de tomar al menos dos muestras de producción escrita como punto de referencia para su evaluación.

Siguiendo esta misma línea, Hughes (1989) nos aconseja incluir “as many different tasks as posible” (p. 290). Este autor sostiene que los estudiantes son mejores en unas tareas que en otras y que cuantas más muestras de producción escrita se obtengan mayor será la posibilidad de aumentar la consistencia de las puntuaciones y de poder generalizar los resultados. Con respecto a este último punto, Henning (1991) manifiesta:

> With tests involving many items, the statistical procedure for equating is well developed, but with writing tests involving a single prompt and a single essay, it is sometimes difficult to ensure that the scores have the same meaning across different prompts and topics. (p.290)

Esta última preocupación es la que recoge la técnica del *portfolio*. Dicha técnica recopila varias muestras de producción escrita a lo largo de un

período de tiempo determinado y de acuerdo con las especificaciones del contexto académico particular. Las ventajas de la técnica del *portfolio* son evidentes ya que permite entre otras cosas evaluar la producción escrita como un *proceso* en el que se tienen en cuenta aspectos como la repetición del borrador (*redrafting*) y la auto-reflexión. Por ello, los *portfolios* se consideran instrumentos válidos:

> Portfolios, because they contain several samples, and because they can be constructed so that texts written under different conditions are included, allow a more complex look at a complex activity, and are therefore generally considered to be more valid. (Hamp-Lyons 1991e: 263)

No obstante, la escasa fiabilidad que se les atribuye a los *portfolios*. dificulta su implementación en evaluaciones nacionales e internacionales en las que conseguir puntuaciones fiables se plantea como un objetivo prioritario.

Como ya dijimos, el concepto de fiabilidad quedaba relegado a un segundo plano dentro del paradigma comunicativo. Este enfoque se mostraba más interesado en asegurar la obtención de un gran nivel de validez en las pruebas. En este sentido, la inclusión del ejercicio del ensayo en la prueba de Inglés de Selectividad constituye un ejemplo de la importancia que se le concede a la validez de la prueba dado que este ejercicio goza de una gran validez de constructo. No obstante, esta última ventaja no debe conducirnos a descartar la adopción de medidas destinadas

a conseguir resultados fiables en la evaluación de la competencia comunicativa en las PAAU. Sobre todo, si se tiene en cuenta que los resultados de las pruebas de Selectividad van a tener importantes repercusiones para la futura vida académica de los estudiantes.

5.5. La factibilidad

La cualidad de la factibilidad o aspecto práctico de las pruebas de evaluación no hace referencia al uso que se hace de las pruebas sino a la posibilidad de su implementación. La factibilidad de una prueba implica el diseño de pruebas económicas y de fácil administración, corrección e interpretación.

Bachman y Palmer (1996) definen la factibilidad como la relación que se establece entre los recursos necesarios para el diseño, desarrollo y uso de la prueba y los recursos disponibles para su puesta en práctica. Estos autores expresan dicha relación a través de la siguiente figura:

Figura 5.2.Factibilidad (Bachman y Palmer 1996: 36)

$$\text{Practicality} = \frac{\text{Available resources}}{\text{Required resources}}$$

If practicality ≥ 1, the test development and use is practical.
If practicality ≤ 1, the test development and use is not practical.

De acuerdo con la representación gráfica anterior, una prueba factible va a ser aquella que no requiera para su construcción y uso más recursos de los que haya disponibles.

Cumplir con el requisito de la factibilidad en la construcción de pruebas comunicativas puede resultar problemático. Weir (1993) nos explica los motivos:

> Tests of this type are difficult and time-consuming to construct, require more resources to administer, demand careful training and standardisation of examiners and are more complex and costly to mark and report results on. The increased per-capita cost of using communicative tests in large-scale testing operations may severely restrict their use. (p. 37)

Sin embargo, Hughes (1989) resalta que el coste económico de una prueba no es el criterio primordial a tener en cuenta a la hora de construirla. Este autor aconseja comparar el coste económico de una prueba con la pérdida de tiempo y esfuerzo que supone para los profesores y los estudiantes llevar a cabo actividades poco apropiadas a los objetivos de aprendizaje marcados. Tampoco cabe olvidar la consecuente pérdida económica a nivel nacional que representaría el no disponer de gente formada y competente en lenguas extranjeras. Desde esta perspectiva: “we are likely to decide that we cannot afford *not* to introduce a test with a powerful beneficial backwash effect.” (Hughes 1989: 47).

El criterio de la factibilidad es una de las mayores ventajas que presentan las pruebas de producción escrita directas como el ensayo. Estas

pruebas son fáciles de diseñar y administrar, no requieren instrucciones complicadas ni papel impreso y los resultados son fáciles de interpretar. Sobre todo, si se les facilita a los candidatos el peso o valor de los diferentes aspectos del ensayo de forma analítica.

No obstante, si se pretende conseguir un nivel de fiabilidad aceptable se va a necesitar una inversión de tiempo considerable. En aquellas pruebas nacionales o internacionales que requieran la formación previa de los correctores para garantizar la aplicación de un criterio de evaluación común en la corrección de las producciones escritas, la inversión de tiempo va a ser mayor.

En las PAAU no se contempla el proceso de formación de los correctores en técnicas de evaluación, aunque se establecen reuniones con carácter obligatorio entre los diferentes correctores y los coordinadores de las diversas áreas para asegurar la correcta aplicación de unas orientaciones y criterios básicos. Desafortunadamente, el hecho de que se asigne un solo corrector para un grupo determinado de ensayos propicia, en ocasiones, la aplicación de criterios individuales que no constan en los criterios evaluativos que se han establecido para la corrección de las pruebas. El proceso de formación de los correctores en la evaluación de las pruebas de Selectividad reduciría la cualidad de la factibilidad dado que encarecería considerablemente el coste económico de su desarrollo y puesta en práctica. Asimismo, se requeriría una mayor inversión de tiempo para poder formar a los correctores de forma adecuada. No obstante, como

ya dijimos, la cualidad de la factibilidad no debiera ser el requisito fundamental a tener en cuenta en el diseño de las pruebas.

5.6. La relación entre la fiabilidad y la validez: ¿unión o tensión?

De acuerdo con la teoría clásica de la evaluación, la validez de una prueba se puede obtener únicamente si se garantiza un nivel mínimo de fiabilidad. En palabras de Weir (1993): "...a test can only be valid if it is also reliable". Sin embargo, una prueba consistente y, por lo tanto, fiable no tiene por qué ser válida.

Como se recordará (ver capítulo 1, epígrafe 1.4.), Davies (1978) nos habla de la existencia de una *tensión* entre los conceptos de fiabilidad y validez. Dicha *tensión* se produciría en aquellas ocasiones en las que se hace necesario sacrificar la consistencia de las medidas para garantizar una mayor validez. El caso contrario no es posible, o sea, no se puede sacrificar la validez a expensas de la fiabilidad ya que entonces se obtendrían medidas fiables de un constructo que no es el que se desea medir.

Garantizar la fiabilidad de las pruebas comunicativas resultaba especialmente problemático. Esto se explica porque:

> Test reliability is increased by adding to the stock of discrete items in a test: the smaller the bits and the more of these there are, the higher the potential reliability. Validity, however, is increased by making the test truer to life, in this case more like language in use. (Davies 1982: 131)

A la vista de estos resultados, Davies (1982) concluye: “...we must recognise that reliability and validity are often at odds with one another.” (p. 30).

La solución propuesta para resolver esta *tensión* tanto dentro del paradigma comunicativo como dentro del campo del *educational measurement* consistirá en relegar el concepto de fiabilidad a un segundo plano (ver Rea, 1978). No obstante, la necesidad de obtener medidas precisas y resultados que permitan extrapolarse a otros contextos ha motivado que la validez se vea marginada en el diseño de pruebas estandarizadas (i.e. las pruebas de elección múltiple) que enfatizan el concepto de fiabilidad.

En estos últimos tiempos, el movimiento llamado *performance-based assessment* trata de equilibrar la *tensión* que se establece entre ambos conceptos. Harlen (1994) sugiere tratar esta *tensión* primando la calidad de la prueba. Ello se consigue a través del logro del máximo nivel de fiabilidad posible de acuerdo con el propósito de la prueba y a través del mantenimiento de altos niveles de validez, tanto de contenido como de constructo.

Según Nuttall (1987), la posibilidad de generalizar o extrapolar los resultados obtenidos en la muestra de la prueba al conjunto global de habilidades que definen el constructo que se persigue nos ofrece la oportunidad de unir el concepto de validez al de fiabilidad. Según nos explica Nuttall (1987), para poder generalizar los resultados, las pruebas

deben poseer ambas cualidades: validez y fiabilidad. En caso contrario no es posible llevar a cabo este proceso.

La *tensión* entre la fiabilidad y la validez se hace especialmente visible en las pruebas de producción escrita directas. Como hemos explicado a lo largo de este capítulo, dichas pruebas gozan de un alto nivel de validez tanto de tarea como de contenido. Los candidatos, además, le atribuyen una gran validez aparente, lo que les predispone positivamente hacia la tarea.

Asimismo, el efecto rebote que producen las pruebas de producción escrita directas es altamente positivo ya que su implementación se traduce en una inversión de tiempo y esfuerzo considerables en el logro de producciones escritas de calidad dentro del aula. Si lo que se pretende es, entonces, construir pruebas de expresión escrita que sean válidas la mejor forma de conseguirlo es pidiendo a los candidatos que escriban (ver Hughes, 1989).

Desafortunadamente, el incremento de la validez de la prueba viene acompañado, generalmente, por un descenso en la posibilidad de especificar lo que se está midiendo. Ello se traduce en una reducción de su fiabilidad.

Hoy en día existen numerosos métodos para conseguir una mayor fiabilidad en la corrección de las pruebas de producción escrita directas. Hughes (1989), por ejemplo, consigue niveles de fiabilidad por encima de los 0,9 en la evaluación de los ensayos. A pesar de ello, Hamp-Lyons (1990) asegura que la posibilidad de conseguir resultados totalmente consistentes

en la evaluación de los ensayos es prácticamente nula dado que los correctores son personas humanas y, por lo tanto, constituyen una fuente potencial de error.

El factor humano, según afirma Hamp-Lyons (1990), representa un problema tanto de fiabilidad como de validez. Esta autora afirma que la inconsistencia que se observa en las valoraciones del rendimiento de los candidatos se debe a que los correctores no comparten un mismo constructo de dominio de la producción escrita. Hamp-Lyons (1990) argumenta:

> To the extent that ranking differ, either different constructs are being measured, or measurement error is intervening. In my view, both of these intrusions into valid, reliable writing assessment are occurring (p. 81)

De acuerdo con esta premisa, la única posibilidad que existe de llegar a una unión entre los conceptos de validez y de fiabilidad es conseguir el acuerdo unánime de la definición del constructo o los constructos de la habilidad de producción escrita que se están midiendo o que se desean medir.

CAPÍTULO 6. LOS MÉTODOS DE EVALUACIÓN DE LAS PRUEBAS DE EXPRESIÓN ESCRITA

6.1. Introducción

La elección del método de evaluación que utilicemos en las pruebas de expresión escritas resultará esencial para poder justificar los resultados obtenidos en dichas pruebas y poder hacer inferencias sobre la habilidad de producción escrita que se desea medir.

Según Bachman y Palmer (1996), en la elección de cualquier método de evaluación que deseemos utilizar se han de tener en cuenta dos aspectos fundamentales. En primer lugar, se ha de prestar especial atención a la definición del constructo que se pretende medir y que determinará las diferentes habilidades que se van a evaluar. En segundo lugar, se han de considerar las especificaciones de la tarea que determinarán a su vez la respuesta que se espera en las pruebas y el tipo de evaluación que se vaya a realizar.

Debido a los diferentes usos que se hacen de los resultados de las pruebas y de la gran variedad de tareas que se pueden utilizar en su diseño se han desarrollado numerosos métodos para la evaluación de las pruebas de expresión escrita. No obstante, no es nuestra intención exponer aquí una lista detallada de todos los métodos evaluativos propuestos hasta el día de

hoy. Más bien vamos a centrarnos en la discusión de los dos principales métodos que se han creado para la evaluación de los ensayos en segundas lenguas: el método holístico y el método analítico.

En este capítulo se comentarán, en primer lugar, las principales características del método holístico y del método analítico y las filosofías que inspiran su puesta en práctica. Finalmente, se contrastarán ambos métodos y se analizarán las causas que determinan la elección del método en los diferentes contextos evaluativos.

6.2. Valoración por impresión recibida

El término *impresión recibida* (*impressionistic scoring*) hace referencia a los enfoques evaluativos que dependen principalmente de las valoraciones subjetivas de los correctores sin ningún tipo de orientación o "guides or controls" (White, 1984). Según Hamp-Lyons (1991f), la valoración por impresión recibida se vincula con frecuencia a la llamada valoración holística ya que este último tipo de valoración, al igual que la primera, produce como resultado una valoración única y global basada en la impresión general que produce el ensayo en los correctores.

Algunos especialistas consideran recomendable hacer una distinción entre ambos métodos dado que, actualmente, el método holístico tiende a evitar la total subjetividad y a hacer más explícitos los criterios de corrección que se aplican mediante el uso de una escala de valoración u otros criterios

orientativos. En este caso, la evaluación se denomina “focused holistic scoring” (ver Hamp-Lyons, 1991f).

A pesar de no ser equivalentes, los conceptos de valoración por impresión recibida y valoración holística se han utilizado indistintamente en la literatura de la evaluación. Brooks (1980), por ejemplo, no establece ninguna distinción entre ambos conceptos y entiende que la corrección del ensayo utilizando la valoración por impresión recibida implica el uso de dos o más correctores para llegar a una sola nota o puntuación global basada en la impresión general que produce el ensayo. La utilización de una escala de valoración u otros criterios específicos, a pesar de tenerse en cuenta mentalmente, no se especifican.

La valoración por impresión general en su forma más pura requiere que los correctores lean primeramente entre un 10% y un 25% de los ensayos como muestra. A partir de ahí, los correctores establecen un criterio evaluativo en su mente y lo aplican al resto de los ensayos que se han de leer a continuación de forma rápida. Este último punto es esencial para garantizar que los ensayos se evalúan de acuerdo con las primeras impresiones que producen en los correctores. McColly (1970) y Myers (1980) aconsejan formar previamente a los correctores para conseguir este objetivo. Asimismo, McColly (1970) recomienda el mantenimiento de un ritmo de lectura estable que se situaría en torno a unas 400 palabras por minuto.

6.3. El método holístico: definición y características

Jacobs *et al*. (1981) definen el método holístico como: “Any procedures which stop short of enumerating linguistic, rhetorical or information features of a piece of writing” (p. 92). Según estos autores, la noción de método holístico excluye claramente cualquier intento de separar los distintos elementos del ensayo para evaluarlos de forma independiente.

Dentro de esta misma línea, White (1984) manifiesta que el término holisticismo se utiliza para describir la visión de las cosas como un todo y utiliza este argumento para calificar al método analítico de reduccionista. Este último método, como veremos, descompone el ensayo en diversos aspectos del uso de la lengua. White (1984) afirma que, precisamente, el método holístico:

> …emerged as a reaction to the spirit of analytic reductionism of the age...it is associated with several movements in the field such as process research and post-structural literary criticism which reject product analysis. (White 1984: 20)

En este sentido, Hamp-Lyons (1990) argumenta que el método holístico: “...is based on the view that there are inherent qualities of written text which are greater than the sum of the text’s countable elements.” (p. 79).

A pesar de que tanto el método holístico como el método analítico se han convertido en los dos principales métodos de evaluación de los ensayos, hasta el día de hoy, ninguno de ellos ha logrado demostrar su superioridad frente al otro. Por el contrario, recientemente, algunos autores han empezado a cuestionar algunos de los principios teóricos y prácticos que

rigen ambos métodos (Raimes 1990; Freedman 1991). Con respecto a este punto, Cumming (1995) señala:

> theoretical and empirical validation is needed to support and refine the widespread uses of holistic, analytic or other impressionistic methods to assess students' written performance in second languages. (p. 72)

6.4. El método analítico: definición y características

El método analítico fue ideado por Diederich *et al.* (1961). Estos autores, a través del estudio y análisis de cincuenta y tres comentarios realizados por correctores ingleses nativos y utilizando la técnica del análisis factorial aplicada a trescientos ensayos de estudiantes universitarios, identificaron cinco componentes que los correctores aseguraban valorar en la evaluación de los ensayos. Estos componentes son: *ideas*, *forms*, *flavour*, *mechanics* y *wording*. De este modo, Diederich (1974), consiguió diseñar la primera escala analítica para la evaluación de la producción escrita en inglés como primera lengua (L1).

Dentro de la evaluación de la producción escrita en segundas lenguas (L2), el *ESL Composition Profile* diseñado por Jacobs *et al.* (1981) y formado por cinco components: contenido, organización, vocabulario, uso de la lengua y mecánica constituye: "...one of the first attempts to develop an analytic type of scale." (Sasaki 1999: 459).

Brooks (1980) define el método analítico del siguiente modo:

> Analytical marking refers to a method whereby each separate criterion in the mark scheme is awarded a separate mark and the total mark is arrived at by the addition of these marks. (p.6)

Perkins (1983) explica que la utilización del método analítico conlleva la separación de diversos componentes del ensayo para poder evaluarlos individualmente. Esta visión de la habilidad del dominio de la lengua susceptible de ser dividida en diferentes componentes es la que defienden Bachman y Palmer (1996) en el diseño de sus escalas de valoración:

> In designing rating scales, we start with componential construct definitions and create *analytic* scales, which require the rater to provide separate ratings for the different components of language ability in the construct definition. (p. 211)

6.5. El Método holístico *versus* el método analítico

El debate suscitado en torno a la utilización del método holístico versus el método analítico en la evaluación de los ensayos nos obliga a examinar ambos métodos con mayor detenimiento.

Una de las mayores críticas y, quizás, la más persistente que se le atribuye al método holístico es la escasa fiabilidad de las puntuaciones que se obtienen. Los resultados de las investigaciones han sido, sin embargo, algo contradictorios. Así, Cooper (1977), Perkins (1983) o Hughes (1986) obtienen altos niveles de fiabilidad inter-corrector e intra-corrector en las valoraciones holísticas de sus estudios llegando a alcanzar coeficientes de fiabilidad tan altos como 0,90.

Por el contrario, Hartong *et al*. (1936) demuestran que los correctores producen resultados holísticos o globales muy diversos debido, en parte, a la severidad o indulgencia con la que juzgan la producción escrita. Al parecer, algunos correctores se muestran inflexibles y basan sus calificaciones en textos ideales no existentes[1] mientras que otros evalúan los ensayos desde perspectivas más amplias. Este último grupo de correctores, consecuentemente, es el que produce los resultados más realistas en cuanto a la habilidad de los estudiantes que se pretende medir (Cumming, 1990).

Hartog *et al.* (1936) y Sweedler- Brown (1993) consideran que la discrepancia que manifiestan los diversos correctores en sus valoraciones globales del rendimiento de los candidatos se puede reducir hasta cierto punto utilizando el método de corrección analítico. Algunos autores aseguran que el método holístico puede inducir a los correctores a basar sus puntuaciones finales en las características superficiales de los ensayos. Así, Charney (1984) afirma que los altos niveles de fiabilidad entre los correctores que se consiguen en las evaluaciones holísticas, se debe a que estos últimos se dejan influir por características de los ensayos: “easy to pick out but irrelevant to ‘true writing’” (Charney, 1984). En este sentido, Song y Caruso (1996) postulan que el método analítico es: “a grading system that is both more reliable an unbiased in evaluating the essays of ESL students.” (p. 176).

Como dijimos (ver capítulo 5, epígrafe 5.4), algunos autores consideran que el problema de la escasa fiabilidad atribuida al método

[1] Ver Zamel (1985)

holístico puede minimizarse a través de la técnica denominada evaluación múltiple (i.e. *múltiple scoring*). Hamp-Lyons (1990) nos explica la filosofía sobre la que se fundamenta esta idea:

> The rationale generally given for multiple scoring is that multiple judgments lead to a final score that is closer to a "true" score than any single judgment. (p. 79)

Algunos partidarios del uso de la técnica de la evaluación múltiple como Wiseman (1949) demuestran que la suma de las puntuaciones holísticas de varios correctores puede reducir considerablemente la inconsistencia de las puntuaciones. Wiseman (1949) utiliza este argumento para defender la superioridad del método holístico frente al analítico. El autor afirma:

> ...the estimate of the probable correlation of averaged marks with 'true' marks is 0.92. This is very much higher than we could expect from <u>one</u> analytic marker. (Wiseman 1949: 205)

No obstante, Hamp-Lyons (1990) nos advierte de que la suma de dos o tres puntuaciones provenientes de correctores independientes plantea sus inconvenientes ya que, según esta autora: "how do we know that the result is in fact a "true" score?" (Hamp-Lyons 1990: 80). Esta técnica requiere que los correctores sean todos ellos consistentes ya que si uno de ellos no lo es la suma de las dos puntuaciones obtenidas será inconsistente y, por tanto, la

puntuación final será menos válida que la que se hubiera obtenido con un único corrector consistente.

A pesar de estas objeciones, una gran mayoría de autores coincide en afirmar que la suma de las puntuaciones de dos correctores independientes produce resultados más consistentes que los que se obtienen a partir de la puntuación de un único corrector. Raymond (1982), por su parte, opina que la diversidad de valoraciones entre los correctores es inevitable pero también deseable.

Otra de las alternativas que se proponen para aumentar el nivel de fiabilidad de los métodos holísticos consiste en utilizar escalas de valoración. (Vaughan 1991; Brown 1992; Shavelson *et al.* 1992). Chaplen (1970) aboga por la elaboración de una escala en la que cada puntuación se asocie a un nivel diferente de dominio lingüístico del candidato[2]. Un ejemplo del uso de escalas de este tipo lo encontramos en el *Test of Written English* (TWE) que se incluye en el *Test of English as a Foreign* Language (TOEFL). A esta prueba se le aplica una escala de valoración de seis niveles. Estos niveles contienen una detallada descripción de la actuación que se espera del candidato en cada momento.

En ocasiones se utilizan bandas de valoración de dominio lingüístico como sucede en el sistema de evaluación desarrollado por el *British Council*

[2] Como vimos en el epígrafe 6.2. este tipo de enfoque holístico “where readers are prefocused on some key facets of textuality” se denomina *focused holistic* (Hamp-Lyons, 1991e).

en el *English Language Testing Service* (ELTS)[3]. Este último sistema, incorpora cinco componentes que se evalúan de forma independiente de acuerdo con el criterio básico de lograr la eficacia comunicativa.

A pesar de la creciente utilización de las escalas de valoración, tanto holísticas como analíticas, algunos autores consideran que dichas escalas son herramientas ineficaces y poco fiables para la evaluación de los ensayos. Entre alguno de sus inconvenientes cabe citar la mayor inversión de tiempo y de personal que requiere su diseño y puesta en práctica (Cooper, 1977).

Las mayores críticas, sin embargo, hacen referencia a la vaguedad de los descriptores que se aplican a dichas escalas especialmente a las bandas de valoración holísticas que suelen ser muy confusas. Estas últimas se prestan a distintas interpretaciones lo que favorece el desacuerdo entre los correctores a la hora de aplicarlas.

No obstante, en opinión de Hughes (1989), el mayor inconveniente que presentan las escalas de valoración holísticas es: “the potential lack of fit in individuals between performance in the various subskills”. (p. 91). Bachman y Palmer (1996) sintetizan las principales desventajas de estas escalas en tres puntos:

1. problems of inference
2. difficulties in assigning levels, and
3. differential weighting of components

(Bachman y Palmer 1996: 209)

[3] Las bandas que se aplican al ELTS incorporan cinco elementos siguiendo el ejemplo del *ESL Composition Profile* de Jacobs *et al.* (1981) que evalúa también cinco aspectos:

Las objeciones anteriores impulsa a algunos investigadores a reconsiderar el uso de métodos de evaluación analítica. En las escalas analíticas cada criterio evaluativo se asigna a un nivel de la escala independiente, por lo que se evita el inconveniente de tener criterios colapsados en un mismo nivel. Bachman y Palmer (1996) explican que las escalas analíticas presentan una doble ventaja:

> First, it allows us to provide a 'profile' of the areas of language ability that are rated...A second advantage is that analytic scales tend to reflect what raters actually do when rating samples of language use (p. 211)

Asimismo, Song y Caruso (1996) y O'Loughlin (1994) defienden que las escalas analíticas contribuyen a hacer más explícitos los criterios de evaluación y las impresiones de los distintos correctores. De este modo, se consigue incrementar el nivel de fiabilidad. Por otra parte, Hughes (1989) afirma que: "the very fact the scorer has to give a number of scores will tend to make the scoring more reliable." (p. 94). Desde esta perspectiva, el método analítico podría considerarse como una alternativa válida a la técnica de evaluación múltiple que se utiliza en los métodos holísticos para incrementar su fiabilidad. Cabe destacar, además, que el uso de escalas analíticas ha demostrado ser de gran utilidad en la formación de los correctores en el proceso de la evaluación.

---

contenido, organización, vocabulario, uso de la lengua y mecánica.

Desde el punto de vista pedagógico, probablemente, la principal ventaja que presenta el método analítico es la de poder facilitar información a los correctores sobre los aspectos de la producción escrita que los estudiantes deberían mejorar. Esto se consigue gracias al útil y adecuado sistema de retroalimentación (*feedback*) que este método proporciona. En este sentido, Carlson y Bridgeman (1986) manifiestan que el método analítico es potencialmente más eficaz que el método holístico para la intervención educativa ya que permite considerar a los estudiantes de forma individual y no simplemente a nivel de grupo.

Hamp-Lyons (1991f) afirma, además, que el método analítico permite la creación de perfiles de la habilidad de la producción escrita de los diversos estudiantes. A partir de ellos se pueden llegar a diseñar cursos de recuperación muy productivos. De este modo, se establece una clara relación entre el proceso de enseñanza y el proceso de evaluación.

Sin embargo, los resultados de los estudios que defienden la superioridad del método analítico frente al método holístico no son concluyentes. Así, mientras algunos autores aseguran que el método analítico resulta mucho más efectivo que el método holístico a la hora de reducir la inconsistencia de las puntuaciones de los diversos correctores, otros autores critican el hecho de que este método no consiga un nivel de discriminación suficiente entre los estudiantes. O'Loughlin (1994) y Bacha (2001) obtienen una fiabilidad más alta utilizando el método holístico. No obstante, O'Loughlin opina que el método analítico es más válido. Según

este autor: “reliability is necessary but not a sufficient condition for test validity” (p. 41).

Otra de las mayores objeciones que se le atribuyen al método analítico se relaciona con el concepto de validez. Se cuestiona, por ejemplo, el que la comunicación escrita esté constituida por diferentes elementos susceptibles de ser evaluados de forma independiente. También se advierte que la concentración en los diversos aspectos puntuales del ensayo puede acabar desvirtuando el efecto global (*halo effect*) del texto (Henning, 1987).

Hamp Lyons (1991f) critica este razonamiento y afirma:

> ...I have never sat in a discussion of an assessment essay that did not move from holistic comments to discuss specific facets or traits in the essay. (p. 245)

Esta autora continúa diciendo que cuando se evalúa una pieza de producción escrita siempre se analizan los diversos elementos por separado aún cuando se hace uso de métodos de evaluación holísticos:

> Research conducted on the reading process show that readers do sometimes read to separate out features of essays they are trying to score. (Hamp-Lyons 1991f:248)

Este planteamiento sugiere que los correctores leen los ensayos de forma multidimensional y *nonholistically* (Hamp-Lyons, 1991). Bachman y Palmer (1996) comparten esta misma opinión y afirman que:

> …even when expert raters are asked to sort writing samples into levels on the basis of *overall* quality, they report that they take into consideration specific areas of language ability (such as grammar, vocabulary, content) when they do so. (p. 211-212)

Como vemos, los resultados de los estudios no son definitivos, por lo que creemos conveniente no defender el uso exclusivo que se haga sobre cualquiera de los dos métodos. Brooks (1980) explica que:

> The continued status and popularity of both of these methods is largely due to the fact that research has been unable to demonstrate conclusively the superiority of either method. The two methods have been found to be of roughly equal merit when judged by the criterion of reliability. (p. 40)

Según Hughes (1989), la elección entre el método de evaluación holístico o analítico estará condicionada, en parte, por el propósito de la prueba y por las circunstancias que rodean el proceso de evaluación. El autor nos sugiere tener en cuenta las siguientes recomendaciones:

> If it is being carried out by a small, well-knit group at a single site, then holistic scoring, which is likely to be more economical of time, may be the most appropriate. But if scoring is being conducted by a heterogeneous, possibly less well trained group, or in a number of different places, analytic scoring is probably called for. (Hughes 1989: 97)

La evaluación del ejercicio del ensayo en la prueba de Inglés de Selectidad PAAU sigue una orientación holística si bien se especifican brevemente algunos de los aspectos de la lengua que se deberían considerar en su evaluación.

Como ya explicamos en el capítulo anterior (epígrafe 6.5), la adopción del enfoque holístico obedece al criterio de la factibilidad y a consideraciones básicamente de carácter práctico. No cabe duda de que las evaluaciones holísticas son económicas en cuanto a tiempo y a esfuerzo. Esto las hace especialmente atractivas. Según White (1985): “the evolution of holistic scoring made direct measurement an economically feasible alternative to multiple choice testing.”

Sin embargo, el criterio de la factibilidad no garantiza que los métodos holísticos sean siempre los más idóneos. Hamp-Lyons (1991f: 244) se muestra especialmente crítica ante la adopción de métodos holísticos y señala: “I have become convinced that the writing of some writers cannot be encapsulated into a single score.” Esta autora defiende la idea de que una nota global y única no define de forma adecuada la habilidad del candidato.

No obstante, dentro del marco comunicativo de la lengua en el que nos movemos, la utilización del método holístico para evaluar los ensayos de las PAAU se considera una decisión acertada dado que el objetivo último que se persigue es el de la evaluación de la eficacia comunicativa de la producción escrita.

No por ello se descarta el estudio de medidas destinadas a minimizar la variabilidad de las puntuaciones obtenidas en las pruebas. Entre ellas cabe destacar la elaboración de criterios de corrección más claros y operativos que faciliten la llegada a un consenso entre los correctores.

Sea cual sea el método de evaluación elegido, la nueva visión de validez propuesta por Messick (1993) exige recoger evidencia empírica: “to

support the adequacy and appropriateness of inferences and actions based on test score" (p. 13). Este es uno de los objetivos básicos que nos hemos planteado en este trabajo.

CAPÍTULO 7. PRINCIPALES VARIABLES DE LAS PRUEBAS DE EXPRESIÓN ESCRITA

7.1. Introducción

En este capítulo nos ocuparemos de las principales variables de la producción escrita. En primer lugar, se analizarán las investigaciones llevadas a cabo en torno a la fiabilidad asociada con la figura del corrector. Dado que este punto constituye el eje central de este trabajo nos detendremos en él especialmente. A continuación, examinaremos las variables relativas a la tarea que pueden resultar más relevantes en el contexto de las PAAU. Finalmente, se introducirán las variables relativas a los candidatos como escritores de los textos.

7.2. Variables relativas a la figura del corrector

7.2.1. Introducción

En los siguientes subepígrafes se realizará un estudio detallado de los aspectos más destacados relacionados con el estudio tanto de la fiabilidad inter-corrector como de la fiabilidad intra-corrector. Se prestará especial atención al estudio de este primer tipo de fiabilidad (i.e fiabilidad inter-corrector) ya que dicha variable se considera una de las fuentes potenciales de error más importante en la evaluación de la producción escrita.

Nos adentraremos en este terreno exponiendo, en primer lugar, las principales técnicas de investigación que se utilizan para analizar la actuación de los correctores durante el proceso de evaluación de los ensayos. A continuación, expondremos los resultados de los estudios que tratan de explicar el comportamiento de los correctores a través de diversas variables como son: la disciplina académica y la experiencia profesional, la experiencia lingüística, y diversos aspectos de la personalidad y afectividad (i.e. el género, la edad). Entre estos aspectos destacaremos la influencia del factor género y de la situación laboral de los correctores cuyo estudio se abordará en la parte empírica de nuestro trabajo.

Posteriormente, analizaremos las distintas *facetas* de la producción escrita que parecen afectar en mayor medida la fiabilidad de las puntuaciones de los ensayos. Aquí, se investigará la influencia de los niveles de dominio lingüístico de los ensayos, los elementos sobresalientes, la evaluación del error y el estudio de la forma *versus* el contenido.

Cerraremos el análisis de las variables relativas a la figura del candidato con el estudio de la filosofía sobre la que se fundamentan algunas de las técnicas que se utilizan para formar a los correctores en las distintas técnicas de evaluación.

No cabe duda de que la figura del corrector va a jugar un papel esencial en la evaluación de las pruebas *subjetivas* entre las que se incluye el ejercicio del ensayo. En dichas pruebas se observan grandes fluctuaciones en los juicios emitidos por los diversos correctores tanto entre

los distintos correctores (i.e. fiabilidad inter-corrector) como entre el mismo corrector en diferentes ocasiones (i.e. fiabilidad intra-corrector). Hamp-Lyons (1991f), por ejemplo, afirma que la fiabilidad de las pruebas basadas en los ensayos se va a ver muy afectada por las variaciones en las percepciones y en las actitudes de quienes las corrigen.

La disyuntiva que plantea el tener que reconciliar la subjetividad de los correctores con la objetividad que requiere cualquier prueba de producción escrita se ha recogido extensamente en la literatura. (Diederich French y Carlton 1961; Oller 1979; McNamara y Adams 1991/1994). La fiabilidad inter-corrector, especialmente, se considera una de las mayores fuentes de error en la medición y evaluación de la producción escrita del candidato. Tanto es así, que Moss (1994) califica este tipo de fiabilidad como: “the greatest bugbear in assessment” . Ya en su momento, Edgeworth (1980) acusaba la gravedad que representa la falta de fiabilidad de las puntaciones en las evaluaciones públicas:

> I find the element of chance in these public examinations to be such that only a fraction_ from a third to two-thirds_ of the successful candidates can be regarded as safe, above the danger of coming out unsuccessfully if a different set of equally competent judges had happened to be appointed. (Edgeworth 1980:653)

Algunos de los estudios conducidos hasta ahora parecen indicar que las distintas valoraciones que establecen los correctores obedecen a diversos aspectos o *facetas* de la producción escrita (Diederich, French y

Carlton, 1961). Por otro lado, diversos autores manifiestan que el comportamiento de los correctores en la evaluación de las pruebas escritas se atribuye, en parte, a variables como son: el género, la experiencia profesional, la situación laboral, el tiempo de exposición a la producción escrita en L2, etc. (ver Hamp-Lyons 1990).

Gamaroff (2000) considera que la falta de consenso que se acusa entre los diversos correctores no debiera ser en sí sorprendente dada la compleja naturaleza de las actividades que entran en juego en el proceso de la evaluación. En este sentido, conviene recordar que tanto el desarrollo como la evaluación de la producción escrita son actividades cognitivas que conllevan complejos procesos de pensamiento, inferencia, expectación, actitud, percepción y valoración:

> Language is closely connected to human rationalities, imaginations, motivations and desires, which because they each comprise and extremely complex network of biological, cognitive, cultural and educational factors, could easily compromise the quest for objectivity. (Gamaroff 2000: 34)

A la complejidad de este proceso hay que añadirle la posibilidad de que se produzca algún tipo de interacción entre un determinado corrector y un aspecto concreto de la situación de evaluación. En nuestro caso, dado el carácter específico de la tarea que se les pide evaluar en las PAAU, la interacción puede darse entre el corrector y el posible planteamiento del tema. Dicha interacción, que se define como sesgo o *bias* en el método de evaluación de facetas múltiples, consiste en los siguiente:

> Raters may display particular patterns of harshness or leniency in relation to only one group of candidates, not others or in relation to particular tasks, not others, or on one rating occasion, not the next. That is, there may be an *interaction* involving a rater and some other aspect of the assessment setting. (Lumley y McNamara 1995: 56)

A pesar de ello, la gran y, hasta cierto punto sorprendente, variabilidad observada en los juicios emitidos por los correctores sobre los diversos ensayos ha motivado que varios investigadores se replanteen el estudio del constructo que se está midiendo. En opinión de Milanovic *et al.* (1996), la diversidad de puntuaciones se debe a que los correctores están midiendo distintas habilidades o constructos de la producción escrita:

> These findings indicate that markers do not seem to measure one common construct that might be termed writing ability while carrying out an assessment. It seems clear then that marking is not simply a matter of reliability, but also concerned with the issue of construct validity. (p. 93)

Gamaroff (2000) utiliza este último argumento para unir los conceptos de validez y de fiabilidad. El autor argumenta que si los correctores no pueden llegar a un acuerdo sobre lo que se está midiendo, es decir, si no se puede establecer la fiabilidad inter-corrector, entonces carece de sentido hablar de validez. Según esta aproximación, los conceptos de fiabilidad y validez representarían dos caras de la misma moneda.

> It is in rater (un)reliability that matters of validity and reliability come to a head, because it brings together in a poignant, and often humbling and humiliating way, *what* is being (mis)measured, which is the concern of validity, and *how* it is (mis)measured, which is the concern of reliability. (p. 47)

Por su lado, Davies (1990) opina que la subjetividad de los correctores en la evaluación de los ensayos es, probablemente, una variable imposible de controlar. Ello se debe, en parte, a que cada corrector dispone de su propio criterio de corrección. Así, y a pesar de la puesta en práctica de diversas técnicas destinadas a promover el acuerdo entre los diversos correctores y reducir la variabilidad de las puntuaciones (i.e. *moderation workshops*, formación de los correctores, etc.), los diferentes estilos de corrección (*rating styles*) individuales acaban manifestándose a lo largo del proceso de evaluación.

Ante la imposibilidad de que los correctores coincidan en la aplicación de sus criterios, Oller (1979) justifica el uso de un único corrector durante todo el proceso de evaluación de los ensayos. Según nos explica Gamaroff (2000): “Oller’s point is that because it is difficult to get raters to agree, one should do the next best thing and try to agree with oneself (intrarater reliability).” (p.45). Esta teoría se intentará contrastar en la parte empírica de nuestro estudio en la que se analizarán las puntuaciones individuales de cada uno de los correctores en la primera y en la segunda ocasión (PRE y POST respectivamente).

Gamaroff (2000) rebate el argumento de Oller (1979) y subraya las dificultades que la presencia de un solo corrector plantea: “…But then, as we

know, we cannot be sure that the rater will not mark differently before breakfast (a good or bad one) than after." (Gamaroff 2000: 45).

Como veremos, el debate en torno a la posibilidad de controlar la subjetividad de los correctores sigue abierto y generando polémica.

7.2.2. Técnicas de investigación del comportamiento de los correctores durante el proceso de evaluación

En los siguientes subepígrafes expondremos las principales técnicas de investigación utilizadas para averiguar los criterios de actuación que siguen los correctores en el trascurso de las evaluaciones holísticas de los ensayos. Entre ellas destacaremos: los *think-aloud protocols*, las observaciones etnográficas, entrevistas y comentarios y, por último, otras medidas de carácter objetivo. Estas dos últimas técnicas se aplicarán en la parte empírica de nuestro estudio.

Algunos estudios tratan de descubrir las posibles causas que explican la falta de fiabilidad en las puntuaciones de los correctores. Para ello, se intenta analizar lo que piensan los correctores en el mismo momento de la evaluación de los ensayos: "Stated simply, if we do not know what raters are doing (and why they are doing it), then we do not know what their ratings mean." (Connor-Linton 1995: 763). La observación directa del proceso de la evaluación permite capturar los procesos mentales que siguen los correctores en el momento de evaluar los ensayos. Es decir: "what goes on in trained raters' minds when they are evaluating essays holistically."

(Vaughan 1991: 113). Esta premisa no lleva de lleno al terreno del cognitivismo. Se tratará de saber qué paradigmas conceptuales aplican los diversos correctores durante el proceso de evaluación.

Para iluminar este complejo proceso, se han propuesto diferentes métodos de investigación que se comentarán a continuación.

Entre ellos cabe citar los siguientes: *think-aloud protocols*, observaciones etnográficas, entrevistas y comentarios y, por último, la observación de la relación que se establece entre las puntuaciones y otras medidas de carácter objetivo (ver Connor-Linton, 1995).

7.2.2.1. *Think-aloud protocols*

Este primer método consiste en la observación directa del proceso de evaluación a través del análisis del registro de los pensamientos que los correctores realizan en voz alta (i.e. *think-aloud protocols*) en el momento de evaluar los ensayos. Cumming (1990) nos ofrece en su estudio un ejemplo ilustrativo de la implementación de dicha técnica destinada a aportar evidencia empírica sobre la actuación que siguen los correctores durante el proceso de evaluación. Las conclusiones a las que llega Cumming (1990) cuestionan la posibilidad de obtener resultados homogéneos en las valoraciones de los correctores aún cuando estos últimos se guíen por escalas y criterios de corrección específicos o bien cuenten con una experiencia profesional considerable:

> The sheer quantity of interrelated decisions which occur in this process testify to the difficulty of obtaining homogeneous ratings on composition exams, even among skilled raters. (Cumming 1990: 44)

Vaughan (1991) decide también utilizar la técnica de los *think-aloud protocols* en su estudio. La autora coincide con Cumming (1990) en señalar la discrepancia que se observa en el comportamiento de los correctores. Al parecer, y a pesar de la utilización de escalas de valoración, los correctores expresan sus dudas sobre el modo de aplicar ciertos criterios evaluativos establecidos en la escala. Ello provoca que, en la mayoría de los casos, los correctores acaben estableciendo sus propios criterios de evaluación y se guíen por estos últimos. De ahí, que Vaughan (1991) concluya que la evaluación holística se concibe como un acto individual:

> The data show that raters are not a *tabula rasa*, and do not, like computers, internalize a predetermined grid that they apply uniformly to every essay. Despite their similar training, different raters focus on different essay elements and perhaps have individual approaches to reading essays. Holistic assessment is a lonely act. (p. 120)

Los resultados obtenidos en los estudios anteriores nos han permitido trazar los diferentes estilos de lectura que demuestran los correctores en el transcurso de sus evaluaciones holísticas. Cumming (1990) llega a identificar veintiocho tipos distintos de comportamiento que manifiestan los correctores durante la evaluación de los ensayos. Estos comportamientos se recogen en la Tabla 7.1.

Tabla 7.1 Comportamiento de los correctores en la evaluación de los ensayos

(Cumming 1990:37)

Self-control Focus	content focus	language focus	organization focus
Interpretation strategies			
1. scan whole text to obtain initial impression	3. interpret ambiguous phrases	5. classify errors	7. discern rhetorical structure (s)
2. envision situation of writing and writer	4. Summarize propositions	6. edit phrases	
Judgements strategies			
8. establish personal response to qualities of items	14. Count propositions to assess total output	19. establish level of comprehensibility	25. assess coherence
9. define, assess revise own criteria & strategies	15. assess relevance	20. establish error values	26. identify unnecessary repetition
10. read to assess criteria	16. assess interest	21. establish error frequency	27. assess helpfulness in guiding reader
11. compare compositions	17. assess development of topics	22. establish command of syntactic complexity	28. rate overall organization
12. distinguish interactions between categories	18. rate content overall	23. establish appropriateness of lexis	
13. summarize judgments collectively		24. rate overall language use	

En opinión de Cumming (1990), la actuación o comportamiento de los correctores sigue dos estrategias principales: las estrategias interpretativas

(*interpretation strategies*), utilizadas para leer los textos (del ítem 1 al 7) y las estrategias de valoración (*judgement strategies*), utilizadas para evaluar las cualidades de los textos (del ítem 8 al 28).

Estas estrategias contienen, a su vez, cuatro focos principales que se localizan en:

1. el control del propio corrector sobre los procesos de lectura y evaluación (ítems 1, 2 y 8-13)
2. el contenido sustancial de los textos (ítems 3, 4 y 14-18);
3. el uso de la lengua en los textos (ítems 5, 6 y 19-24) y, por último,
4. la organización retórica de los textos (ítems 7 y 25-28).

De igual modo, Vaughan (1991) recoge diversas estrategias o estilos de lectura que observa en los correctores durante los *think-aloud protocols*. Los comportamientos que se registran evidencian las distintas actuaciones individuales que siguen los correctores en el proceso de la evaluación de los ensayos. Vaughan (1991) distingue y clasifica varios estilos de lectura (i.e. *reading styles*) aunque no los estudia en profundidad. Estos estilos son: *the single-focus approach*, *the "first impression dominates" approach*, *the "two-category" strategy*, *the laughing rater* y *the grammar-oriented rater* (p.118).

Siguiendo el mismo procedimiento, Milanovic *et al.* (1996) afirman que la actuación de los correctores durante la evaluación de los ensayos sigue cuatro estrategias básicas: 1) *Principled two-scan/read*; 2) *Pragmatic two-*

scan/read; 3) *Read through* y 4) *Provisional mark*. A continuación, se explican sus rasgos más característicos:

1. *Principled two-scan/read*_ los correctores examinan el texto detenidamente o lo leen dos veces antes de decidir la nota final que le conceden.
2. *Pragmatic two-scan/read*_ los correctores leen el texto dos veces antes de asignarle la nota final. La segunda lectura únicamente se realiza cuando se han encontrado dificultades en el texto o en el proceso de evaluación que les ha obligado a releer el texto para poder determinar la nota final con confianza.
3. *Read through*_ Se lee el texto una sola vez para extraer los aspectos positivos y los aspectos negativos.
4. *Provisional mark*_ Se hace una sola lectura del texto pero con una breve interrupción en el proceso de evaluación. Esta interrupción se realiza normalmente en la primera parte del ensayo. El corrector establece, entonces, una nota inicial hipotética sobre los méritos del candidato antes de reanudar la lectura de la parte restante del texto. Esta última será la que confirme o contradiga la evaluación inicial.

Como vemos, los estudios citados hasta ahora ilustran la diversidad de lecturas individuales que siguen los correctores. Cumming (1990) comenta al respecto:

> These tendencies suggest that the thinking processes involved in evaluating ESL compositions consist in many discrete decisions, which are relatively variable from person to person. (p. 37)

Gamaroff (2000) también se hace eco de la *guerra de estilos* o *style wars*[1] que se debate entre los distintos correctores. Según este autor, se trata simplemente de estilos de evaluación distintos:

> There are different learning styles, teaching styles and also rating styles. One rater, as indeed one learner or one teacher, may be mainly interested in the big picture, i.e. in coherence, while another may be mainly interested in systematicity and structure. (p.44)

No obstante, la variedad de estilos que se acusa adquiere una especial relevancia en situaciones de evaluación como las PAAU dada la trascendencia que van a tener los resultados para los candidatos que concursan en dicha pruebas.

7.2.2.2. Observaciones etnográficas, entrevistas y comentarios

Un segundo grupo de técnicas utilizadas para investigar la actuación de los correctores en el transcurso de las evaluaciones holísticas de los ensayos, consiste en realizar observaciones etnográficas o entrevistas a los correctores para que ellos mismos nos expliquen su comportamiento durante el proceso de evaluación.

[1] Ver Oxford *et. al*., (1991) y Dreyer, (1998).

Hamp-Lyons (1991e) pone en práctica algunas de estas técnicas cualitativas en su estudio como son el debate y la discusión. Así, la autora les pide a los correctores que argumenten y expliquen personalmente las razones que les han impulsado a asignar las distintas puntuaciones a un mismo grupo de ensayos. El objetivo que se persigue con ello es: “to make external some of the internal values that good readers of essays apply in making their judgements.” (Hamp-Lyons 1991e: 135). Esto es, el estudio de Hamp-Lyons (1991e) trata de analizar los procesos cognitivos, sociales y evaluativos que utilizan los correctores en el momento de evaluar las producciones escritas. Estos últimos procesos son los que, precisamente, intentaremos averiguar en el análisis de los comentarios positivos y negativos que cada uno de los sujetos de nuestro estudio realiza sobre los ensayos (ver cuestionario en Apéndice 1).

En su estudio, Hamp-Lyons (1991e) estudia diversas cintas grabadas que recogen el debate, los comentarios y los motivos que impulsan a los correctores a asignar las distintas puntuaciones a los ensayos. La riqueza de opiniones que obtiene la autora con la implementación de algunas técnicas cualitativas le permiten confirmar que los correctores varían considerablemente en la forma de llegar a establecer sus valoraciones individuales. Estudios posteriores confirmarán estos resultados (Brown 1991; Connor-Linton 1995; Gamaroff 2000).

La técnica de la entrevista con preguntas abiertas le permite a Johns (1991) averiguar los distintos aspectos que conforman lo que los profesores denominan *academic literacy*. Según Johns (1991) los tres criterios básicos

que los profesores enfatizan en la evaluación de la producción escrita académica son:

> 1) Use of Readings for Writing Assessment: Testing for Audience Awareness;
> 2) Exploitation of Common Academic Genres: Argumentation and Problem/Solution y
> 3) Testing of Subject Matter, Conceptual Control and Planning (Johns 1991: 172)

Como vemos, este tipo de investigaciones más abiertas y cualitativas permiten captar puntos de vista interesantes y sutiles que van más allá de la identificación de aspectos puntuales que se obtienen a través de las técnicas cuantitativas. En este sentido, Zemelman (1979) afirma que la utilización de técnicas cualitativas ofrece la ventaja de permitir a los entrevistados crear sus propias categorías sobre las que comentar y discutir. Asimismo, la inclusión de preguntas lo suficientemente abiertas en la entrevista facilita la posibilidad de que surjan nuevas ideas y conceptos que no se le habrían ocurrido al entrevistador de otro modo.

En nuestro trabajo, intentamos compatibilizar y enriquecer la investigación cuantitativa rigurosa y empírica con la técnica cualitativa del cuestionario (Ver Cuestionario, Apéndice 1). Esta última técnica nos permitirá establecer los diferentes perfiles docentes, actitudinales y evaluadores de los correctores.

7.2.2.3. Relación entre las puntuaciones y otras medidas de carácter objetivo

Un tercer y último tipo de metodología que recomienda Connor-Linton (1995) en la investigación del proceso de la evaluación es el que ponen en práctica aquellos estudios que analizan las relaciones entre las valoraciones de los candidatos y las medidas objetivas que nos ofrecen ciertas características textuales. Dentro de este tipo de estudios destaca principalmente la investigación en torno al error (*error gravity research*) que estudiaremos en el subepígrafe. 7.2.4.3.

También cabe mencionar aquí los estudios cuyo objetivo primordial es identificar los distintos componentes del ensayo que los correctores subrayan. Entre ellos cabe señalar de manera especial la importancia que se le concede a los aspectos más sobresalientes del ensayo (*salient elements*). Todos estos aspectos se abordarán en el epígrafe (7.2.4).

7.2.3. Factores que influyen en los correctores

En los siguientes subepígrafes abordaremos el estudio de los principales aspectos que afectan la figura del corrector. Estos son:

1. la disciplina académica y la experiencia profesional
2. la experiencia lingüística: nativos *versus* no nativos
3. la personalidad y afectividad: el género y la edad de los correctores.

7.2.3.1. La disciplina académica y la experiencia profesional

La identidad de los correctores desempeña un papel decisivo en la evaluación de los ensayos. Dentro de las investigaciones llevadas a cabo en torno al error, Connor-Linton (1995) asegura, por ejemplo, que: "Judgments of error importance are intimately related to and dependent upon the rater's educational and sociocultural context." (p. 112).

Siguiendo esta misma línea de investigación, Vann Lorenz y Meyer (1991) afirman que la tolerancia que manifiestan los correctores ante los errores cometidos por los candidatos en el ensayo varía de forma considerable según la disciplina académica a la que pertenezcan. En su estudio, estos autores demuestran que los errores de los estudiantes: "...may be tolerated less by faculty in the hard sciences than by those in ESL[2] (included in the Humanities, Social Sciences, and Humanities group in this study)." (Vann Lorenz y Meyer 1991: 193). Vann Lorenz y Meyer (1991) concluyen que la respuesta ante el error no es fortuita y que la disciplina académica, a diferencia de otras variables contextuales potenciales (i.e. edad, el género, etc.), es una medida de predicción significativa en el análisis de las puntuaciones que asignan los correctores a la producción escrita de los candidatos.

Janopoulus (1992) obtiene resultados similares en su estudio. El autor manifiesta que los profesores que trabajan en los departamentos de Ciencias Sociales son, en general, más tolerantes ante los errores cometidos por los estudiantes de segundas lenguas que los profesores pertenecientes

[2] La abreviatura ESL se utiliza para referirse al estudio del inglés como segunda lengua (i.e. *English as a Second Language*).

a otros departamentos como son: Física, Biología, Humanidades, Educación, etc. Janopoulos (1992) concluye:

> These findings lend support to the contention that although university faculty members may be in general agreement concerning the relative gravity of certain error types found within samples of NNS undergraduate writing, they are by no means equally tolerant of those errors. (p. 116)

Por su parte, Brown (1995) descubre patrones de comportamiento distintos en los dos grupos de correctores que investiga en su trabajo (i.e. correctores pertenecientes al sector de la enseñanza y correctores pertenecientes al sector de la industria), aún cuando ambos grupos han experimentado una misma fase de formación previa. La autora afirma que los correctores perciben los criterios evaluativos de forma distinta. Del mismo modo, la aplicación de las escalas de valoración varía de un grupo de correctores a otro, a pesar de que las puntuaciones globales no muestran diferencias significativas. Brown (1995) concluye:

> These different patterns of behaviour appear to reflect different perceptions of the importance of particular features of language and different perception of the tasks. (p. 9)

Como dato a señalar, cabe decir que la mayoría de los estudios contrastivos concluyen que los correctores pertenecientes al sector de la enseñanza, concretamente, los profesores de Inglés alcanzan un grado de fiabilidad mayor que el que obtienen los correctores de otros departamentos

(i.e. Comercio, Leyes, etc.). Este mayor consenso se observa tanto en la toma de decisiones como en la asignación de puntuaciones (Diederich French y Carlton 1961; Michael *et al.* 1980).

Meuffels (1989) corrobora este argumento en su estudio. El autor investiga la fiabilidad de tres grupos de ocho correctores cada uno formados por profesores de lengua, profesores de matemáticas y directores de empresa. Meuffels (1989) demuestra que los profesores de lengua alcanzan un mayor grado de acuerdo en la emisión de sus juicios que el que logran el resto de los grupos, esto es, el grupo de profesores de matemáticas y el grupo de jefes de empresa. Asimismo, los datos parecen indicar que el grupo formado por los profesores de lenguas es el que puntúa más errores lingüísticos y, por consiguiente, es el grupo de correctores más fiable.

Aparte de la disciplina académica, algunos autores destacan los años de experiencia profesional como uno de los factores que afectan de forma más significativa la evaluación de los ensayos. En opinión de Song y Caruso (1996):

> ...the number of years of faculty experience in teaching and holistic rating, rather than background and training seems to be a significant factor affecting holistic scores. (p. 176)

Según estos autores, cuantos más años de experiencia tengan los correctores mayor será la indulgencia o benevolencia que demuestren en la evaluación de los ensayos:

> It seems appropriate to conclude that in this study a rater's experience in teaching and holistic evaluation was a significant factor affecting his or her holistic scores. In specific terms, a rater having a greater number of years in teaching and assessment tended to be more lenient in rating essays by either NES[3] or ESL students. (Song y Caruso 1996: 172)

Song y Caruso sugieren la siguiente explicación para estos resultados:

> However, because the specific reasons for this finding are not clear, the authors can only speculate that as faculty become more experienced in teaching and assessment, they also become more realistic in their expectations of students' performance and less stringent in rating essays. (p. 172)

De modo similar, Cumming (1990) aborda el estudio del efecto de la variable años de experiencia en su estudio. Para ello, el autor analiza el comportamiento de seis profesores expertos que cuentan con una considerable trayectoria profesional en la enseñanza de la expresión escrita en segundas lenguas y el comportamiento de siete estudiantes sin ningún tipo de experiencia previa en la evaluación de la producción escrita. Los resultados indican que las decisiones a las que llegan ambos grupos de correctores muestran diferencias cualitativas considerables. Cumming (1990) concluye:

[3] A partir de ahora las abreviaturas NES se utilizarán para referirse a los estudiantes ingleses nativos (i.e. *native English speakers*).

> Overall, expert teachers appear to have a much fuller mental representation of 'the problem' of evaluating student compositions, using a large number of very diverse criteria, self-control strategies, and knowledge sources to read and judge students' texts. (p. 43)

Sin embargo, Schoonen *et al.* (1997) consideran que en el estudio de los diferentes juicios o valoraciones de aquellos correctores que cuentan con largos años de experiencia profesional frente a aquellos correctores que carecen de esta última (i.e. correctores expertos y correctores legos) hay que tener en cuenta dos factores importantes. Estos factores son: la distinta naturaleza de las tareas que se evalúan y la determinación de los aspectos que se consideren.

Según estos autores los correctores expertos producen, en general, resultados más fiables que los correctores legos en la evaluación del uso de la lengua en las producciones escritas relativamente libres o poco restrictivas, como, por ejemplo, el ensayo. No obstante, los correctores legos producen resultados fiables y consistentes comparables al que pudiera producir cualquier grupo de correctores expertos en la evaluación de aspectos incluidos en tareas específicas y más controladas como, por ejemplo, el *contenido* de los ensayos:

> In sum, the more restrictive the writing and rating task and the easier the text quality to be rated, the smaller the difference we can expect in reading reliability between lay and expert readers. (Schoonen *et al.* 1997: 164)

Por consiguiente, Schoonen *et al.* (1997) concluyen que es necesario matizar algunas de las afirmaciones que sugieren que los correctores expertos son más fiables que los correctores no expertos en la evaluación de la producción escrita:

> Our analyses showed that the findings depend upon (the combination of) the kind of writing and rating task that are used and on the aspect of the text that has to be rated (in our case Content and Usage). (p. 180)

Como explicaremos más adelante, en este trabajo, se intentó abordar el estudio de la variable años de experiencia profesional tal y como se recoge en el cuestionario (ver Apéndice 1). No obstante, los grupos de sujetos eran poco homogéneos y el reducido tamaño de la muestra impedía obtener resultados válidos. Ello motivó que descartásemos su estudio que, sin embargo, queda abierto para futuras investigaciones.

7.2.3.2. La experiencia lingüística: nativos *versus* no nativos

Otra dimensión importante que cabe tener en cuenta en el estudio de la identidad de los correctores es el de su experiencia lingüística. Esto es, la actuación de correctores nativos *versus* no nativos en las evaluación de los ensayos en segundas lenguas.

Como sabemos, en el desarrollo de las pruebas de lengua internacionales como el *First Certificate of English* (FCE), el *Cambridge Advanced Examination* (CAE), el *Test of English as a Foreign* language

(TOEFL), etc. no se permite la formación ni la participación de los correctores no nativos en lengua inglesa.

En las PAAU que se realizan en el estado español, la mayoría de correctores de la prueba de Inglés son profesores no nativos si bien la participación de los correctores nativos es, igualmente, posible. De hecho, dos de los treinta y dos correctores que participaron en nuestro estudio eran profesores nativos. Obviamente, el tamaño de la muestra nos impide realizar cualquier tipo de inferencia seria sobre el comportamiento de estos correctores con respecto al del resto de los correctores que participaron en este estudio. No obstante, no se descarta la ampliación de la muestra en un futuro para la realización de investigaciones posteriores.

La desestimación de la participación de los correctores no nativos en la evaluación de las producciones de lengua internacionales obedece a que su dominio de la lengua inglesa es inferior al que poseen los correctores nativos. Gamaroff (2000) critica, sin embargo, la falta de evidencia empírica que respalde este argumento:

> With regard to the level of English proficiency of raters, it does not follow that because a rater (or anybody else) is not a mother-tongue speaker (of English in this case) that his or her English proficiency is necessarily lower than a mother-tongue speaker of English. (p. 45)

En opinión de este autor, muchos hablantes no nativos poseen un dominio de la lengua inglesa superior al de los hablantes nativos. Ello se debe a que:

> …A major reason for this is not a linguistic one, but because these non-mother-tongue speakers are more academically able, i.e. they have better problem-solving abilities and abilities for learning, and in the case of raters, for assessment. (Gamaroff 2000: 44)

La dicotomía establecida entre hablante nativo / hablante no-nativo fue muy criticada durante los años 70, y a principios de los años 90, de acuerdo con ciertos principios lingüísticos, ideológicos y pragmáticos. No obstante, este binomio ha permanecido hasta nuestros días como un modelo orientativo cuyo último objetivo persigue la adquisición de una competencia lingüística similar a la del hablante nativo. A pesar de que esta dicotomía se considera: "useless as a measure" (Davies 1995: 157), algunos autores argumentan que los correctores nativos y los correctores no nativos enfocan las tareas y, por tanto, evalúan los ensayos desde perspectivas diferentes.

Según estipula Medgyes (1994) en su libro titulado *The Non-native Teacher*, ambos grupos de profesores, nativos y no-nativos, se diferencian básicamente en cuatro aspectos fundamentales:

> 1. Native English speaking teachers (NESTs) and non-native English speaking teachers (non-NESTs) differ in terms of their language proficiency;
> 2. They differ in terms of their teaching behaviour;
> 3. The discrepancy in language proficiency accounts for most of the differences found in their teaching behaviour;
> 4. They can be equally good teachers in their own terms.
> (Medgyes 1994: 357)

La siguiente tabla (Tabla 7.2) nos resume algunas de los diferencias más notables que Medgyes (1994) observa en el estilo de enseñanza de estos dos grupos de profesores (Medgyes 1994: 357):

Tabla 7.2. Diferencias en el estilo de enseñanza que manifiestan los profesores nativos y los profesores no nativos

NESTs	non-NESTs
Own use of English	
Speak better English	Speak poorer English
Use real language	Use 'bookish' language
Use English more confidently	Use English less confidently
General attitude	
Adopt a more flexible approach	Adopt a more guided approach
Are more innovative	Are more cautious
Are less empathetic	Are more empathetic
Attend to perceived needs	Attend to real needs
Have far-fetched expectations	Have realistic expectations
Are more casual	Are more strict
Are less committed	Are more committed
Attitude to teaching the language	
Are less insig*htful*	Are more insightful
Focus on	Focus on
Fluency	accuracy
Meaning	form
Language in use	grammar rules
Oral skills	printed word
Colloquial registers	formal registers
Teach items in context	Teach items in isolation
Prefer free activities	Prefer controlled activities
Favour groupwork / pairwork	Favour frontal work
Use a variety of materials	Use a single textbook
Tolerate errors	Correct / punish for errors
Set fewer tests	Set more tests
Use no / less L1	Use more L1
Resort to no / less translation	Resort to more translation
Assign less homework	Assign more homework
Attitude to teaching culture	
Supply more cultural information	Supply less cultural information

En un estudio posterior, Árva y Medgyes (2000) van a confirmar que, efectivamente, los profesores nativos y no-nativos muestran comportamientos claramente diferentes en el aula. Medgyes (1994) comenta, sin embargo, que: "Different does not imply better or worse." Según

Árva y Medgyes (2000) las diferencias que se observan en la actuación de ambos grupos de correctores obedece, en parte, a la clara superioridad del dominio de la lengua inglesa que tienen los profesores nativos. Los autores concluyen:

> In any case, it is reasonable to assume that the respective teaching behaviour of NESTs[4] and non-NESTs is connected with linguistic matters, and at least some of divergences perceived between the two cohorts are determined by their divergent language backgrounds. (p. 364)

A la luz de estos resultados, y desde un punto de vista pedagógico, Árva y Medgyes (2000) aseguran que: "it appears to be a fair assumption that even untrained NESTs can be used effectively for certain teaching purposes." (p. 369). Estos autores señalan que los correctores no-nativos adecuadamente formados en el manejo de las técnicas de evaluación pueden llegar a convertirse en mejores correctores que los correctores nativos no formados (ver van Essen, 1994). Sin embargo, es cierto que: "Poorly qualified NESTs can do a decent job as long as they are commissioned to do what they can do best: converse." (Árva y Medgyes 2000: 369). En cualquier caso, se postula que los profesores debieran contratarse por su calidad profesional y no, simplemente, por su dominio lingüístico.

[4] Las abreviaturas NESTs (*native English second language teachers*) y non-NESTs (*non-native English second language teachers*) se utilizan para referirse a los profesores de segundas lenguas nativos y a los profesores de segundas lenguas no nativos respectivamente.

Los resultados que nos ofrece la literatura en torno a la participación de los correctores nativos *versus* no nativos en la evaluación de los ensayos han sido hasta ahora ambiguos y contradictorios. Como vimos, la hipótesis principal que se defiende es que los dos grupos de correctores poseen diferentes percepciones sobre la calidad de las producciones escritas de los candidatos y, en consecuencia, sus valoraciones producirán puntuaciones distintas.

Sin embargo, Carlson *et al.* (1985) afirman que las puntuaciones de los profesores nativos y no nativos en la evaluación de los ensayos escritos, tanto por estudiantes nativos como por estudiantes no nativos, no registran, en general, diferencias significativas. Las correlaciones entre ambos grupos de correctores oscilan entre 0,65 y 0,72.

Estos resultados se confirman en estudios más recientes (Brown 1991; Purpura 1992) que subrayan, en general, la similitud de las puntuaciones otorgadas por ambos grupos de correctores. Debido a ello, algunos autores reivindican la admisión y participación de los correctores no nativos en la evaluación de las pruebas de producción escrita internacionales. Brown (1991) expone al respecto:

> The results show that on the most crucial issues pertaining to acceptability as raters (reliability, bias, ranking of candidates), given adequate training and explicit assessment criteria, there is little evidence that native speakers are more suitable than non-native speakers. (p. 13)

Es importante hacer notar que no en todas las investigaciones llevadas a cabo sobre el efecto que ejerce la experiencia lingüística de los correctores en la evaluación de los ensayos se tienen en cuenta las mismas variables (i.e. la lengua objeto, la formación previa de los correctores, etc.). A pesar de ello, la mayoría de los estudios concluyen que los correctores no nativos se muestran menos tolerantes ante los errores cometidos por los estudiantes de segundas lenguas que los correctores nativos (James 1977; Sheorey 1985; Santos 1988).

Así, por ejemplo, Fayer y Karsinki (1987), en un estudio sobre las reacciones de los correctores ante las producciones escritas en lengua inglesa de los hablantes puertorriqueños, concluyeron que el grupo de correctores no-nativos, formado por un grupo de correctores españoles, se mostró menos tolerante que el grupo de correctores ingleses nativos. Asimismo, Santos (1988) pide a un grupo de profesores universitarios que evalúen la producción escrita de dos estudiantes no nativos haciendo uso de una escala analítica. De nuevo, se demuestra que los hablantes no nativos son más severos en la evaluación de las producciones escritas de los estudiantes.

Por su parte, Hill (1996), en un interesante estudio, nos recuerda la necesidad de interpretar los resultados de los estudios dentro del contexto adecuado. La autora investiga las puntuaciones de un grupo de correctores no nativos (i.e. correctores indonesios) frente a un grupo de correctores nativos (i.e. correctores australianos) en la evaluación de la competencia de la lengua inglesa en Indonesia. Hill (1996) concluye:

> ...although the Australian group were more experienced, the Indonesian group were more in agreement (i.e. had a narrower range of harshness estimates). (p. 287)

Esta autora enmarca su trabajo dentro del contexto específico que representa Indonesia. En dicho país se intenta defender la autonomía de la variedad indonesa de la lengua inglesa y se lucha en contra del modelo de lengua inglesa imperante basado en el hablante nativo ideal y la denominada *native-like competence*. Hill (1996) asegura que si se aceptan otros modelos lingüísticos como son las diversas variedades de la lengua inglesa más allá del *Inner Circle*[5], o países donde la lengua inglesa es la lengua materna o L1[6] (por ejemplo, Gran Bretaña, Estados Unidos, Canadá y Australia), como modelos apropiados para evaluar la lengua inglesa, entonces, los resultados de muchos de los estudios hasta ahora realizados pueden aportarnos conclusiones diferentes. Hill (1996) observa:

> Contrary to the findings of earlier studies comparing native and nonnative speakers raters, this study found that, when using a non-native speaker ideal and trained raters, it was the native speakers who tended to be harsher overall. (p. 288)

7.2.3.3. Personalidad y afectividad

Cuando hablamos de factores personales y afectivos nos referimos a aspectos de la persona, emociones y actitudes que afectan nuestro

[5] Ver Kachru (1888)

comportamiento. Dentro de este grupo, se incluyen reacciones a factores como son: la edad, la apariencia física, la personalidad o el género, tanto de los correctores como de los candidatos.

La influencia que dichas variables ejercen en las puntuaciones de los correctores ha sido investigado extensamente en la literatura, especialmente en la evaluación de la producción oral de los estudiantes (Brown y Levinson 1978; Porter 1991a; 1991b; Coates 1993; Celce-Murcia 1997; O'Sullivan 2000,etc.).

En el campo de la producción escrita, sin embargo, no se han podido obtener datos concretos sobre la interacción directa que se produce entre los correctores y los candidatos aunque los efectos de algunas variables (i.e. el género, la edad, la experiencia personal...) sobre las valoraciones de los correctores han sido objeto de numerosos estudios (Baird 1988; Vann, Lorenz y Meyer 1991; Herrera 2000a). Los resultados obtenidos no han sido, sin embargo, concluyentes.

En los subepígrafes siguientes comentaremos algunos de los estudios realizados en torno a las variables género y edad de los correctores. Como ya dijimos, el reducido tamaño de la muestra impidió el estudio de la variable edad en este trabajo. No obstante, la variable género fue relevante en nuestra investigación. Los resultados obtenidos se discutirán en la parte empírica de este trabajo.

[6] La abreviatura L1 hace referencia a la lengua materna o nativa. La L2 se utiliza para referirse a la segunda lengua.

7.2.3.3.1. El género

El efecto que ejerce el factor género en la evaluación, tanto de la producción oral como de la producción escrita, se ha abordado desde perspectivas diferentes. La teoría más extendida es que las diferencias atribuidas al género de los correctores vienen motivadas fundamentalmente por dos tipos de factores: los factores innatos o los factores psicológicos o socioculturales (ver Herrera, 2000a).

Dentro del análisis de la conversación, varios estudios como el de Locke (1984) o Porter (1991a; 1991b) demuestran que, en las pruebas de producción oral, los participantes de la interacción obtienen resultados más altos cuando son entrevistados por hombres que cuando, por el contrario, son entrevistados por mujeres. Conviene hacer notar que todos los sujetos que participan en estos estudios comparten la misma nacionalidad y, por tanto, se mueven en contextos más o menos homogéneos. Se trata de estudiantes árabes en el caso de Locke (1984) y Porter (1991a) y de estudiantes argelinos en el caso de Porter (1991b).

Una vez que se alteran los factores contextuales y se mezclan sujetos de diversas nacionalidades se obtienen resultados algo distintos. Se mantienen, por una parte, la relevancia del factor género, que es significativo desde el punto de vista estadístico, pero, por otra parte, los candidatos consiguen puntuaciones más altas al ser entrevistados por mujeres (O'Sullivan, 2000). Según O'Sullivan (2000), la explicación de resultados tan contradictorios se debe a las diferencias culturales de los candidatos:

> This difference in results may, of course, be ascribed to differences in rater behaviour, though it seems likely that the difference in cultural background of the subject is the more probable explanation _ particularly in view of the reaction of different cultural groups to mixed-sex interactions. (p. 375)

Celce-Murcia (1997) comparte este punto de vista y asegura que las diferencias que se hallan en las interacciones entre hombres y mujeres se vinculan tanto a aspectos relacionados con el factor género como a aspectos socio-culturales.

Dentro de la investigación de la producción oral, otro grupo de estudios analiza las diferencias de estilo en la expresión oral de hombres y mujeres (Fishman 1978; Coates 1993). Algunas investigaciones (Gass y Varonis 1986; Pica *et al*. 1989; Shehadeh 1999) demuestran que el hombre posee un mayor dominio de la conversación que el que posee la mujer.

Por su parte, Shehadeh (1999) sugiere que los hombres y las mujeres utilizan la conversación con fines distintos. Así, los hombres harían uso de la conversación para promocionar su habilidad de producción mientras que las mujeres lo harían para promocionar su habilidad de comprensión. Otros autores como Berninger *et al.* (1996) señalan, por el contrario, que el discurso oral de las mujeres posee una mayor fluidez que el de los hombres. Sin embargo, la calidad del discurso de los hombres no es, por ello, inferior.

Asimismo, Brown y Levinson (1978) observan que las mujeres tienden a expresarse con un mayor nivel de educación (i.e. *politeness*) y ofrecen un mayor grado de apoyo (i.e. *support*) al interlocutor en la conversación. Esto se demuestra a través del reconocimiento y de la reconstrucción de las

intervenciones producidas por los distintos participantes en la interacción oral. En esta línea de trabajo, cabe situar los resultados obtenidos por O'Sullivan (2000) que concluye que los sujetos no tan sólo obtienen resultados más altos al ser entrevistados por mujeres sino que, además, dichos sujetos aumentan la calidad y la corrección de su producción oral.

Según nos explica Porter (1991a), es muy probable que algunos de los factores que son significativos en la evaluación de la producción oral de los candidatos sean también relevantes en la evaluación de la producción escrita. Cabe, por otro lado pensar, que en la investigación de la producción escrita pueden surgir otro grupo de factores afectivos, distintos a los estudiados en la producción oral, que sean significativos en la evaluación de los ensayos.

En cualquier caso, la influencia de los aspectos afectivos, tanto en la producción oral como en la producción escrita de los candidatos, no puede obviarse:

> Nevertheless, a description of a learner's linguistic performance which ignored this dimension of complexity would be leaving out of account something important. (Porter 1991: 39)

En las PAAU, conscientes de la importancia que juegan los factores afectivos en el proceso de evaluación, se omite cualquier tipo de información sociométrica o personal relativa a los estudiantes ya sea, el nombre, el género, la edad de los candidatos, datos familiares, etc.

En el campo de la producción escrita, la influencia que ejerce el factor género en las puntuaciones de los correctores no se ha podido determinar con exactitud. Goddart-Spear (1983) afirma en su estudio que tanto los hombres como las mujeres otorgan puntuaciones más bajas a una misma pieza de producción escrita de carácter científico cuando su autoridad se atribuye a las mujeres que cuando se atribuye a los hombres. Estos últimos grupo obtienen puntuaciones más altas.

Vann Lorenz y Meyer (1991) definen ciertas variables, entre las cuales se incluye el género, como: “weaker predictors of response”. Sin embargo, estos autores observan una tendencia general de mayor tolerancia en la actitud de las mujeres frente a los errores de los estudiantes. Según Vann Lorenz y Meyer (1991):

> Women reported being less irritated by and more accepting of the erroneous language than did their male colleagues. Although not statistically significant, this pattern of greater tolerance suggests a potentially interesting area for future research. (p. 191)

Los resultados de este estudio contrastan con los de Herrera (2000a). Este autor en un estudio sobre la influencia del factor género y situación laboral de los correctores en la evaluación de las pruebas de Selectividad demuestra que, a pesar de la similitud observada en el comportamiento del profesorado de universidad, tanto en hombres como en mujeres: “leniency is greater among men than among women both in the objective and in the subjective subtests.” (Herrera 2000a: 18)

Por otra parte, autores como Baird (1988) consideran que no existe evidencia empírica suficiente que avale la influencia de la variable género en el comportamiento de los correctores. Por el contrario, Gyagenda y Engelhard (1998) obtienen resultados que demuestran la influencia significativa de esta variable en la evaluación de la producción escrita.

A la luz de resultados tan contradictorios, podemos señalar que la relevancia del factor género en las puntuaciones de los correctores está aún por determinar. En la parte empírica de nuestro trabajo intentaremos corroborar su grado de significación en las puntuaciones holísticas y analíticas de los correctores en la evaluación de los ensayos de Selectividad.

7.2.3.3.2. La edad

Otra de las variables afectivas que se ha investigado en la literatura de la evaluación es la influencia del factor edad en el comportamiento de los correctores. A pesar de que los resultados de los estudios no han sido concluyentes se han observado ciertas tendencias. Así, Vann Lorenz y Meyer (1991) exponen en su estudio que, aunque el factor edad no resultó ser estadísticamente significativo: “there was a clear pattern of increasing tolerance with age.” (p. 191).

Asimismo, Santos (1988) demuestra que la edad de los correctores es un factor significativo en la evaluación de algunos aspectos lingüísticos de los ensayos de los estudiantes. Según Santos (1988): “the older professors rated the language less irritating than did the younger professors.” (p. 84). La hipótesis que se apunta es que:

> One can only speculate why this might be so, for example, perhaps professors, as they become older, become more realistic in their expectations of students' performance and thus more tolerant. (p. 85)

Diversos autores han admitido y reconocido la relevancia de las variables personales y afectivas en el proceso de evaluación (Newcomb 1977; Vann Lorenz y Meyer 1984; Hamp-Lyons 1990; Herrera 2000a). No obstante, se necesitan investigaciones posteriores para determinar la medida en que dichas variables afectan las puntuaciones de los correctores. A pesar de ello, Porter (1991a) insiste en la necesidad de tener en cuenta los factores afectivos en cierto tipo de pruebas entre las que podríamos incluir las PAAU. Su postura se justifica en la siguiente cita:

> ...where the test is a large one, where the results can affect the course of lives or entail the expenditure of large sums of money, and where specifiable affective factors are known to have significant effects on linguistic performance, it could be costly to ignore them. (Porter 1991a: 40)

7.2.3.4. El orden de los ensayos o efecto contraste

El efecto significativo que ejerce sobre los correctores el orden en el que aparecen los ensayos (i.e. el efecto contraste) es otra de las variables que se consideran críticas en la evaluación de la producción escrita (Hales y Tokar 1975; Hughes, Keeling y Tuck 1980).

Algunos autores señalan que los ensayos se comparan inevitablemente los unos con los otros. De este modo, el hecho de que un *buen* ensayo aparezca en primer o último lugar afectará sensiblemente la evaluación del resto de los ensayos ya que estos últimos tienden a compararse con el primero. En este sentido, se puede decir que el primer ensayo determina, hasta cierto punto, el nivel de actuación de los correctores o los criterios de evaluación que se aplicarán al resto de las producciones escritas.

Milanovic *et al*. (1996) observan el efecto contraste en su estudio y comentan:

> Raters effects which are well documented elsewhere in the literature were also observed in this data. Sequencing of scripts seemed to have an influence on some markers who appeared to be judging the level by comparing it to the previous one marked. (p.106)

Vaughan (1991) argumenta que los criterios comparativos que se establecen en la lectura holística de los ensayos son inevitables:

> …Furthermore, when papers are read quickly, one after another, as they are in a holistic assessment session, they become, in the rater's mind, one long discourse. For example, seven of these raters made comparative statements as they read, such as, "This essay is better/worse than the previous one or the others. (p. 121)

Por este motivo, la autora nos recomienda considerar esta variable y el efecto potencial de la misma en la evaluación de las diferentes producciones escritas de los candidatos:

> Another aspect of the reading session that should therefore be taken into consideration is the environment of the essays_ the effect of the papers taken as a whole on each other. (Vaughan 1991: 121)

La observación del efecto contraste en los estudios anteriores explica el hecho de que algunos autores soliciten que la evaluación de los ensayos escritos por los estudiantes de segundas lenguas se realice de forma independiente a la de los ensayos escritos por estudiantes nativos. Sweedler-Brown (1993) expone:

> If NS[7] and ESL essays are graded separately, ESL writers will be compared with other ESL writers, and graders will more easily recognize degrees of language control among them. But when ESL essays are graded side by side with NS essays, the language abilities of the two are inevitably compared, and frequency of error in ESL essays becomes particularly conspicuous. (p. 13)

El efecto rebote no se ha investigado en nuestro trabajo dado que ello sería motivo de una nueva tesis. No obstante, se han observado constantes referencias contrastivas y algunos criterios comparativos que establecen los correctores de nuestro estudio entre unos ensayos y otros. Estos criterios se reflejan de forma clara en la descripción de los elementos positivos y negativos que se realizan sobre cada uno de los ensayos (ver Apéndice 2).

[7] La abreviatura NS (*native speakers*) se refiere a los hablantes nativos.

7.2.4. Elementos de los ensayos que los correctores destacan en la evaluación

La escasa fiabilidad que se observa en las puntuaciones holísticas de los correctores (i.e.fiabilidad inter-evaluador) se ha tratado de subsanar o, al menos, minimizar a través de formas diversas. Como ya explicamos en el capítulo 6, entre las distintas medidas que se proponen, algunos autores aconsejan la utilización de escalas de valoración holísticas. Estas escalas servirían de guía a los correctores en sus evaluaciones de forma que se lograrían aumentar los niveles de consistencia y de fiabilidad de las puntuaciones (Vaughan 1991; Brown 1992; Shavelson e*t al.* 1992).

No obstante, para poder construir una escala estándar que se adapte a cualquier situación específica de evaluación, se hace necesario identificar los elementos de los ensayos que los correctores destacan en sus evaluaciones. Es decir: “what raters *really* pay attention to” (Pollit y Murray, 1996).

En opinión de Schoonen *et al*. (1997), las producciones escritas de carácter restrictivo facilitan la elaboración de escalas de valoración. Sin embargo, la construcción de escalas de valoración para pruebas de producción escrita abiertas, como es el caso del ensayo, constituye una tarea mucho más compleja:

> When a restricted range of ‘written responses’ can be expected, detailed scoring guides can be developed. In free-writing assignments such scoring guides can hardly be imagined. (Schoonen *et al*. 1997: 163-164)

La construcción de una escala común de valoración que se pueda aplicar a los diferentes contextos evaluativos requiere tener en cuenta las percepciones que poseen los correctores sobre los diversos elementos de los ensayos. Este argumento se ha utilizado en defensa de los métodos de evaluación analíticos ya que estos últimos facilitan la identificación de las distintas orientaciones que manifiestan los correctores en el proceso de evaluación (Brown 1991; O'Loughlin 1994).

Pollit (1996) considera que la percepción de los elementos de los ensayos que contribuyen, en mayor o menor medida, a definir el dominio de la producción escrita es un fenómeno totalmente subjetivo en el que entran en juego distintos constructos individuales. Con respecto a este último punto, Kelly (1955), padre de la psicología constructista, sostiene que la realidad de cada individuo lo constituye el universo tal y como este último lo percibe, es decir: "reality is subjective rather than objective" (Kelly, 1955). De acuerdo con esta premisa, lo que los investigadores tratan de averiguar es de qué forma los distintos correctores perciben o construyen el mundo:

> ...The repertory of constructs tells us what a subject sees in the world, what is salient, and so offers an insight into the thinking processes in a procedure like assessment. (Pollit 1996: 77)

Sin duda alguna, tratar de discernir los distintos constructos personales que aplican los correctores en sus evaluaciones de la producción escrita resulta complejo. Como vimos (subepígrafe 7.2.2.), han sido varias las técnicas que se han diseñado para la colección y búsqueda de los

elementos de los ensayos que centran la atención de los correctores: entrevistas personales, entrevistas colectivas, cuestionarios retrospectivos, *think-aloud protocols*, métodos cuantitativos (i.e. correlaciones, técnicas de regresión), etc.

No obstante, la gran variedad de respuestas y de resultados que se obtienen con la implementación de estas técnicas dificulta su sistematización. Algunos estudiosos se muestran asombrados ante la diversidad de elementos que se recogen:

> What is most striking about the composition elements markers focused on is their diversity...Some markers concentrated almost exclusively on a paper's bad points while others were equally prepared to mention those aspects which credited a candidate. (Milanovic *et al.* 1996:100-101)

A título ilustrativo, cabe señalar la gran discrepancia que se observa en los resultados de algunos estudios que tratan de identificar los componentes del ensayo que determinan sus puntuaciones finales. Freedman (1979), por ejemplo, demuestra que el *contenido* es el factor más significativo en la valoración global del ensayo. Grobe (1981), por su parte, manifiesta que los correctores se muestran mayormente afectados por la riqueza y diversidad léxica del ensayo. Esta idea es la que defiende Santos (1988) que afirma que la presencia de errores léxicos determina las puntuaciones finales de los correctores.

Por el contrario, Stewart y Grobe (1979) manifiestan que los dos factores decisivos en la evaluación de los ensayos son: la extensión del texto

y la corrección u omisión de errores en la escritura. Este último aspecto se confirma en otros estudios que señalan que la omisión de errores constituye el factor más relevante en la evaluación de la expresión escrita (Mullen 1980; Homburg 1984;McDaniel 1985; Sweedler-Brown 1993).

En la parte empírica de nuestro trabajo trataremos de confirmar los resultados de algunos de los estudios anteriores. Se investigarán los elementos de los ensayos que afectan en mayor medida la puntuación holística final de los ensayos y la influencia del factor error.

Milanovic *et al.* (1996) identifican hasta once elementos que los correctores subrayan en los ensayos:

1. Length
2. Legibility
3. Grammar
4. Structure
5. Communicative effectiveness
6. Tone
7. Vocabulary
8. Spelling
9. Content
10. Task realisation
11. Punctuation

(Milanovich *et al.*1996: 101)

No obstante, la puntuación holística final que asignen los correctores a los ensayos vendrá condicionada no tan sólo por los diversos elementos que se destaquen sino también por el peso o valor individual que se le atribuya a cada uno de ellos.

Weir (1988) investiga los aspectos de la producción escrita que centran la atención de los correctores en una prueba de inglés para fines académicos (*Test in English for Educational Purposes*, TEEP). A partir de una serie de cuestionarios y entrevistas realizadas a diversos profesores y estudiantes, Weir (1988) confecciona una lista de los elementos que definen los principales constructos de producción escrita que se evalúan en el mundo académico.

La lista de elementos sobre la que los profesores deben pronunciarse es la siguiente:

1) grammatical accuracy
2) variety of grammatical structures employed
3) appropriateness of grammatical structures employed
4) appropriateness of vocabulary
5) range of vocabulary
6) the subject content
7) clarity of expression
8) arrangement and development of written work
9) spelling
10) punctuation
11) handwriting
12) tidiness

(Weir 1988)

Los resultados que obtiene Weir (1988) revelan la gran disparidad de criterios que se utilizan para juzgar la importancia que se le concede a los distintos elementos de los ensayos. Entre las diversas tendencias que se observan, este autor destaca la variedad de concesiones que los correctores otorgan especialmente a los estudiantes extranjeros en la evaluación de la producción escrita. Weir (1988) manifiesta que existe un claro contraste entre aquellos correctores que realizan concesiones a los estudiantes y

aquellos que, por el contrario, las descartan. Esta diversidad de concesiones impide que se puedan establecer generalizaciones sobre la importancia que los correctores atribuyen a los elementos concretos de la producción escrita:

> We would argue strongly that this variety precludes the possibility of making any valid generalizations concerning tolerance levels that operate in the written medium on the part of staff and must bring into question the findings of both Carroll (1977) and Munby (1977). In practice establishing these tolerance conditions is by no means as easy as they had assumed. (Weir 1988: 24).

La conclusión a la que llega Weir (1988) es poco optimista: “It is an unrealistic task to speculate on the tolerance conditions that will apply in the assessment of written work.” (p. 25).

Siguiendo la misma línea de investigación, Brown (1991) les pide a un grupo de correctores que indiquen el mejor y el peor aspecto de cada uno de los ensayos a medida que los van corrigiendo. Para ello, se les ofrece una lista de elementos de entre los que poder elegir: *cohesión*, *contenido*, *mecánica*, *organización*, *sintaxis* y *vocabulario*. El estudio de estos elementos le permite a la autora comprobar que los correctores emiten sus puntuaciones desde perspectivas muy diferentes, a pesar de que los resultados holísticos obtenidos son, en general, similares desde el punto de vista cuantitativo. Entre sus resultados, Brown (1991) descubre que el *contenido* se considera un elemento fundamentalmente positivo tanto para los profesores nativos como para los profesores no nativos de lengua inglesa. No obstante, la valoración de otros elementos revela grandes discrepancias entre ambos grupos de correctores. Brown (1991) concluye:

> The best and worst feature analyses indicate that both faculties attend to Content as a primary positive feature. Yet, the English faculty members appear to pay more attention to Cohesion and Syntax than do ESL (English second language) raters, while the latter group appears to consider Organization more important. In terms of negative features, both groups seem to attend to Syntax as a primary negative feature with Mechanics being of somewhat more interest to English faculty and Content being of more interest to ESL raters. (p. 601)

Estos resultados le permiten a autora inferir que la similitud de las puntuaciones holísticas no implica que el camino escogido para llegar a definirlas sea el mismo:

> It appears that, on average, English and ESL faculty assign very similar scores_ regardless of differences in background or training. They may, however arrive at those scores from somewhat different perspectives. (Brown 1991: 601)

Como veremos en la parte empírica de nuestro trabajo, el estudio de Brown (1991) nos va a servir de punto de referencia para investigar e identificar los mejores y los peores elementos de los ensayos de Selectividad que los correctores destacan. Las puntuaciones holísticas que se otorgan al ejercicio del ensayo de Selectividad no nos permiten obtener dicha información dado que estas últimas no reflejan las distintas orientaciones que manifiestan los correctores.

En un estudio similar al de Brown (1991), Connor-Linton (1995) asegura que la consistencia de las puntuaciones holísticas de los correctores: “does not guarantee that they arrived at those same scores by the same route, by applying the same standards.” El autor añade:

> In fact, an important methodological implication of this result is the question whether *quantitative similarities* (in average ratings, for example) may mask *qualitative differences* in the ways by which those ratings were determined. (Connor-Linton 1995: 103)

Gamaroff (2000) confirma estos últimos resultados en un estudio reciente. El autor investiga la relación que se da entre los resultados holísticos de las pruebas y los criterios individuales que aplican los correctores. La conclusión a la que llega es la siguiente:

> Similar scores between raters do not necessarily mean similar judgements, and also, different scores between raters do not necessarily mean different judgements. (Gamaroff 2000: 42)

En los siguientes subepígrafes analizaremos los estudios que examinan la relevancia de los elementos de los ensayos según su nivel de dominio lingüístico. Asimismo, se comentarán los resultados obtenidos en torno a la incidencia de los elementos sobresalientes (*salient elements*) en la producción escrita. Posteriormente, se abordará el estudio del error y la reacción de los correctores ante el mismo. Finalmente, analizaremos los aspectos formales y de *contenido* que los correctores subrayan en la evaluación de los ensayos. Todos estos aspectos son relevantes desde el punto de vista pedagógico ya que nos permiten adecuar el proceso de retroalimentación o *feedback* a las necesidades de nuestros estudiantes.

7.2.4.1. Niveles de dominio lingüístico de los ensayos

Algunas investigaciones parecen demostrar que la atención que prestan los correctores a los elementos o categorías de los ensayos se adecua a los diferentes dominios de la lengua que poseen los candidatos. Según este planteamiento, los correctores obviarían o enfatizarían ciertos elementos en las distintas producciones escritas manifestando comportamientos distintos en cada caso.

Con el fin de ratificar esta hipótesis, Brown (1991) clasifica los ensayos de su estudio en tres niveles distintos de dominio de la lengua inglesa: bajo (*low*), medio (*middle*) y alto (*high*). Los resultados que obtiene esta autora señalan que, por lo que respecta a los elementos positivos de los ensayos, la *cohesión* tiende a valorarse como un elemento positivo, principalmente en los ensayos con un nivel de dominio de la lengua bajo. El *contenido*, por otro lado, destaca como un elemento positivo en los ensayos con un dominio de la lengua alto. Por último, la *mecánica* únicamente se juzga en los ensayos que demuestran un nivel de competencia lingüística bajo o medio.

Por lo que respecta a los elementos negativos, el *contenido* y la *mecánica* son los que mayor consideración reciben en la evaluación de los ensayos. Además, estos elementos son los únicos que muestran diferencias significativas. El *contenido* aparece como el elemento más destacado en los ensayos con un nivel de dominio de la lengua inferior. La *mecánica*, por su parte, se considera relevante en los ensayos con un nivel de dominio de la lengua bajo y medio.

Siguiendo esta misma línea de actuación, Milanovic *et al.* (1996) clasifican los ensayos según los diferentes niveles de dominio lingüístico que demuestran. Las conclusiones a las que llegan estos autores contradicen los resultados de Brown (1991). Según Milanovic *et al.* (1996) las categorías que mayor atención reciben en los diferentes ensayos son las siguientes:

> With regard to proficiency level it was noted that, with higher level scripts (CPE), markers focused more on vocabulary and content, whereas with the intermediate level scripts (FCE)[8], markers focused more on communicative effectiveness and task realisation. (p. 106)

Connor-Linton (1995) utiliza la misma metodología de trabajo en su estudio y de nuevo divide los ensayos en tres niveles de dominio lingüístico: bajo, medio y alto. Este autor afirma que se observa una mayor homogeneidad en las razones cualitativas apuntadas por los correctores con respecto a aquellos ensayos que muestran un nivel de competencia lingüística bajo o alto. Por el contrario, el grado de desacuerdo entre los correctores es elevado en la evaluación de los ensayos con un nivel de competencia lingüística medio. Connor-Linton (1995) concluye que la diversidad numérica de las razones cualitativas atribuidas a los diferentes ensayos sugiere que los ensayos con un dominio de la lengua medio son los que presentan una evaluación más problemática:

[8] La abreviatura CPE se refiere a la prueba: *Cambridge Proficiency Examination*. Por su parte, la abreviatura FCE se refieren al *First Certificate of English*.

> This suggests that it was easier for both groups of teachers to provide explicit reasons for their ratings of the weakest (and, to a lesser extent, the strongest) essays. And this in turn, suggests that writing from students with a mid-range proficiency may be most difficult to assess and diagnose. (Connor-Linton 1995: 110)

Vaughan (1991) comparte esta última opinión y subraya la dificultad que conlleva la evaluación de los ensayos con un dominio lingüístico medio o *bordeline cases*:

> It may be that, although some essays can fairly clearly judged according to holistic assessment guidelines, many fall between the cracks, perhaps having clear organization but simplistic sentence structure and weak content, for instance. With these bordeline cases, raters may be more apt to fall back on their own styles of judging essays. (p. 121)

En nuestro trabajo seguiremos el planteamiento metodológico de los estudios anteriores. De este modo, tendremos la oportunidad de comprobar si los criterios evaluativos que se aplican se adecuan al nivel de dominio lingüístico de los ensayos.

7.2.4.2. Elementos sobresalientes

Diversos autores defienden que la clasificación de los ensayos en sus distintos niveles de dominio de la lengua se vincula estrechamente a la aparición de ciertos elementos de los ensayos que destacan o *sobresalen* (*salient elements*) (Stewart y Grobe 1979; Charney 1984; Jonopoulos 1992; Song y Caruso 1996). Entre estos últimos cabe citar: la presentación

(*handwriting*), la extensión del ensayo (*length*), el número de errores de escritura (*spelling*), etc. Dichos elementos influirían decisivamente en la evaluación holística de los ensayos.

Paradójicamente, la mayoría de los elementos que *sobresalen* o que destacan en los ensayos se consideran irrelevantes desde el punto de vista del aprendizaje de la producción escrita. Para evitar la influencia de este fenómeno, algunos autores recomiendan leer los ensayos de forma rápida. McColly (1970), por ejemplo, aconseja un ritmo de lectura de unas 400 palabras por minuto.

Vaughan (1991) confirma en su estudio que los correctores mencionan de forma constante ciertos elementos sobresalientes de los ensayos a pesar de que dicho elementos no figuran en los criterios de evaluación que se aplican en ese momento concreto:

> In this study, handwriting was one of the most frequently cited problems; the longest essay was passed by everyone; and the major reason cited by most raters for passing essay B was its unique use of an extended metaphor. But these features are not mentioned in the guideline characteristics. (p. 121)

Dentro de la evaluación de la producción oral, Pollitt y Murray (1996), en un interesante estudio, mantienen que la relevancia que adquieren ciertos aspectos viene determinada por el nivel de dominio lingüístico de los interlocutores. Si aplicamos este razonamiento a la evaluación de la producción escrita, probablemente podamos entender el motivo por el que el uso correcto de una forma léxica *sobresale*, por ejemplo, en un ensayo de

dominio lingüístico bajo y, sin embargo, esa misma forma léxica no atrae la atención de los correctores en un ensayo de dominio lingüístico alto. Pollitt y Murray (1996) explican:

> It seems likely that those performance characteristics associated by judges with the lower end of the scale become increasingly less evident at the higher end because they are less problematical for candidates and therefore less salient...Any increase in form-focus will be proportional to decreases in content-focus. (p. 76)

Como se ha podido apreciar, son muchas las variables que parecen afectar las evaluaciones holísticas de los correctores. La utilización de métodos holísticos como los que se aplican en la corrección de los ensayos de Selectividad no nos permite identificar los componentes que determinan las puntuaciones globales. Esto se debe a: "...the 'uncommunicative' nature of holistic rating." (Elbow 1993). Este hecho nos ha llevado a reconsiderar la introducción de categorías analíticas en este trabajo (ver Apéndice 2). Dichas categorías facilitan la identificación de los elementos que centran el interés de los correctores a la vez que nos suministran información sobre el valor o el peso que se les otorga a cada una de ellas en la evaluación del ensayo.

7.2.4.3. La evaluación del error

La evaluación del error ha suscitado un interés especial dentro del campo de LT. De hecho, son muchos los estudios que confirman que la falta u omisión de errores constituye el factor más decisivo en la evaluación de la

expresión escrita (Stewart y Grobe 1979; Mullen 1980; Homburg 1984; McDaniel 1985; Sweedler-Brown 1993).

La evaluación de los errores se aborda desde perspectivas diferentes. Algunos estudios se muestran interesados en la medición de la reacción que manifiestan los correctores ante los errores cometidos por los candidatos (Santos 1988; Vann Meyer y Lorenz 1991). Desde esta perspectiva, la evaluación del error se entiende como la reacción subjetiva de los profesores frente a los errores cometidos por los estudiantes no nativos en la producción escrita. Santos (1988) define la evaluación del error, en términos generales, como: "NSs' reactions to NNSs' errors." (p. 70) [9].

Los tres principales criterios que se han estudiado para evaluar los errores son los siguientes:

1) *Comprensibilidad*. Este criterio mide hasta qué punto el interlocutor entiende lo que se le dice o está escrito;

2) *Irritabilidad*. Se define como:

> the result of the form of the message intruding upon the interlocutor's perception of the communication...The irritation continuum ranges from unconcerned, undistracted awareness of a communicative error to a conscious preoccupation with form (Ludwig 1982: 275).

3) *Aceptabilidad*. Este criterio determina hasta qué punto el interlocutor considera que el mensaje verbal o escrito del sujeto no nativo se aproxima a las normas establecidas de la lengua objeto.

Con estos criterios en mente, se pretende elaborar un marco de referencia que nos ayude a determinar la seriedad de los diversos errores y la importancia que se le concede a cada uno de ellos. Así, se logrará establecer la gravedad de cada error de forma individual. No obstante, la selección de los criterios evaluativos que se han de aplicar para juzgar la importancia de los errores ha suscitado cierta polémica.

Autores como Politzer (1978), Albrechtsen *et al*. (1980) y Chastain (1980) consideran, por ejemplo, que el criterio más relevante para determinar la gravedad de un error es el de la *comprensión*. Dordick (1996) defiende este criterio con vehemencia:

> In order to determine which types of error require greater attention, a sound frame of reference for error gravity must be established. To do so, we must first consider the ultimate purpose of language, which simply put, is to communicate meaning. We, as teachers of ESL, need to focus on communication. Therefore, the seriousness of a given error is directly relate to the effect such error has on the ability to successfully communicate. (p. 299)

Esta última aproximación había sido defendida, ya en su momento, por Oller (1979) que subrayó la necesidad de dar prioridad a la efectividad de la comunicación y al significado en la producción escrita de los candidatos. Este autor también apuntó que se debía restar importancia a la corrección de las estructuras formales de los ensayos.

---

[9] NS y NNS son las formas abreviadas que se utilizan para referirse a *native speaker* (hablantes nativos) y *non-native speaker* (hablantes no nativos) respectivamente

Establecer la relación entre significado y forma se considera necesario dado que, generalmente: “all mistakes are corrected with equal vigor.” (Burt y Kiparsky 1974: 71). Este hecho puede resultar contraproducente para los estudiantes ya que se les insiste de forma reiterada sobre la importancia de ciertos errores sin hacer distinciones sobre el valor que se le atribuye a cada uno de ellos. Sweedler-Brown (1993), por ejemplo, denuncia las altas correlaciones que se dan entre los errores cometidos por el uso incorrecto de los artículos y la evaluación holística de los ensayos:

> This strong correlation between article errors and the score on sentence structure (and the correspondingly strong relation between sentence structure and holistic scores on corrected essays) is particularly disturbing because article errors are_ even by the retrospective admission of the graders in this study_ very low distortion errors which do not affect comprehension. (p. 11)

En opinión de otro grupo de autores, el criterio más importante a tener en cuenta para juzgar la seriedad o gravedad de los errores es el criterio de la *irritabilidad*. Algunos estudiosos opinan que el criterio de la *irritabilidad* se encuentra inexorablemente ligado al criterio de la *comprensión*. Así, la *irritabilidad* se juzga en la medida en que esta última contribuye a la incomprensión del texto (Piazza 1980; Ludwig 1982). Sin embargo, Vann *et al.* (1984) y Santos (1988) no comparten esta visión y diferencian ambos criterios: “…regarding irritation more as a function of the expectations and characteristics of interlocutors, who may become irritated by errors even when the message is comprehensible to them.” (Santos 1988: 70).

El tercer y último criterio que los correctores consideran decisivo en la evaluación de los errores es el de la *aceptabilidad* (Chastain 1980; Galloway 1980; Brownig 1982). Dordick (1996), sin embargo, niega que este criterio sea válido para juzgar la gravedad de los errores. Las razones que argumenta para ello son las siguientes:

> To measure acceptability, (…) respondents must give their personal opinion as to their reaction to sample errors, an inherently weak method of error gravity analysis due to its indirectness and subjectivity (Dordick 1996: 300)

A pesar de que Chastain (1980) juzga el criterio de la *comprensibilidad* como el más importante desde el punto de vista estrictamente comunicativo, el autor destaca el papel crucial que puede llegar a desempeñar el criterio de la *aceptabilidad*. Esta visión le lleva a clasificar lo errores en tres categorías distintas de acuerdo con la tolerancia lingüística, el interés y la paciencia que demuestre el interlocutor nativo ante estos últimos. Estas categorías son: 1) *comprensible y aceptable*, 2) *comprensible pero no aceptable* e 3) *incomprensible*, en caso de que se produzca una falta absoluta de comprensión. Chastain (1980) argumenta que:

> Although comprehensibility is the most important goal for non-native speakers, in interpersonal communications they must also ultimately be concerned that their language does not lead to a negative reaction on the part of the native speakers with whom they are communicating (p. 212)

De esta forma, Chastain (1980) une el criterio de la comprensión lingüística al criterio de la compatibilidad socio-cultural.

En opinión de Connor-Linton (1995), la relativa importancia atribuida a los errores no se puede medir a lo largo de una única dimensión sea esta la de *comprensibilidad / inteligibilidad*, *aceptabilidad* o *irritación*. De igual modo, ninguna dimensión puede describir y contener todos y cada uno de los posibles errores cometidos por los estudiantes.

Connor-Linton (1995) manifiesta que algunos errores resultan relevantes para unos correctores y, sin embargo, no afectan en modo alguno a otros correctores. De hecho, el desacuerdo entre los diversos correctores sobre la importancia de los errores es un hecho que se repite de forma constante. Nagy (1988), por ejemplo, señala: “...one of the main sources of disagreement among judges is the importance they place on errors” (p. 364). La interpretación y clasificación de los errores también resulta problemática (Oller 1979; Ingram 1985). Esto se debe a que: “at times the classifying of errors becomes a matter of individual interpretation and judgment.” (Santos 1988: 74). Por su parte, Ingram (1985) afirma: “It is often a matter of judgement whether, for example, an error is merely spelling (to be disregarded) or phonological or grammatical.” (Ingram 1985).

Como vemos, la variedad de opiniones que defienden los autores dificulta la sistematización y categorización de los errores así como la

determinación del valor o peso que se le ha de conceder a cada uno de ellos.

En el siguiente subepígrafe expondremos las principales técnicas utilizadas en la evaluación de los errores. El análisis de sus ventajas e inconvenientes nos guiarán en la selección de las técnicas de evaluación de los errores que utilizaremos en nuestro trabajo.

7.2.4.3.1. Técnicas en la evaluación de los errores

La mayoría de las investigaciones realizadas sobre la importancia que asignan los profesores a los errores cometidos por los estudiantes se basan en el análisis de categorías previamente establecidas. Estas categorías se crean, generalmente, a partir de los resultados obtenidos a través de la implementación de técnicas cualitativas como son, por ejemplo, las entrevistas realizadas a distintos miembros del profesorado. A través de ellas, se identifican los errores más comunes cometidos por los estudiantes de segunda lenguas (Tomiyama 1980; Van y Meyer 1984).

Otra de las técnicas más populares que se utilizan para determinar la importancia de los errores se basa en el recuento del número de errores de acuerdo con criterios específicos establecidos de antemano. Esta técnica fue la que utilizó Evans (1979) en su estudio para clasificar los distintos errores. El autor llegó a producir una lista exhaustiva de 19 categorías si bien posteriormente redujo su número (Evans, 1981). En esta nueva lista, los errores se clasifican del siguiente modo: 1) ortografía; 2) *major sentence error*, que incluye el error de puntuación y, por último, 3) errores

gramaticales, que se encuentran subcategorizados en errores de (a) pronombre*;* (b) verbo y (c) otros.

Algunos autores introducen pasajes de textos manipulados o frases descontextualizadas en las que se insertan los diversos errores que se pretenden analizar para juzgar su relevancia (Van *et. al.* 1984; Santos 1988). Estas últimas técnicas han sido duramente criticada por su falta de validez (Davies 1983; Ludwig 1982). Se cuestiona, por ejemplo, el que la presentación de los diversos errores se produzca en frases individuales y desconectadas. Este proceso se considera artificial dado que raramente se leen las frases de esta forma. También se critica el hecho de que las frases se muestren fuera de contexto ya que los errores acaban siendo juzgados de forma más severa. Estos últimos pueden incluso crecer fuera de proporción ya que el contexto consigue mitigar el efecto que cualquier error puede tener sobre la comprensión del texto (Palmer 1973; Dordick 1996). Como veremos, el estudio de estas consideraciones se intentará corroborar en la parte empírica de nuestro trabajo.

Santos (1988) nos resume brevemente las principales desventajas que presentan estos últimos métodos de investigación citados:

> Artificially prepared passages allow for maximum control of the variables by the researcher, but they also sacrifice the natural quality of unaltered connected discourse. Furthermore, they do not allow the NS judges to decide for themselves which errors are the most glaring. Finally, selectively inserted errors give equal weight to each error type by presenting them only once each, an unrealistic condition that ignores the frequent recurrence of certain error types and the relatively infrequent occurrence of others. (p. 74-75)

Conscientes de la importancia que juega el contexto en la evaluación de la producción escrita, en nuestro trabajo decidimos combinar dos tipos distintos de metodología. Por una parte, nos propusimos estudiar los distintos errores contenidos en frases descontextualizadas aunque inalteradas, dado que nos basamos en las muestras de texto escrito originales (i.e. producidas por los estudiantes en el desarrollo de sus ensayos en las pruebas de Selectividad). Por otra parte, decidimos investigar el efecto de los errores en el discurso conectado, es decir, en los ensayos originales de los estudiantes. La combinación de ambas metodologías nos permitirá comprobar si los resultados de los correctores en el análisis de las frases descontextualizadas y en el discurso conectado se mantienen a pesar de las distintas técnicas de investigación utilizadas.

7.2.4.4. Forma versus contenido

Fathman y Whalley (1990) afirman que hay una gran controversia entre los profesores e investigadores por lo que respecta a la adecuación del *feedback* que se ha de procurar a los estudiantes en la evaluación de los ensayos. Esto afecta especialmente a los aspectos relacionados con la *forma* vs. el *contenido*. Los autores aseguran:

> Much of the conflict over teacher response to written work has been whether teacher feedback should focus on form, (e.g., grammar, mechanics) or on content (e.g., organization, amount of detail). (Fathman y Whalley 1990: 178)

El debate suscitado en torno a la evaluación de estos dos últimos aspectos no es nuevo. Recordemos que, a partir de 1945, dentro de la enseñanza de segundas lenguas se aprecia un mayor interés por los aspectos comunicativos. Este nuevo planteamiento se tradujo en enfoques basados principalmente en el significado (i.e. *meaningful-based approaches*) y en el desarrollo de tareas (i.e. *task-based approaches*). Dentro de la producción escrita, se observa una tendencia general entre los profesionales a estudiar y subrayar el proceso de composición que siguen los estudiantes en la elaboración de los ensayos frente a la exclusiva consideración del producto final. Desde esta perspectiva, el texto se contempla como un elemento secundario cuya forma se deriva del *contenido* y de la necesidad de comunicación (Miller y Judy 1978).

Sin embargo, y a pesar de que el enfoque basado en el proceso ha sido muy bien acogido dentro del mundo de la enseñanza, algunas investigaciones recientes demuestran que el *feedback* que ofrece el profesorado a sus estudiantes se centra exclusivamente en los elementos del ensayo relacionados con la *forma* (i.e. gramática, ortografía). Por el contrario, en la mayoría de los casos, se presta escasa atención al *contenido*. (Applebee 1981; Perkins 1981; Homburg 1984; Ziv N. 1984; Zamel 1985; Amengual y Herrera 2000).

Como vimos en el capítulo 4 (epígrafe 4.4.), dentro del campo de la investigación de la producción escrita va a surgir un renovado interés por el producto final elaborado por los estudiantes. Dichos estudiantes han de operar necesariamente dentro de un contexto académico que se rige por

unas normas estrictas y que les exige unas tareas de producción escrita muy estructuradas (Hortowitz 1986a). Estas circunstancias son las que animan a algunos autores a buscar un compromiso entre el proceso y el producto. Estos aspectos se hallan estrechamente vinculados a los de *contenido* y *forma* respectivamente (Connor 1987; Hamp-Lyons 1990). Sweedler-Brown (1993) expone esta idea:

> Certainly, correctness should not be the exclusive or even predominant concern in ESL writing. We must continue to teach writing as a meaningful communicative activity, but accuracy of language has been given short shrift in the wake of the process phenomenon, and ample research suggests that particularly in postsecondary academic settings we may be doing our students a disservice if we are not willing to become language teachers as well as writing teachers to our ESL students. (p. 15)

Mientras la controversia en torno a los aspectos del ensayo que el profesorado debe primar continua, Mohan y Low (1995) destacan la tensión que dicha situación produce en el profesorado de segundas lenguas: “ ...as teachers question the *what* and *how* of assessing student writing in an academic context.” (Mohan y Low 1995: 28).

Dentro del estudio concreto de las evaluaciones holísticas, la excesiva atención que se concede a los aspectos formales se denuncia reiteradamente. Así, Sweedler-Brown (1993) asegura que el peso que se le concede a los errores gramaticales es mucho mayor que el que se le otorga a los aspectos retóricos del ensayo:

> In this study, sentence-level error was the only significant influence on holistic score and was, furthermore, the critical factor in pass/fail decisions in these ESL essays. It is discouraging to note that the quality of organization and paragraph development had no observable effect on the essays' holistic scores. (p. 12)

No obstante, algunos investigadores, contradicen dicho supuesto y subrayan la relevancia de los aspectos temáticos o retóricos (i.e. el contenido, la organización o el desarrollo de las ideas) y su influencia en las puntuaciones holísticas de los correctores (Freedman 1979; Freedman y Pringle 1980; Santos 1988). Santos (1988), por ejemplo, afirma en su estudio que los correctores, sin formación previa en técnicas de evaluación, juzgan el aspecto de contenido de forma más severa que el aspecto gramatical o de habilidad lingüística de los ensayos:

> The findings for several of the research questions seem to lead to the conclusion that professors are willing to look beyond the deficiencies of language to the content in the writing of these NNS. This conclusion would account for the fact that the content of the essays was rated significantly lower than the language. (p. 84)

Song y Caruso (1996) comparten esta última opinión y sostienen que:

> That English raters assigned a higher holistic mean score to the ESL essay, which is weaker in language use but stronger in content and rhetorical features in comparison with the NES essay, clearly indicates that content and rhetorical features were the major factors influencing English rater' judgement of writing, whether by ESL or NES students. (p. 173)

Sin embargo, Sweedler-Brown (1993) señala que ciertas cualidades de los ensayos que se subrayan en la enseñanza de la producción escrita entendida como proceso, tan sólo se tienen en cuenta cuando los estudiantes demuestran un alto grado de fluidez lingüística. O sea, se trata de estudiantes que únicamente necesitan que se les recuerde las normas de la escritura formal en lengua inglesa durante la práctica de los ejercicios de composición. Este hecho explica la efectividad del enfoque basado en el proceso cuando se utiliza con estudiantes nativos. Sweedler-Brown (1993) afirma:

> This assessment orientation has worked well with NS student essays, where it is rare that a high level of rhetorical ability is not accompanied by acceptable sentence-level control. (p. 4)

Sweedler-Brown (1993) continúa diciendo que cuando los estudiantes no poseen un dominio de la lengua inglesa lo suficientemente alto o bueno, los correctores hacen caso omiso de los criterios evaluativos. La corrección gramatical pasa a ser considerada, entonces, un elemento decisivo en la evaluación de los ensayos:

> ...it appeared to be the case that the rhetorical quality of an essay had little or no effect on the score assigned by these graders if a noticeable level of ESL error was evident. (Sweedler-Brown 1993: 5)

En esta misma línea de investigación, Amengual y Herrera (2000) argumentan que en los ensayos de los estudiantes con un dominio lingüístico inferior, los correctores enfatizan principalmente el aspecto formal de la lengua en sus evaluaciones. Tan sólo en los ensayos con un dominio de la lengua inglesa alto es cuando los correctores prestan atención a ambos aspectos, es decir, a la *forma* y al *contenido*. De ser así, a los estudiantes con un dominio lingüístico inferior se les está negando una información que pudiera serles útil ya que son, precisamente, estos estudiantes los que se muestran más ansiosos por recibir algún tipo de *feedback* positivo sobre algún aspecto del ensayo que les anime a producir trabajos mejores (Cohen y Cavalcanti, 1990).

Es importante destacar que autores como Cummins y Swain (1986) y Cumming (1990) afirman que la habilidad de producción escrita y la habilidad lingüística no son tan interdependientes en la producción escrita de los estudiantes de segundas lenguas como en la de los estudiantes nativos. Cumming (1990) explica que, a pesar de que los correctores conciben la competencia lingüística en una segunda lengua y la habilidad de producción escrita en dicha lengua como “separate, non-interacting factors”, en muchas ocasiones:

> Both skills are being evaluated in conjunction and are not logically distinguished from each other within evaluators’ ratings. This may disadvantage two groups of students in different ways in evaluation practices. For minority language students, such as ESL learners in English-dominant settings, analytic evaluations of their written compositions may be biased against their language proficiency. Conversely, for unskilled writers in second language programmes, the same kinds of evaluation may be biased against their lack of expertise in composition. (p. 42)

En este sentido, los enfoques analíticos se presentan como una alternativa válida que trata de subsanar las deficiencias de los enfoques holísticos vistas hasta ahora. Los métodos analíticos aseguran la contemplación tanto de los aspectos de la *forma* como del *contenido*, dado que se les asigna un valor individual e independiente a cada uno de ellos en las evaluación de los ensayos.

La escala analítica de Diederich *et al.* (1961), por ejemplo, otorga 10 puntos a los dos primeros elementos denominados: *ideas* y *forms* y 5 puntos a los elementos restantes: *flavour*, *mechanics* y *wording*. Por su lado, Jacobs *et. al.* (1981) en su *ESL Composition Profile* proponen la distribución del valor de los elementos del siguiente modo: contenido*,* 30; organización*,* 20; vocabulario, 20; uso de la lengua*,* 25; y mecánica*,* 5. Según esta distribución, los aspectos relacionados con la *forma* y con el *contenido* representan cada uno de ellos el 50% del valor total que se le asigna a los ensayos. Song y Caruso (1996) consideran que la mayoría de los correctores, tanto nativos como no nativos, son partidarios de asignar el 60% del valor total a los aspectos relacionados con el contenido y la organización de los ensayos frente al 40% del valor total que se le otorgaría al uso de la lengua.

En un intento por reconciliar ambos aspectos, Kroll (1990d) nos propone la construcción de una escala, “a rhetoric/syntax split scale”, en la que los aspectos de *forma* y *contenido* reciben el mismo valor. Esto es, el

50% del valor total que se le asigna a los ensayos. Sweedler-Brown (1993) defiende la utilización de la escala propuesta por Kroll (1990d) dado que dicha escala garantiza el reconocimiento de los aspectos retóricos de los ensayos: “A divided scale would ensure that the students’ rhetorical abilities were acknowledged even in the presence of frequent ESL error.” (Sweedler-Brown 1993: 14).

Sweedler-Brown (1993) continúa diciendo que la escala de Kroll (1990d) cuenta con la ventaja adicional de facilitar información a los estudiantes sobre los aspectos retóricos que utilizaron correctamente, aún cuando los estudiantes no lleguen al nivel de competencia lingüístico mínimo requerido. Esta información se les niega en las evaluación holísticas:

> Furthermore, even if a student’s language skills were not at a required level of competency, his or her rhetorical skills would be clearly and separately acknowledged. (p. 14)

A pesar de los resultados obtenidos en los estudios citados, se requieren investigaciones futuras que nos ayuden a establecer el conveniente equilibrio entre los aspectos relacionados con la *forma* y con el *contenido* de los ensayos. En cualquier caso, y mientras continúa el debate en torno a los elementos que se deben valorar en la producción escrita de los candidatos, parece ser que la mayoría de los autores se muestran partidarios de evaluar ambos aspectos, es decir, *forma* y *contenido* de forma conjunta (Taylor 1981; Raimes 1983; Krashen 1984; Fathman y Whalley 1990; Kroll 1990d).

Fathman y Whalley (1990) sostienen que la formación más adecuada que se les puede procurar a los estudiantes para que logren mejorar los aspectos tanto de la *forma* como del *contenido* de sus ensayos se consigue con la técnica del *rewriting*. Esta técnica, como su nombre indica, consiste en la producción de diferentes borradores antes de la entrega final del texto definitivo. La implementación de dicha técnica debería ir acompañada de *feedback* simultáneo sobre aspectos de la *forma* y del *contenido*:

> The results of this study also suggest that when grammar and content feedback are presented at the same time, the content of rewrites improves approximately as much as when content feedback only is given. Focus on grammar does not necessarily affect the content of the writing. This would suggest that students can improve their writing in situations where content and form feedback are given simultaneously. (Fathman y Whalley 1990: 186)

Mohan y Low (1995) defienden de forma decidida la evaluación conjunta o integral de los aspectos de *forma* y *contenido* en las producciones escritas. Para conseguir este objetivo, los autores sugieren formar a los correctores a través de la técnica denominada *collaborative assessment*. Dicha técnica requiere que los profesores de un mismo curso colaboren en la definición conjunta de los criterios de evaluación que se aplicarán a un grupo de pruebas. Según nos explican Mohan y Low (1995), la técnica de *collaborative assessment* que se implementó en su estudio puso en evidencia la dificultad de establecer la relación que se da entre los aspectos de lengua (i.e. *forma*) y de *contenido*. Estos autores afirman que los profesores carecen de un criterio común para llegar a definir ambos

conceptos. A la mayoría de los correctores les resulta difícil extraer el contenido del ensayo a partir de una forma textual deficiente. La pregunta que se formulan estos autores es la siguiente: “Can the assessor interpret what students mean from what they say?” (Mohan y Low 1995: 3).

Song y Caruso (1996) responden afirmativamente a dicha pregunta. Estos últimos autores demuestran que los correctores son capaces de distinguir perfectamente entre aquellos aspectos que se relacionan con la *forma* y aquellos que se relacionan con el *contenido* del ensayo. Song y Caruso (1996) concluyen: “The English faculty were able to make a distinction between the quality of the rhetorical and language aspects of the ESL writing.” (p. 172).

Santos (1988) considera también positiva la distinción que se establece entre los aspectos de *forma* y *contenido* en la evaluación de los ensayos. No obstante, cuestiona que dicha actuación sea siempre posible:

> …It is a mark of their tolerance that although they regarded errors as linguistically unacceptable, the professors still judged content and language independently, to the extent that this was possible. (Santos 1988: 84)

Así, Santos (1988) nos explica que los errores léxicos al afectar directamente sobre el contenido de los ensayos impiden que los aspectos de *forma* y *contenido* se juzguen de modo independiente:

> It is precisely with this type of error (the lexical error) that language impinges directly on content; when the wrong word is used, the meaning is very likely to be obscured. (p. 74)

La relación entre *forma* y *contenido* cobra una especial relevancia dentro del contexto de la evaluación de la producción escrita en segundas lenguas dado que nuestros estudiantes escriben y se expresan en una lengua que no es la suya propia. Desde una perspectiva pedagógica, se señala la necesidad de ofrecer a los estudiantes información sobre el valor o peso que los correctores atribuyen tanto a la calidad de ideas como a la calidad de la expresión escrita. Taylor *et al.* (1988) declaran que los profesores deberían ayudar a los estudiantes a desarrollar lo que se denomina *academic literacy*. De esta forma, los estudiantes podrían lograr el éxito académico deseado.

Las investigaciones de Norton y Starfield (1997) apuntan a que, en general, la gran mayoría de los estudiantes desea que el *feedback* que les facilita el profesorado contemple los aspectos de *forma* y de *contenido* de forma conjunta. Tan sólo una escasa minoría prefiere que el profesorado responda únicamente a los aspectos relacionados con el *contenido*, aún cuando este último aspecto forma parte de una disciplina académica distinta a la de lengua inglesa.

Los criterios de evaluación que facilitan los coordinadores de las pruebas de Inglés de Selectividad suelen indicar algunos aspectos de la *forma* y del *contenido* que se deberían valorar en la evaluación de los ensayos. A título ilustrativo, podemos citar algunos ejemplos: corrección

ortográfica, corrección de la estructura sintáctica (i.e. *forma*) y organización de ideas y creatividad (i.e. *contenido*).

Asimismo, se suele señalar, de modo indicativo, la distribución de las puntuaciones para determinados aspectos del ensayo. Hay que hacer notar que no siempre se asignan valores individuales para cada aspecto sino que se indica una valoración general que puede incluir dos o tres aspectos al mismo tiempo. A modo de ejemplo, citamos la distribución de algunas puntuaciones sometidas para evaluar los ensayos de la prueba de Inglés en la convocatoria de junio de 2000 en la *Universitat de les Illes Balears* (UIB): "1,5 puntos por la corrección de la estructura sintáctica y organización de ideas"; "1,5 puntos por la adecuada utilización del léxico, su riqueza y creatividad."

Como vemos, los criterios de evaluación que se establecen son muy abiertos y flexibles. Este hecho favorece que los correctores se dejen guiar por sus propias orientaciones metodológicas en cuanto a la importancia que se ha de prestar a los aspectos de la *forma* y del *contenido* de los ensayos.

7.2.5. La formación de los correctores en técnicas de evaluación

En el campo de LT existe un especial interés por corroborar las valoraciones o juicios que emiten cada uno de los responsables de los distintos niveles y etapas de la elaboración y el diseño de las pruebas. Lo que se pretende con ello es garantizar la obtención de resultados fiables:

> The most important task of statistics in behavioural research_ and, of course, in research on language behaviour_ is to protect the drawing of conclusions against possible biases resulting from sampling and measurement error. (Schils *et al.*1991: 125)

No obstante, la fiabilidad inter-corrector raramente se mantiene. Los correctores tienden a variar en la elaboración de sus juicios dependiendo de múltiples factores que hemos venido analizando hasta ahora. Entre estos factores cabe destacar: la benevolencia o exigencia de los correctores (Cason y Cason 1984; Linacre 1989; Wigglesworth 1993) y la aparición de elementos de error o elementos fortuitos en la evaluación.

Lumley y McNamara (1996) afirman que las diferencias observadas en las puntuaciones de los correctores obedecen a diversas causas. Así, por ejemplo, un corrector puede mostrarse más *indulgente* (*lenient*) que otro, de forma general. También puede que se apliquen diferentes niveles de exigencia o de condescendencia a aspectos concretos de la evaluación (i.e. un determinado grupo de candidatos, una tarea específica o diversos criterios de evaluación establecidos). Una tercera posibilidad que apuntan Lumley y McNamara (1996) es que los correctores interpreten la escala y los criterios de evaluación de forma distinta. Por último, Lumley y McNamara (1996) reconocen la posibilidad de que los correctores no sean consistentes consigo mismo y modifiquen su nivel de exigencia en las distintas ocasiones.

Esta gran variedad de aspectos se puede englobar en una única variable denominada “características de los correctores” (*rater characteristics*) McNamara y Adams (1991/1994). Dicha variable contempla tanto los aspectos relacionados con la severidad de los correctores como la de otros aspectos más específicos.

Según Cason y Cason (1984) las diferencias relacionadas con las distintas muestras de severidad de los correctores ejerce la misma influencia en las puntuaciones de los ensayos que las diferencias que se atribuyen a las diversas habilidades de producción escrita de los candidatos. Asimismo, Wigglesworth (1993) señala que, desafortunadamente, el factor suerte ejerce un papel decisivo en la evaluación de los candidatos:

> From the point of view of the candidate, it becomes a matter of luck whether they are assessed by particular raters. A candidate may draw the most lenient member of the rating team and benefit as a result or, alternatively, s/he may draw the harshest member and may suffer the consequences of this. (p. 305)

Entre las varias medidas que se proponen para tratar de resolver o paliar estas situaciones, cabe destacar la formación de los correctores (*teacher training*) en técnicas de evaluación. Este proceso se desarrolla, generalmente, en varias sesiones siguiendo una serie de etapas como las que se detallan a continuación:

> 1. Read and discuss scales together.
> 2. Review language samples which have been previously rated by expert raters and discuss the ratings given.
> 3. Practice rating a different set of language samples. Then compare the ratings with those of experienced raters. Discuss the ratings and how the criteria were applied.
> 4. Rate additional language samples and discuss.
> 5. Each trainee rates the same set of samples. Check for the amount of time taken to rate and for consistency.
> 6. Select raters who are able to provide reliable and efficient ratings.
> (Bachman y Palmer 1996: 222).

La eficacia real de la técnica de formación de los correctores se ha cuestionado en numerosas ocasiones. Dentro del ámbito de la evaluación de la producción escrita se critica, por ejemplo, su naturaleza artificial (Charney 1984; Barrit Stock y Clarke 1986; Huot 1990). Asimismo, autores como Charney (1984), Constable y Andrich (1984) o Huot (1990) consideran que se otorga un excesivo énfasis a la consecución de resultados fiables, lo que en determinadas ocasiones va en detrimento de la validez de las pruebas. Estos autores también destacan la gran distorsión que se produce en la interacción entre el lector y el texto escrito en el proceso de lectura cuando el objetivo primordial que se persigue se reduce a la simple búsqueda del consenso entre los correctores. Huot (1990) explica que los correctores reaccionan de forma individual y personal ante el texto escrito y que dicha actuación es intrínseca al propio proceso de lectura. El autor confiesa su preocupación en la siguiente cita:

> a personal stake in reading might be reduced to a set of negotiated principles, and then a true rating of writing quality could be sacrificed for a reliable one. (Huot 1990: 211)

En opinión de Henning (1996), el proceso de formación de los correctores presupone la siguiente lectura:

> Implicit in this approach is the assumption that rater agreement is a reflection of reliability of evaluation, and rater disagreement is an indication of unreliability of evaluation. (p. 53)

Henning (1996) nos advierte del riesgo que supone identificar estos dos tipos de constructo. Esto es: acuerdo entre los correctores (i.e. *rater agreement*) y fiabilidad de puntuaciones (i.e. *score reliability*). A pesar de que ambos constructos comparten características similares no son idénticos. Henning (1996) también puntualiza que el consenso entre los correctores no puede entenderse como sinónimo de aproximación a la habilidad de producción escrita real de los candidatos.

Por su parte, Charney (1984) argumenta que la obtención de puntuaciones homogéneas sólo es posible si se evalúan los aspectos más superficiales del texto. Coffman (1971) y Cooper (1977), señalan, además, que las puntuaciones fiables que se consiguen durante las sesiones de formación se explican por la presión que ejercen los compañeros del grupo para que se evalúe de forma similar.

A pesar de estas críticas, la formación de los correctores en el proceso de evaluación se ha considerado, tradicionalmente, una técnica decisiva y eficaz capaz de garantizar niveles óptimos de fiabilidad en las puntuaciones de los correctores (Cooper 1977; Jacobs *et al.* 1981; Homburg 1984; Bachman y Palmer 1996).

Entre algunos de los principales efectos positivos que se mencionan (ver Weigle, 1994) cabe destacar que la formación de los correctores junto con el uso de escalas de valoración clarifica, presumiblemente, los criterios de evaluación que se han de aplicar (Charney, 1984). Se considera, además, que la formación de los correctores consigue minimizar las posibles

diferencias relacionadas con los aspectos educacionales o contextuales de los correctores (Jacobs *et al*., 1981).

Asimismo, se favorece el que los correctores se centren en los criterios evaluativos establecidos (Follman y Alderson, 1967) y se logran modificar las expectativas concernientes a lo que se entiende por una producción escrita de calidad. Esto último se consigue gracias a la clarificación de las exigencias que presentan las diversas tareas y al análisis de las características concretas de los candidatos (Freedman y Calfe 1983; Huot 1990). McIntyre (1993) concluye que la formación de los correctores reduce las diferencias extremas entre los correctores. Además, los correctores extremos (*outliers*) son fácilmente identificables tanto por su desmesurada exigencia como por su excesiva benevolencia en la aplicación de los criterios de corrección. Esto nos va a permitir prescindir de su participación en evaluaciones futuras.

Sin embargo, los numerosos beneficios que se citan no logran suprimir la totalidad de las diferencias que hay entre los correctores. La mayoría de las diferencias persisten aún después de haberse finalizado la etapa de formación, por lo que el éxito de esta técnica se considera únicamente parcial (Stahl y Lunz 1991; Vaughan 1991; Lumley y McNamara 1995). Weigle (1998) observa este fenómeno en su estudio:

> …the process of training was effective in reducing some of the differences in rater severity, but major differences among raters remain despite training. (p. 277)

puesta en práctica de esta nueva tecnología de facetas múltiples para la investigación en el campo de LT:

> By producing rater calibrations that are independent of the data used to derive them, comparison across different rating occasions becomes possible. Multifaceted measurement has made possible the close examination of an issue that has long been recognized. (Lumley y McNamara 1995: 69)

La formación de los correctores para los proponentes del modelo Rasch de facetas múltiples se considera importante para lograr el común entendimiento entre los correctores sobre la aplicación de la escala de valoración a los ensayos. No obstante, a diferencia de otros enfoques, este modelo no pretende que los correctores lleguen a conseguir niveles de severidad similares, tarea que, por otra parte, ha resultado ser imposible hasta ahora. El objetivo que se persigue es el de la fiabilidad intra-corrector:

> The implication is that the function of training is not, or should not necessarily be, to force raters into agreement with each other (inter-rater reliability), but rather to train raters to be self-consistent (intra-rater reliability). (Weigle 1998: 264-265)

El estudio de Weigle (1998) confirma los resultados obtenidos por Lunz, Wright y Linacre (1990). Estos autores afirman que el proceso de formación de los correctores consigue aumentar, principalmente, la consistencia o fiabilidad de los correctores consigo mismo (i.e. fiabilidad intra-corrector). La consistencia entre los diversos correctores (i.e. fiabilidad

Hamp-Lyons (1990) considera que la influencia de esta técnica en la evaluación de la producción escrita no ha sido suficientemente investigada. Esta autora subraya la necesidad de realizar estudios tanto previos como posteriores sobre el proceso de formación de los correctores (pre- and post-training) que permitan contrastar sus efectos y consecuencias. Hamp-Lyons (1990) sostiene que la multiplicidad de factores que intervienen en el proceso de formación de los correctores dificulta esta tarea:

> The context in which training occurs, the type of training given, the extent to which training is monitored, the extent to which reading is monitored, and the feedback given to readers all play an important part in maintaining both the reliability and the validity of the scoring of the essays. (p. 82)

Es importante señalar que la literatura cuantitativa (*measurement literature*) aborda la formación de los correctores desde una perspectiva algo diferente a la tradicional. Así, el modelo Rasch de facetas múltiples desarrollado por Linacre (1989) no establece como objetivo último la eliminación de las diferencias entre los correctores. Este modelo contempla dicho fenómeno como algo inevitable y hasta cierto punto beneficioso ya que ofrece la variabilidad suficiente para poder establecer la estimación probabilística de tres aspectos claves en la misma escala linear. Esto es: la severidad de los correctores (*rater severity*), la dificultad de la tarea (*task difficulty*) y la habilidad del candidato (*examinee ability*).

Lumley y McNamara (1995), al igual que autores como Elder (1993) o Brown (1995), ensalzan, en este sentido, el enorme potencial que supone la

inter-corrector) que, en opinión de Lunz, Wright y Linacre (1990), pretende convertir a los correctores en duplicados los unos de los otros se ve sensiblemente menos afectada. Asimismo, Lumley y McNamara (1995) admiten que el logro mayor que se consigue con el proceso de formación de los correctores es el siguiente: “The main contribution of rater training is to reduce the random error in rater judgements.” (p. 57).

Wigglesworth (1993) aplica la técnica del modelo Rasch de facetas múltiples a la investigación y análisis de posibles sesgos (*bias analysis*) entre los correctores durante su actuación en una prueba de producción oral. Gracias a la puesta en práctica de esta técnica, la autora descubre los efectos beneficiosos que supone el obtener información particularizada sobre el comportamiento de cada uno de los correctores. Se consigue, de este modo, trazar el criterio de actuación individual de los distintos correctores.

Wigglesworth (1993) señala que los correctores responden positivamente al *feedback* que se obtiene sobre sus diversas actuaciones en el proceso de evaluación. Estos correctores son a la vez capaces de incorporar dicha información en sus evaluaciones posteriores por lo que se favorece la reducción de posibles sesgos que pudiera haber. La obtención de estos resultados que, bien pudieran confirmarse en la evaluación de la producción escrita, es extremadamente positiva ya que, tal y como afirma Wigglesworth (1993):

> This type of analysis enables trainers to build up a profile of rater characteristics, and to obtain continuing feedback on their individual performances. The decision to include bias analysis as a feature in the training of raters on the oral interaction subskill of ‘access” is evidence of the importance of incorporating bias analysis into rater training. (p. 318- 319)

Una de las preocupaciones constantes relativas al proceso de formación de los correctores es descifrar la duración o mantenimiento de los efectos de dicho proceso a lo largo del tiempo. Lumley y McNamara (1995) se formulan esta pregunta:

> If a rater’s characteristics are successfully modified by training, are these changes stable over time, or does the rater revert to old habits? How often do raters need to be retrained? (p. 59)

Los resultados de varios estudios confirman que los cambios producidos en el comportamiento de los correctores como resultado de su formación en los proceso de evaluación no se mantienen estables a lo largo de extensos periodos de tiempo. Sobre todo, si el proceso de formación conlleva la corrección de un gran número de ensayos (Coffman y Kurfman 1968; Lunz y Stahl 1990). Así, Lumley y Mcnamara (1995) descubren grandes discrepancias en el comportamiento de los correctores a lo largo de dos periodos de tiempo: uno de ellos asociado al período correspondiente a la formación de los correctores y, el otro, mayormente vinculado al período en el que transcurre la administración y producción de la prueba. Estos datos aconsejan el desarrollo de sesiones previas de formación antes del inicio de

cada una de las pruebas. De este modo, se consigue que los correctores asimilen e interioricen los criterios de corrección aprendidos:

> It seems that at every administration, new calibrations of rater characteristics are required; failing that, the traditional technique of double and if necessary multiple ratings seems amply justified. (Lumley y Mcnamara 1995: 69)

A pesar de los inconvenientes que plantea el proceso de formación de los correctores, pensamos que la implementación de esta técnica pudiera resultar beneficiosa en el desarrollo de las PAAU. Como hemos visto, la formación de los correctores permite que la aplicación de los criterios de evaluación se realice de forma uniforme y homogénea y, por consiguiente, se logran reducir los factores subjetivos que influyen en las decisiones finales de los correctores. Weigle (1994) nos advierte, en este sentido, que resulta difícil obtener medidas fiables y válidas por parte de los correctores que no han sido previamente formados en las técnicas de evaluación.

No obstante, y reconociendo que aún utilizando tecnologías innovadoras como el modelo Rasch de facetas múltiples, existen múltiples factores fortuitos y no controlables que impiden la obtención de resultados plenamente fiables, creemos conveniente recordar en este punto las palabras de Lumley y Mcnamara (1995) cuando concluyen que, en definitiva: “no one judgement may be said to be definitive” (p. 57).

7.3. Variables relativas a la tarea

En este epígrafe introduciremos, en primer lugar, los resultados de algunos estudios que explican los efectos producidos por el método en las puntuaciones holísticas de las pruebas de producción escrita. A continuación, analizaremos la incidencia que ejerce la elección del tema en la elaboración de los ensayos. Finalmente, estudiaremos la influencia del factor tiempo en la calidad del producto final de los ensayos. Estas dos variables son especialmente relevantes en el diseño y desarrollo del ensayo en las PAAU.

7.3.1. Introducción

En realidad, no existe ningún método ni teoría que explique hasta qué punto los componentes del método (i.e *method effects*) influyen en los resultados de las pruebas.

De hecho, en los modelos psicométricos que se aplican a la evaluación objetiva se distingue entre los procesos de evaluación, por una parte, y el proceso de corrección, por otra. De esta forma se asume la ausencia de los posibles efectos producidos por el método.

Sin embargo, cuando se asume la presencia de estos efectos como ocurre en la evaluación *subjetiva* de la producción escrita, la puntuación global obtenida no se considera dependiente únicamente de las habilidades de los candidatos sino que se atribuye, parcialmente, a un grupo de características relacionadas con la tarea (Upshur y Turner, 1999).

7.3.2. Efectos producidos por el método

Los efectos que produce la elección del método en las puntuaciones globales que asignan los correctores han sido extensamente reconocidos. En general, la gran mayoría de autores aboga por el necesario control de las diversas fuentes potenciales de error (Alderson y Urquhart 1985; Bachman *et al.* 1995; Lumley y McNamara 1995; Herrera 2002).

Según nos comentan Bachman y Palmer (1996):

> The characteristics of the tasks used are always likely to affect test scores to some degree, so that there is virtually no test that yields only information about the ability we want to measure. (p. 46)

Bachman y Palmer (1996) afirman que las tareas que aparecen en las distintas pruebas son, normalmente, de naturaleza muy variada, lo que probablemente condicione los resultados que se obtengan. Asimismo, una misma tarea como el ensayo puede mostrar características diversas en cuanto a: los distintos tipos de tema a desarrollar, la audiencia a la que se dirige el texto, el objetivo que se persigue o los diferentes aspectos discursivos que exige el desarrollo del tema.

Por este motivo Bachman y Palmer (1996) rechazan la idea de caracterizar las tareas de un modo holístico o global. En su lugar, estos autores presentan un marco descriptivo (i.e. *a framework of language task characteristics*) que contempla las diversas características de las tareas a tener en cuenta. En este marco se incluye los siguientes aspectos: *setting*,

the test rubric, *the input*, *the expected response* y *the relationship between input and response*.

La aplicación de esta estructura nos ofrece dos ventajas fundamentales:

1) la posibilidad de describir las tareas. Esto nos permite comparar las características de las situaciones reales de uso de la lengua con las de las tareas que se incluyen en las pruebas y
2) la creación de nuevas tareas para futuras pruebas utilizando el marco de referencia creado para este propósito.

Horowitz (1991) expone que uno de los aspectos principales a tener en cuenta en la evaluación de la producción escrita es la consideración del contexto en el que se realiza la tarea. Así, las tareas deben reflejar la realidad de las instituciones en las que se trabaja. Con este objetivo, Horowitz (1991) elabora una taxonomía de *essay examination prompts* cuya aplicación, si bien resulta más apropiada dentro del marco de la producción escrita del Inglés para Fines Específicos (IFE), también se considera útil y productiva en otras situaciones.

Horowitz (1991) incluye cuatro categorías básicas en su taxonomía de tareas de producción escrita: 1) mostrar familiaridad con un concepto 2) mostrar familiaridad con la relación entre conceptos 3) mostrar familiaridad con un proceso y 4) mostrar familiaridad con la argumentación. Esta última categoría que incluye la noción de pensamiento crítico es, probablemente, la

que mejor ilustra el tipo de producción escrita que deben acometer nuestros estudiantes en el desarrollo de los ensayos de Selectividad.

Cabe hacer notar que las diferentes escalas o criterios de corrección que se utilizan en la evaluación de la producción escrita no siempre reflejan en sus consideraciones finales los efectos potenciales producidos por las diversas tareas realizadas en las pruebas. Upshur y Turner (1999) nos sintetizan las tres nociones básicas que se asumen en la elaboración de estas escalas.

Una primera noción denominada evaluación del dominio de la lengua absoluto (*absolute proficiency rating*) obvia los efectos potenciales producidos por la tarea o, bien asume que la escala de valoración puede compensar cualquier fuente posible de error. Una segunda noción, más popularmente conocida como evaluación del dominio de la lengua en las diferentes tareas (*task proficiency rating*), juzga y analiza el grado de dificultad de las tareas. A pesar de que se utiliza una escala única que se aplica a todas las tareas que informan la misma habilidad (por ejemplo, el conjunto de tareas de producción escrita), se intenta ajustar la dificultad de las diferentes tareas a la habilidad que demuestran los candidatos. Por último, la escala denominada evaluación según una unidad o tarea individual (*rating according to a task / scale unit*) reconoce los efectos producidos por las diversas tareas. Por consiguiente, las escalas de valoración se desarrollan para un grupo de tareas determinado de forma específica.

No obstante, por ahora, carecemos de estudios que hayan investigado los efectos que estas tres escalas de valoración producen en la evaluación de las diferentes habilidades.

7.3.3. La elección del tema

De entre todas la variables que se han identificado como fuentes posibles de error en la evaluación de la producción escrita de los candidatos (i.e. el factor tiempo, el propósito que persigue el texto, la audiencia a la que se dirige el texto, etc.), la variable que ha suscitado una mayor polémica ha sido, sin duda, la de la elección del tema del ensayo. Esto se explica, en parte, por la dificultad que supone elegir temas que sean representativos para los candidatos y que les permitan demostrar su habilidad de producción escrita de forma eficiente.

Las investigaciones llevadas a cabo en el campo de las primeras lenguas han demostrado que el contenido y la extensión de los textos se ven notablemente afectados por la elección del tema que se desarrolla en el ensayo (Hartog *et al.* 1941; Applebee 1983; Pollitt *et al.* 1985).

Por el contrario, otros estudios contradicen dichos resultados y demuestran que el tema sobre el que se basa la producción escrita de los estudiantes ejerce una escasa influencia en las puntuaciones finales. Así, Carlson *et al.* (1985) en un estudio conducente al *Test of Written English* (TWE) afirman no encontrar diferencias significativas en el modo en que la elección del tema clasifica a los candidatos. No obstante, Greenberg (1986) advierte que los resultados descubiertos en el estudio de Carlson *et al*

(1985) se deben a la naturaleza de los criterios de corrección que se aplican. Según nos explica este autor: “the lack of variation in scores across the topic types may be due to the scorers or the scoring procedures rather than to the performance of the examinees.” (Greenberg 1986).

Hamp-Lyons (1990), se suma a la opinión de Greenberg (1986) y argumenta que, como de costumbre, la conclusión a la que uno llega depende de la elección de los instrumentos estadísticos que se elijan para el análisis de los datos y de las expectativas con las que uno inicie la investigación. Asimismo, esta autora subraya la importancia que tiene la elección del tema para los proponentes del Inglés para Fines Específicos. Según estos últimos, los estudiantes de segundas lenguas se ven sumamente favorecidos, tanto en el plano lingüístico como en el plano académico, por los programas de enseñanza y evaluación de lenguas que se centran en el desarrollo de los temas propios de su disciplina. No obstante, carecemos por ahora de estudios empíricos que avalen este argumento.

El dilema que se plantean la mayoría de los estudios anteriores responde a la siguiente pregunta: ¿se debe ofrecer a los candidatos la posibilidad de elegir el tema a desarrollar en el ensayo?. Hasta el momento, los autores no han llegado a un consenso sobre la relevancia e influencia del desarrollo de diferentes temas en las puntuaciones globales de los correctores. Por otra parte, si se opta por ofrecer la posibilidad a los candidatos de elegir el tema, como ocurre en el ejercicio del ensayo de la prueba de Inglés de Selectividad, se asume, entonces, que los temas que se

ofertan son equivalentes. Papajohn (1999) nos advierte de los inconvenientes que plantea dicha premisa:

> Topics necessarily vary from one field to another, yet different topics are also assigned within a given field. Topics within a single field may differ in such areas as topic length, degree of contextualization, distribution of new information and type of information. Various topics in a given field are assumed to be equivalent for the purpose of test evaluation, yet no empirical data has been collected to show that this is true. (p. 52-53)

El argumento de Papajohn (1999) se hace extensivo al desarrollo de temas específicos dentro de un mismo campo de conocimiento ya que existen algunos aspectos concretos que van más allá del desarrollo de los temas que necesariamente se han de considerar y que, de alguna forma, imposibilitan su equivalencia.

Las investigaciones llevadas a cabo en torno a la estimación de la dificultad que plantean los diferentes temas a tratar en los ensayos ha resultado compleja. A modo ilustrativo, podemos citar el estudio de Alderson (1993) que pone en evidencia la falta de consenso que da entre los correctores, todos ellos con una experiencia profesional considerable en el campo de la evaluación de la lengua, a la hora juzgar la dificultad de los diversos ítems que componen una prueba de comprensión lectora. Alderson (1993) concluye:

> Judges were unable to predict with any degree of accuracy the difficulty of test items and subtests. The results established the need to pretest items and subtests, independently of the opinions of test constructors, as to item difficulty. (p. 56)

Hamp-Lyons (1991d) también se muestra bastante crítica ante el papel que desempeñan los correctores en la determinación de la dificultad implícita que supone el desarrollo de los diversos temas de los ensayos. La autora asegura:

> ... once again I realize how improbable it is that we will ever pin down prompt difficulty, since so much of the difficulty is not in the prompt or the writer, but in the reader. (p. 103)

Sin embargo, cabe pensar, que los diferentes temas que se ofertan en los ensayos exigen distintas demandas cognitivas a los candidatos, lo que condiciona el tipo de respuesta que se produce. Las investigaciones realizadas sobre los aspectos retóricos que se plantean en el desarrollo de los temas de los ensayos nos van a permitir estimar su grado de dificultad. No obstante, determinar la dificultad que plantean dichos aspectos retóricos ha suscitado cierta controversia entre los investigadores.

Así, Rushton y Young (1974) consideran que la elección del tema ejerce una poderosa influencia en la selección sintáctica del ensayo. Por el contrario, Reid (1990) demuestra que, independientemente de la tarea de producción escrita que se asigne, los candidatos utilizan construcciones sintácticas muy similares en las diversas tareas y temas. Esto se debe a la presión que supone escribir bajo el tiempo establecido en la prueba y a la

poca disposición de tiempo que se concede a los candidatos para revisar el texto escrito de forma adecuada. Reid (1990) admite, sin embargo, que no sucede así con la selección léxica. Esta última, efectivamente, parece variar de forma significativa según el tema que se desarrolle en el ensayo.

Por su parte, Mohan y Low (1985) observan diferencias en los diversos modos de escritura que utilizan los candidatos pertenecientes a distintas culturas. Estos autores demuestran que los estudiantes japoneses perciben una mayor dificultad en el desarrollo de temas de carácter expositivo y argumentativo que en el desarrollo de temas descriptivos o narrativos. Los resultados obtenidos confirman la hipótesis de otros estudios que señalan que las influencias retóricas de la primera lengua afectan la producción escrita en segundas lenguas (Grabe y Kaplan 1989; Reid 1990).

La investigación realizada por Bridgeman y Carlson (1983) y Carlson *et al.* (1985), que dio lugar a la inclusión de la opción del *Test of Written English* (TWE) en el *Test of English as a Foreign Language* (TOEFL), recomienda incluir dos tipos de texto en los ensayos. Estos textos son: comparar / contrastar un tema o desarrollar un tema a partir de la interpretación de un gráfico o tabla. Carlson *et al.* (1985) afirman que los temas que se elaboran a partir de estos textos correlacionan significativamente. Hamp-Lyons (1991d) o Spaan (1993) subrayan, sin embargo, que las correlaciones obtenidas, situadas en torno a los 0,66 y 0,73, no son lo suficientemente altas como para que ambas opciones se consideren equivalentes.

Spaan (1993) decide investigar el efecto de los aspectos retóricos en las valoraciones globales que se asignan a los ensayos en *The Michigan English Language Assessment Battery* (MELAB). Los textos que se utilizan aquí se clasifican desde un punto de vista retórico y de contenido en: Narrativo / Personal (NP) y Argumentativo / Impersonal (AI). Según Spaan (1993), los ensayos que se elaboran a partir de estas opciones reciben puntuaciones holísticas similares, lo que demuestra que las evaluaciones holísticas no se ven afectadas por la elección del tema. Spaan (1993) afirma, además, que los candidatos se muestran consistentes en el manejo de los aspectos lingüísticos en los dos tipos de ensayo y expone: “one is struck by the linguistic similarity of each writer’s two essays.” (p. 113). Esta autora continúa diciendo:

> Providing a prompt choice, both safe and challenging, may accommodate a broad proficiency range, keeping in mind that what is difficult for one person may not be for another. The choice may prove beneficial from an affective standpoint but remain neutral from the standpoint of performance or scoring. (Spaan 1993: 115)

No obstante, Reid (1990) contradice estos resultados. Este autor declara que la elección del tema es un factor decisivo en la elaboración de los ensayos dado que determina la complejidad sintáctica como la riqueza léxica y organizativa de los mismos. Desde esta perspectiva: “choosing topics should be the teacher’s (as well as the tester’s) most responsible activity.” (Reid 1990: 205).

Al igual que Reid (1990), diversos autores afirman que la calidad del ensayo se atribuye directamente al desarrollo del tema escogido (Hirsch y Harrington 1980; Applebee 1983). Para evitar este problema, Hartog (1936) recomienda adoptar varias medidas. En primer lugar, no se ha de permitir elegir los temas a los candidatos. En segundo lugar, se hace imprescindible limitar el tema que se vaya a desarrollar en el ensayo. Esto se consigue a través de la especificación del objetivo que se persigue en el texto y de la audiencia a la que el texto se dirige. Hamp-Lyons (1991d) se muestra un tanto escéptica ante la efectividad de esta propuesta:

> We are often told that writers write more and better when they are given a purpose for writing, but I know of no studies of this. We must keep in mind that when any writer confronts a writing assessment, even the least sophisticated knows already that the test's purpose is to establish his or her writing ability, and that the audience is the assessor. To establish any other purpose or any other audience seems to be adding confusion to a task which is relatively clear. (p. 92)

Por su parte, los partidarios de la oferta de elección de temas recomiendan seleccionar temas que sean sencillos y similares entre sí, en la medida de lo posible. Para conseguir este objetivo, Hoetker y Brossell (1986) proponen el desarrollo de temas muy restrictivos que van a ser objeto de duras críticas.

Así, Hamp-Lyons (1991d) argumenta que los temas restrictivos generan respuestas restrictivas y que lo que realmente se intenta es que los temas que se traten favorezcan el desarrollo de estructuras complejas y el

fomento de ideas. Asimismo, Teddick (1990) expone que reducir los temas y adecuarlos a un nivel general, que el autor denomina *Everyman*, perjudica a los candidatos que disponen de un conocimiento específico sobre el tema en cuestión. Según este autor, se ha de permitir que los candidatos demuestren su habilidad de producción escrita de forma efectiva y no exigirles, por el contrario, una demostración de un dominio lingüístico mínimo.

La alternativa propuesta por algunos autores de selección de temas de composición libre o de contenido neutral como, por ejemplo, *Red*, también ha sido descartada. A este respecto, Applebee (1981) señala que la producción escrita es un vehículo de comunicación y de descubrimiento de conocimiento por lo que los temas que se ofrecen a los candidatos deberían resultarles familiares.

Una última línea de investigación vincula el estudio de la oferta de temas a los diferentes niveles lingüísticos de los candidatos. Así, Pollit y Hutchinson (1987) demuestran que la diversidad de temas que se ofrecen a los candidatos afecta principalmente a los candidatos que poseen un dominio de la lengua limitado. Los candidatos que disponen de un dominio lingüístico superior no parecen verse afectados por esta medida. También Spaan (1993), niega que los candidatos con un dominio de la lengua alto produzcan textos escritos de mayor calidad según los temas que trabajen. No sucede así con los candidatos con un dominio lingüístico inferior, los cuales se ven afectados por el tema a tratar.

Hamp-Lyons (1991d) opina que la habilidad de seleccionar un tema u otro favorece en mayor medida a los candidatos con un buen dominio de la lengua. La autora afirma:

> ...I see that the students who are most upset by encountering an essay test topic that is not like they had expected are those whose scores suggest that they are weak writers, while those with higher scores seem able to adapt to different topics. (p. 95)

Por su parte, Ruth y Murphy (1984) exponen que la oferta de temas únicamente provoca confusión dado que la elaboración de temas conlleva distintos grados de dificultad y, sobre todo, porque los candidatos pueden no saber escoger el tema que más les conviene. Es decir, los candidatos pueden elegir temas complejos y producir, por ende, ensayos complejos que no por ello se convierten en producciones escritas de calidad. Swinford (1964), por ejemplo, manifiesta en su estudio que los candidatos mejor preparados tienden a elegir los temas más difíciles y, en consecuencia, obtienen puntuaciones más bajas ya que el nivel de dificultad de la tarea es mayor.

En vista de los resultados tan contradictorios que nos ofrece la literatura, Spaan (1993), prudentemente, decide investigar la opinión de los estudiantes de segundas lenguas sobre la oferta de selección de temas a tratar en los ensayos. Los resultados que obtiene son los siguientes: de los setenta y siete sujetos que contestan el cuestionario que se les administra, el 75% afirma que desea disponer de la oportunidad de elegir entre los temas

que se le ofrecen. El 11%, por el contrario, no desea que se le faciliten opciones. Por último, el 4% no indica ningún tipo de preferencia. Por su parte, Papajohn (1999), en un número de entrevistas concedidas a varios estudiantes, observa que éstos identifican varias diferencias entre los diversos temas a tratar y que, de hecho, muestran preferencias por determinados tipos de temas.

En cualquier caso, la falta de resultados concluyentes deja la decisión de ofertar temas en manos de los profesores. Según Hamp-Lyons (1991d):

> We might say that at the moment the answer to the question of whether a choice of prompt is a good or a bad thing depend less on the research evidence and more on our philosophical and pragmatic orientations. (p. 90)

Como explicaremos en el capítulo 8 (subepígrafe 8.4.2), en las PAAU se ofrecen, normalmente, dos opciones de tema a elegir a los candidatos. Estos temas se relacionan con el contenido de un texto de lectura sobre el que se basa el desarrollo de la prueba de Inglés.

7.3.4. El factor tiempo

Otra de las variables que se considera parcialmente responsable de la calidad del producto final del ensayo es la disposición del tiempo con el que cuentan los candidatos para realizar la tarea.

La evaluación de la competencia en lengua inglesa que se lleva a cabo en la prueba de Inglés de Selectividad estipula una duración máxima de 1:30h para el desarrollo global de la prueba para los estudiantes de

LOGSE[10]. Cabe decir que, anteriormente, la duración fijada para la realización de la prueba era tan sólo de 1h. Este período de tiempo decidió extenderse debido a la estrecha relación que se establece, generalmente, entre el factor tiempo, la corrección gramatical y el control de aspectos discursivos como son la organización o la coherencia de los textos. De hecho, algunos profesores consideran que escribir bajo la presión del tiempo establecido constituye una situación artificial que no estimula al candidato a elaborar una buena pieza de producción escrita.

Sin embargo, los estudios llevados a cabo en torno al factor tiempo y su posible influencia en las puntuaciones globales de los correctores nos muestran resultados contradictorios. Kroll (1990c), por ejemplo, expone en su estudio que los 5 grupos experimentales (i.e. sujetos árabes, chinos, japoneses, persas y españoles) mostraron una ligera mejoría en los ensayos, tanto en los aspectos de corrección gramatical como en los aspectos retóricos, tras una concesión mayor de tiempo para elaborarlos.

No obstante, la producción escrita en clase, fijada en un período de 60 minutos, y aquella asignada fuera de clase, en la que los candidatos disponían de un margen de tiempo de 10-14 días para su preparación, no presenta diferencias cualitativas significativas:

> These findings have shown that while the time allowed for the preparation of an essay can contribute to some improvement for the writer both on the syntactic level and the rhetorical level, it does not appear that additional time *in and of itself* leads to a sufficiently improved essay such that there is a statistically significance to the differences between class and home performance. (Kroll 1990c*:* 150)

[10] La abreviatura corresponde a Ley Orgánica General del Sistema Educativo (1992)

El énfasis que pone la autora en las palabras “in and itself” indica que la simple extensión del margen de tiempo que se concede a los candidatos para elaborar los ensayos no garantiza una producción escrita satisfactoria. En otras palabras, la ampliación del plazo de tiempo establecido para el desarrollo del ensayo no es una técnica productiva si no viene acompañada por una formación adicional en el proceso de escritura. De hecho, si los candidatos no disponen del conocimiento suficiente de lo que se entiende por una producción escrita de calidad seguirán mostrando la misma carencia de habilidades, independientemente de las condiciones que se establezcan para su desarrollo.

Kroll (1990c) declara que los candidatos de su estudio pueden haber dedicado menos tiempo a la producción escrita fuera de clase que a la que se les exige dentro del aula. La autora escribe:

> Without any mental formulation of what constitutes good writing or an awareness of the steps involved in producing it, students cannot know how to proceed in the task of writing and time could not buy them anything. To remedy this type of situation, teachers need to train students in a repertoire of strategies for composing as well as to recognize the attributes of effective writing. (Kroll 1990c*:* 152)

Esta perspectiva sugiere que la figura del profesor es fundamental para formar adecuadamente a los estudiantes en el proceso de composición de la producción escrita. La disposición de esta ayuda permitirá a los

estudiantes maximizar el breve período de tiempo que se les ofrece para elaborar sus pruebas escritas en evaluaciones como las PAAU.

7.4. Variables relativas a los candidatos

En este epígrafe expondremos brevemente algunas de las variables más relevantes relacionadas con los candidatos.

7.4.1. Introducción

Los estudios en torno a la figura del escritor o candidato que participa en las pruebas evaluadoras son muy reducidos. Sin embargo, los estudiantes de segundas lenguas que concursan a las pruebas de lengua, especialmente aquellas de carácter internacional, forman un grupo muy heterogéneo en cuanto a aspectos culturales, sociales y económicos se refiere.

7.4.2. La figura del escritor

Hamp-Lyons (1991c) expone que en escasas ocasiones se considera a los candidatos (i.e. escritores) como personas con entidad propia en las discusiones que se establecen en las pruebas de producción escrita. En el mejor de los casos, se analizan los datos demográficos o sociométricos de los candidatos (i.e. lugar de origen, lengua materna, etc.) y se evalúa su producción escrita teniendo en cuenta dichas variables. Hamp-Lyons (1990) considera que esta actuación resulta lamentable dentro de un marco de

evaluación de la lengua que debiera ser defendible desde un punto de vista personal y humanístico. Esta autora manifiesta:

> It is a sad irony that in writing assessment research there is a real tendency for the writer to be forgotten in the difficulties and controversies surrounding such issues such as topic choice, construct validity versus reliability, and the like. If there is any justification for this in L1 writing assessment (which I personally cannot see), there is surely none in L2 writing assessment research. (Hamp-Lyons 1990: p. 76)

No obstante, la clasificación independiente de las variables potenciales que influyen en la figura de los *escritores* resulta problemática. Esto se explica por la multiplicidad de factores que interactúan en la evaluación de la producción escrita. Esto es: el grado de complejidad de las diversas tareas, la fiabilidad de los correctores o los distintos métodos de evaluación que se utilizan.

No obstante, todos somos conscientes de que los candidatos van a verse enormemente afectados por aspectos relacionados con sus emociones, experiencias, ideas y personalidad en la producción escrita dado que ésta se concibe como una tarea personal en la que: “each writer brings the whole of himself or herself to the task at hand” (Hamp-Lyons 1990: 77).

Algunos autores destacan la influencia que ejerce el contexto sociocultural en la expresión personal (Hale-Benson 1986; Kaplan 1987; Hamp-Lyons 1990). Según estos autores, para llegar a entender el proceso de construcción de los textos y su composición necesariamente se han de tener en cuenta los factores contextuales

Por otro lado, los diferentes puntos de vista que poseen los sujetos como miembros de su comunidad se han de acomodar a las convenciones retóricas y estilísticas que exige la comunidad académica. Como dijimos, esta tarea puede resultar especialmente problemática para los sujetos que no logran identificarse con la nueva comunidad. De hecho, algunos candidatos pueden experimentar sentimientos encontrados y contradictorios (McCarthy 1987).

Para resolver este conflicto, Rosen (1969) propone presentar temas a los que los candidatos encuentren sentido responder. Esto significa que la tarea se ha de interpretar como realista y apropiada. De este modo, se consigue cumplir con las expectativas que exige la tarea al tiempo que se favorece la interacción entre la persona del escritor y la de la producción escrita.

Wilkinson (1983) critica los modelos de producción escrita que no incluyen instrumentos de medida para evaluar los aspectos afectivos y morales de los candidatos. Sobre todo, cuando estudios etnográficos han demostrado que los correctores juzgan y evalúan dichos aspectos a no ser que se les forme y enseñe para no tenerlos en cuenta (Hamp-Lyons, 1990).

Uno de los modelos más completos y exhaustivos que aborda el estudio de los aspectos afectivos de los candidatos es el que nos brindan Bachman y Palmer (1996). Estos autores destacan cuatro grupos de características individuales y personales que conviene considerar en el diseño, desarrollo y uso de las pruebas:

1. características personales de los candidatos. Entre ellas se incluye: la edad, el sexo, la nacionalidad, la residencia, la lengua nativa, el nivel de educación y el tiempo de preparación o familiaridad con la prueba evaluadora en cuestión
2. Conocimiento que los candidatos poseen del tema para la elaboración de la prueba (i.e. *topical knowledge*)
3. Esquema afectivo del candidato (*affective schemata*)
4. Habilidad lingüística del candidato

(Bachman y Palmer 1996: 64).

Hamp-Lyons (1991c), por su parte, apuesta por la realización de una observación detallada de la figura de los escritores en el momento de la evaluación. Este proceso nos permitirá estudiar el papel que juegan las diferentes culturas, actitudes y aspectos personales de los candidatos en la producción del texto escrito. La autora denuncia la falta de estudios cualitativos que arrojen algo de luz sobre quienes escriben nuestras pruebas y se pregunta: "...whether we would be better spending our time and research energies talking with students and learning about them *from* them." (Hamp-Lyons 1991c: 63)

Como dijimos, durante el desarrollo de las PAAU, se omite cualquier dato o referencia personal que pueda identificar a los candidatos. Con esta medida se pretende evitar cualquier tipo de sesgo o prejuicio motivado por

determinados aspectos personales de los candidatos (i.e. edad, sexo, nacionalidad, etc.).

Los juicios morales que realizan los correctores a partir de las características del texto escrito es una variable más difícil de controlar. A pesar de ello, en las sesiones previas de coordinación se les recuerda a los correctores la necesidad de emitir juicios imparciales y objetivos.

SEGUNDA PARTE

CAPÍTIULO 8. DESCRIPCIÓN DEL PROCESO METODOLÓGICO

8.1. Introducción

En este capítulo se describen los principales elementos de la muestra de estudio de este trabajo y algunas de las características básicas que enmarcan el desarrollo de las PAAU.

8.2. Sujetos

Los participantes de este estudio fueron treinta y dos correctores procedentes de enseñanzas medias de centros públicos de las comunidades de Madrid y Baleares y de enseñanza universitaria (Universidad Complutense de Madrid, U.C.M. y *Universitat de les Illes Balears* , U.I.B.). La principal consideración en la selección de los sujetos fue la garantía de su participación previa en la corrección de la prueba de Inglés de Selectividad.

Los treinta y dos correctores se clasificaron en cuatro grupos de ocho correctores. Los criterios que se aplicaron para la formación de los distintos grupos responde a dos razones fundamentales:

1) Los grupos seleccionados constituyen, por su bagaje profesional y por la diversidad de experiencias, creencias y expectativas, una muestra representativa de la población objeto de estudio y sobre la cual se desea extrapolar los resultados.

2) Los grupos seleccionados nos permiten investigar la influencia de dos variables principales como son el género y el lugar de trabajo de los correctores en las puntuaciones de los ensayos.

Uno de los principales inconvenientes que planteó el diseño de este estudio fue conseguir un número razonable de correctores que incluyera el mismo número de hombres y de mujeres de enseñanzas medias y enseñanza universitaria. Esto se debe a diversas causas:

1) la ratio es de 5 mujeres por cada hombre que trabaja en el campo de la enseñanza de lenguas.
2) La falta de disponibilidad de los sujetos y la incomodidad que supone el tener que evaluar a otros y ser observado en ese mismo proceso de evaluación *gratiis et amore*. Como detalle, porque en el ámbito académico servicios como este son prácticamente impagables, se gestionó con la UIB la posibilidad de ofrecer una pequeña compensación económica a los participantes de este estudio para que de algún modo se les agradeciera el esfuerzo realizado.

Todos los correctores que accedieron participar en este estudio de acuerdo con los criterios establecidos se incluyeron en la muestra. De este modo, se consiguió la participación total de treinta y dos correctores. El trabajo de la muestra es razonable para cualquier trabajo empírico ya que a

efectos de generalización se requiere una muestra mínima de 30 sujetos. Como ya dijimos, en nuestro caso conseguir este tamaño de muestra es muy difícil debido los factores arriba comentados y al rechazo de los sujetos a ser observados y evaluados en la práctica de sus funciones.

Asimismo, la orientación cuantitativa de este trabajo exigió un incremento en el tamaño de la muestra de los ensayos de acuerdo con el número de correctores disponibles. El trabajo que se les exigió a los correctores fue considerable:

1) Rellenar un cuestionario de 45 ítems
2) Realizar 20 puntuaciones holísticas
3) Realizar 140 puntuaciones analíticas
4) Desarrollar un mínimo de 40 comentarios: 20 positivos y 20 negativos sobre los ensayos

A todos los correctores se les garantizó el anonimato en sus actuaciones.

La distribución de los grupos quedó establecida del siguiente modo:

Fig. 8.1. Correctores del estudio

Grupo 1: Mujeres de enseñanzas medias. Correctores numerados del 1 al 8 (R1-R8)*
Grupo 2: Mujeres de universidad. Correctores numerados del 9 al 16 (R9-R16).
Grupo 3: Hombres de enseñanzas medias . Correctores numerados del 17 al 24 (R17-R24).
Grupo 4: Hombres de universidad. Correctores numerados del 25 al 32 (R25 al 32).

*La abreviatura utilizada para identificar a cada uno de los correctores en este estudio es Rn a efectos de identificación de género y situación laboral.

8.3. Materiales

Los materiales que se utilizaron en este estudio fueron los siguientes (ver Apéndice):

1. Diez ensayos originales procedentes de las PAAU (Apéndice 3)
2. Un cuestionario (Apéndice 1). Este cuestionario consta de tres partes fundamentales:
   - I) Parte: datos sociométricos de los correctores
   - II) Parte: la figura del corrector. Este parte se encuentra subdividida en cuatro apartados:
     - A) Perfil docente: estilos de enseñanza;
     - B) Rasgos de la personalidad;
     - C) Autoevaluación del corrector como tal y
     - D) Identificación de los principales aspectos problemáticos de los ensayos
   - III) Parte: Evaluación de los errores en textos aislados
3. Técnicas de corrección para la evaluación de los ensayos (Apéndice 2). Los correctores debían facilitar la siguiente información sobre cada uno de los ensayos:
   1) puntuación holística;
   2) puntuaciones analíticas y
   3) aspectos positivos y negativos del ensayo

La explicación de cada uno de estos materiales se detalla continuación:

1. *Diez ensayos originales procedentes de las PAAU.*

Se seleccionaron de forma aleatoria diez ensayos originales escritos por los estudiantes que participaron en las PAAU en la convocatoria de junio de 2000. Los ensayos de este estudio proceden de una muestra de ciento noventa y seis ensayos que la comisión de acceso a los estudios universitarios de la UIB asignó a la autora de este trabajo.

La política que aplica dicha comisión para la distribución de las pruebas consiste en la repartición de un número similar de pruebas a cada uno de los correctores siendo este el único criterio que rige su distribución. En este sentido, la aleatoriedad de las pruebas está garantizada.

Los ciento noventa y seis ensayos anteriores se dividieron en dos grupos principales según la opción del tema elegido por los candidatos. La opción sobre la que se basa este estudio fue la opción A) *Write a composition of 100-150 words on the following topic:- A holiday in London, mentioning the places you would visit and why.* Dicha opción incluía un total de ciento treinta y seis ensayos frente a los sesenta ensayos que desarrollaron la opción B.

Los ciento treinta y seis ensayos de la opción A se dividieron nuevamente en tres grupos representativos de tres dominios de la lengua inglesa de los candidatos (i.e. bajo, medio y alto) de acuerdo con las puntuaciones iniciales que esta misma autora les asignó. De este modo, se

obtuvieron tres grupos: el grupo de dominio de la lengua inglesa bajo (incluía ensayos con puntuaciones de 1 hasta 4 puntos*[1]), que contenía 40 ensayos, el grupo de dominio de la lengua inglesa medio (incluía ensayos con puntuaciones de 4 hasta 6 puntos), que contenía 66 ensayos y, por último, el grupo de dominio de la lengua inglesa alto (incluía ensayos con puntuaciones de 7 hasta 10 puntos), que contenía 30 ensayos.

La muestra se completó eligiendo de forma aleatoria tres ensayos del primer y último grupo, es decir, del grupo de dominio de la lengua clasificado como bajo y alto y cuatro ensayos del grupo medio que, por otra parte, era el más numeroso. Se obtuvieron de este modo diez ensayos representativos de tres niveles diferentes de dominio de la lengua inglesa. Una vez reunidos los diez ensayos finales se mezclaron y se estableció su orden de corrección de forma aleatoria. Se obtuvo, de este forma, la muestra definitiva sobre la que se basó este estudio.

2. Cuestionario

Los correctores debían responder a un cuestionario que se incluía en la primera parte del estudio (ver Apéndice 1). Este cuestionario se hallaba dividido en tres partes que recogen información sobre los siguientes aspectos:

I) Parte: **Datos sociométricos de los correctores.** Se investigan las variables género y lugar de trabajo de los correctores y su posible influencia en las puntuaciones de los ensayos. El resto

[1]Tradicionalmente, en las instituciones españolas la asignación de puntos se realiza sobre una base numérica de 10.

de las variables se toma como punto de referencia para futuros estudios ya que el tamaño de las distintas categorías de cada una de las variables (i.e. edad, lengua nativa, etc.) no permitiría generalizarla al resto de los resultados.

II) Parte: **La figura del corrector**. Esta parte se encuentra a la vez subdividida en cuatro apartados:

A) *Perfil docente: estilos de enseñanza*. Basándonos en un estudio de Dreyer (1998) sobre los diversos estilos de enseñanza, y siguiendo la máxima de Kinsella (1995) cuando asegura que: “Although there is probably some truth to the maxim that teachers teach the way they were taught, there is probably a lot more truth in saying that teachers teach the way they learned best in school.” (p. 85), nos proponemos analizar los diversos estilos de enseñanza y aprendizaje que manifiestan los correctores. Con ello, se pretende averiguar hasta qué punto estos estilos afectan las valoraciones del rendimiento de los candidatos.

B) *Rasgos de la personalidad*. Se observan aspectos como: el grado de optimismo, la satisfacción que produce el trabajo, la confianza, la autoestima y la realización personal. El estudio de estos aspectos intenta relacionar la personalidad de los correctores con su comportamiento en la evaluación de los ensayos.

C) *Autoevaluación del corrector como tal*. En este punto, se analiza la opinión de los correctores sobre su faceta evaluadora. Los correctores evalúan su dominio de la lengua inglesa y su grado o nivel de condescendencia, flexibilidad, consecuencia y aptitud.

D) *Identificación de los principales aspectos problemáticos de los ensayos*. Basándonos en un estudio de Johns (1991) sobre la identificación de los principales aspectos problemáticos que presenta la producción escrita académica en segundas lenguas, se confecciona una lista que incluye seis problemas básicos que se han observado en los ensayos de lengua inglesa como segunda lengua. Se les pide a los correctores que elijan tres problemas de la lista que consideren fundamentales y que los clasifiquen en orden de importancia. Asimismo, se anima a los correctores a que apunten cualquier otro criterio u aspecto problemático no incluido en la lista que consideren relevante. Con respecto a este último punto hay que decir que no hubo aportaciones o sugerencias personales. No obstante, este comportamiento se halla justificado dado el esfuerzo del trabajo exigido en este estudio.

III) Parte. **Evaluación de los errores en textos aislados.** Finalmente, se pide a los correctores que puntúen la gravedad de los errores contenidos en una lista de catorce oraciones

originales extraídas de la muestra de ciento treinta y seis ensayos sobre los que se basa este estudio. Cada una de las oraciones incluye una muestra de error común cometida por los estudiantes de las PAAU. Los errores se indican en letra cursiva para facilitar su localización. Estos errores se asocian a siete categorías analíticas[2] que posteriormente se presentarán a los correctores durante la evaluación de los ensayos. Las categorías que se incluyen son: *contenido*, *organización*, *gramática*, *vocabulario*, *registro*, *mecánica* y *presentación*. Se incluyen dos muestras representativas de error para cada una de la siete categorías analíticas.

La información recogida en esta última parte del cuestionario nos permitirá contrastar la actuación de los correctores en la evaluación de los errores contenidos en textos descontextualizados (i.e. las oraciones) con la evaluación de los errores incluidos en el discurso (i.e. los ensayos).

3. *Técnicas de corrección para la evaluación de ensayos* (ver Apéndice 2)

Cada uno de los diez ensayos de la muestra final de este estudio se reproducía en una primera hoja que iba unida a una segunda hoja mediante

[2] Las categorías analíticas que se incluyen en este estudio tienen su punto de partida en la literatura de la evaluación (Diederich *et* al 1961; Jacobs *et al* 1981) y en algunos estudios propios conducidos con anterioridad. Estos estudios nos permitieron identificar los aspectos relacionados con la forma y el contenido que ejercen una mayor influencia en las valoraciones de los correctores (Amengual y Herrera, 2000).

una grapa. En esta segunda hoja se requería a los correctores la siguiente información:

1) **Puntuación holística**. Los correctores tenían que puntuar cada uno de los ensayos de forma individual basándose en la impresión holística o global que éstos le producían. Para ello debían utilizar una escala de valoración de 1(pésimo) a 10 (muy bueno) evitando hacer uso de números decimales. No se les facilitó a los correctores ningún criterio de evaluación específico dado que uno de los objetivos principales de este trabajo consistía en averiguar los criterios de corrección que cada uno de los correctores establecía y aplicaba en la evaluación de los ensayos. Para ayudarles en su tarea, se les recordó a los correctores el nivel de dominio de la lengua de los estudiantes y se les aconsejó que se dejaran guiar por su propia experiencia en la corrección de los ensayos de Selectividad.
2) **Puntuaciones analíticas**. Los correctores debían evaluar cada uno de los siguientes aspectos analíticos del ensayo: contenido, *organización*, *gramática*, *vocabulario*, *registro*, *mecánica* y *presentación*. Para ello, debían utilizar una escala de Likert de 5 puntos: 1 (malo) y 5 (muy bueno).

   El cambio de escala se realizó con la finalidad de evitar, en la medida de lo posible, el que los correctores pudieran relacionar los resultados de los componentes analíticos con la valoración

holística que le habían otorgado a cada una de los ensayos en primer lugar.

3) **Aspectos positivos y negativos del ensayo.** Finalmente, se les pidió a los correctores que anotaran la expresión o estructura que destacarían de forma positiva y negativa en cada uno de los ensayos.

El conjunto del material iba acompañado de una lista de instrucciones contenida en los diferentes sobres que se repartieron a los correctores. Las instrucciones incluían las diferentes normas a seguir para la realización de este trabajo. (ver Apéndice 1)

8.4. Contexto

8.4.1. Las PAAU o pruebas de Selectividad

La prueba de Inglés forma parte junto con otras pruebas pertenecientes a asignaturas diversas como son Lengua española, Lengua propia de la comunidad (Catalán en la comunidad de las Islas Baleares), Historia, Filosofía, etc. de las PAAU o Selectividad. Estas pruebas, de carácter común y obligatorio en todo el estado español, se plantean como objetivo primordial la homogeneización del conjunto de calificaciones obtenidas por los estudiantes al término de la enseñanza secundaria o Formación profesional de segundo grado o grado superior[3]. En este sentido,

[3] Para el acceso a las distintas universidades del estado español, los estudiantes necesitan haber cursado una de las siguientes opciones: COU (Curso de Orientación Universitaria), que se extingue en el curso académico 2002-2003, bachillerato LOGSE (Ley Orgánica

los resultados que se obtengan en las PAAU o pruebas de Selectividad van a condicionar la elección de la carrera universitaria de no pocos estudiantes españoles.

Como dijimos, las PAAU se pueden clasificar como pruebas de dominio dado que miden la capacidad o destreza general de los candidatos sin referirse a ningún programa de estudios en particular. Dichas pruebas, se entienden como pruebas normativas en las que la habilidad del candidato se compara con la del resto de los candidatos. Es decir, su finalidad consiste en discriminar entre los diferentes niveles de aptitud que presentan los estudiantes que concurren a estas pruebas.

Según nos comenta Herrera (1999), el objetivo específico que se plantea la prueba de Inglés es el siguiente: “In the case of the English Test, the target is to discriminate as reliably as possible.” (p. 2). Así pues, garantizar la fiabilidad de las puntuaciones de los correctores resultará esencial para poder proceder a la distribución de los candidatos según el nivel de dominio de la lengua inglesa de forma precisa y adecuada.

En este trabajo nos vamos a centrar concretamente en el análisis de la fiabilidad de las puntuaciones asignadas al ejercicio del ensayo incluido en la prueba de Inglés de Selectividad. Para ello, se tomará como referencia uno de los ensayos administrados en la *Universitat de les Illes Balears* (UIB) durante la convocatoria de junio de 2000.

La pregunta del ensayo requiere que los candidatos a través de una muestra libre de expresión escrita demuestren el dominio de esta última

---

General de Secundaria), Formación profesional de segundo grado o Formación Profesional de grado superior (módulo nivel 3 y ciclo formativo de grado superior).

habilidad. Los candidatos elaboran un ensayo breve (normalmente, se recomiendan entre 100 y 150 palabras) sobre un tema basado en una pieza de lectura original en lengua inglesa. Generalmente, los candidatos escogen el tema a tratar entre dos opciones alternativas.

Cabe decir que el planteamiento que sigue el desarrollo del ensayo en la UIB es común al del resto de las universidades del estado español salvo pequeñas diferencias u posibles excepciones.

8.4.2. El tema del ensayo

En la prueba de Inglés de Selectividad de LOGSE de junio de 2000 que tuvo lugar en la UIB se ofrecieron los siguientes temas opcionales:

Write a composition of 100-150 words on the following topic:

A) - *A holiday in London, mentioning the places you would visit and why*.

B) *Telephone a friend and invite him/her to spend a weekend with you in your home town*

Como se puede apreciar, ambos textos pertenecen a géneros distintos; la opción A) es un texto expositivo, la opción B) se trata de un diálogo o conversación entre amigos. Cada uno de los géneros requerirá estructuras sintácticas de mayor o menor complejidad pero en cualquier caso distintas. Sin embargo, y lejos de la línea de pensamiento que defienden numerosos estudios que aseguran que las características

retóricas y organizativas así como el desarrollo del tema afecta la producción de los candidatos (Ruth y Murphy 1988; Kroll 1990; Hamp-Lyons 1991; Sweedler-Brown 1992; White 1994), los criterios de evaluación que se aplican a los ensayos de Selectividad son los mismos en ambos textos.

En este trabajo, no obstante, debido a que nuestro interés principal era la medición de la fiabilidad de las puntuaciones de los correctores, creímos conveniente analizar únicamente aquellos ensayos que desarrollaran la misma opción temática. Descartamos, de este modo, cualquier otra variable relativa a la tarea que pudiera afectar los resultados obtenidos.

La opción A fue la opción elegida como objeto de estudio de este trabajo, entre otras razones, porque fue mayoritariamente seleccionada por los candidatos.

8.4.3. El factor tiempo

El tiempo específico que se concede a los candidatos para la elaboración del ensayo no viene estipulado en las regulaciones de las PAAU. Se indica, no obstante, la duración máxima que permite la prueba y que según el reglamento de la UIB varía entre una hora (estudiantes de COU) y una hora y media (estudiantes de LOGSE).

Cabe decir que los coordinadores de la prueba de Inglés de la UIB decidieron incrementar en media hora el margen de tiempo concedido para el desarrollo de la prueba de Inglés de los estudiantes de LOGSE a petición de los profesores y de los propios estudiantes. Según los profesores, la

escasez de tiempo asignada a la prueba ejercía una presión negativa en los candidatos que influía en la calidad del producto final de la misma. No obstante, no hay evidencia empírica en este tema que demuestre el efecto beneficioso que la adopción de dicha medida haya podido producir en el desarrollo y calidad de las pruebas.

8.4.4. Instrucciones y criterios de evaluación

Con la finalidad de homogeneizar las valoraciones de los diversos correctores, el coordinador de la prueba de Inglés de Selectividad facilita al panel de correctores pertinente las instrucciones relativas a la distribución de las puntuaciones para cada uno de los apartados de que consta la prueba de Inglés. Los criterios evaluativos que se aplicarán a dicha prueba se comentan conjuntamente. Las sesiones dedicadas al comentario y discusión de los criterios evaluativos se reducen una o dos en la mayoría de las ocasiones. Su duración aproximada es de 20-30m antes del inicio de la prueba y de la distribución del paquete correspondiente de ejercicios a cada uno de los correctores.

Los ensayos se evalúan, entonces, aplicando los criterios de evaluación que se estipulan. Estos criterios incluyen una serie de descriptores que, generalmente, resultan de la adaptación de diversas escalas de evaluación tradicionales (ver Herrera 1999). A modo ilustrativo, incluimos los criterios de evaluación relativos a la corrección del ensayo, que se corresponde con la pregunta número cinco de la prueba de Inglés, perteneciente a la convocatoria de las PAAU de junio de 2000 de la UIB (ver

Apéndice 4). Los criterios evaluativos que se ofrecen en la UCM y en el resto de las universidades españolas son muy similares salvo posibles excepciones.

Como se podrá apreciar, los descriptores evaluativos contienen escasa información sobre la naturaleza de los diferentes componentes del ensayo que se han de evaluar. Generalmente, se mencionan a grandes rasgos diversos elementos que deberían considerarse como, por ejemplo, el uso de vocabulario, la gramática, el contenido, etc. Estos últimos elementos van siempre dirigidos a la consecución de un objetivo básico que es el de la evaluación de una competencia comunicativa razonable en lengua inglesa.

Asimismo, la distribución del peso o valor que se atribuye a cada uno de los elementos del ensayo se establece de forma indicativa. Este hecho propicia el enfrentamiento de opiniones entre los correctores que defienden diversas teorías lingüísticas sobre la enseñanza y la evaluación de la lengua. Este último planteamiento también podría explicar en cierta medida la falta de consenso que se observa entre los diversos correctores a la hora de identificar los aspectos que se han de evaluar y el modo en que éstos han de ser evaluados. Presumiblemente, los diversos criterios de evaluación se aplican para cada ensayo en el mismo momento de la corrección de forma casi impulsiva. Según algunos autores (Herrera 1999.; Vaughan 1991), dicha actuación explicaría la falta de fiabilidad de las puntuaciones de los correctores, quienes, en no pocas ocasiones, acaban dejándose guiar por su experiencia personal.

La pregunta del ensayo representa, generalmente, el 40% de la nota final de la prueba de Inglés, lo que refleja la gran validez de constructo que se le atribuye en las PAAU. De hecho, la pregunta del ensayo es la que determina en muchas ocasiones la puntuación definitiva que se le concede a la prueba de Inglés.

La puntuación máxima de la prueba de Inglés es de 10 puntos y los estudiantes deben obtener una puntuación mínima de 5 puntos para considerar dicha prueba superada. Normalmente, se permite hacer uso de notas enteras y de fracciones de puntos (0,5 y 0,25) por lo que se tiende a redondear las puntuaciones finales en el último minuto cuando se procede a la suma total de calificaciones. Las puntuaciones máximas que corresponden a cada una de las preguntas de la prueba de Inglés se señalan entre paréntesis al final del enunciado de las mismas. La distribución de las puntuaciones señala la importancia que se le atribuye a cada una de las preguntas. El ejercicio del ensayo es la pregunta que mayor puntuación recibe en la prueba de Inglés[4].

Por otro lado, los criterios evaluativos que se aplican al ensayo nos indican la distribución específica de las puntuaciones para sus distintos

[4] Las distintas preguntas que se incluyen en la prueba de Inglés de Selectividad de las distintas universidades del estado español son de naturaleza flexible. No obstante, todas las preguntas se basan generalmente en un texto de lectura original escrito en lengua inglesa. La formulación de las preguntas así como las respuestas de los candidatos son en lengua inglesa. En la prueba de junio de 2000 de la UIB se incluyeron las siguientes preguntas: un resumen con una extensión máxima de 50 palabras (2 puntos), dos preguntas de comprensión abierta (1 punto cada una), una pregunta que incluía una sección de léxico en la que los estudiantes debían encontrar sinónimos de distintas palabras (1 punto) y una sección de sintaxis en la que se debían completar varias frases realizando las modificaciones sintácticas oportunas (1 punto). La última pregunta era la del ensayo (4 puntos). Cabe decir que no se permite el uso de diccionarios ni de cualquier otro material didáctico durante el desarrollo de la prueba.

componentes. Ello nos permite observar el valor que se concede a los aspectos relacionados con la *forma* y con el *contenido* del ensayo.

En este trabajo, sin embargo, hemos decidido no incluir ningún criterio evaluativo que sirva de guía a los correctores para evaluar los ensayos ya que uno de nuestros principales objetivos era el de descubrir los criterios evaluativos individuales que aplican los distintos correctores en la evaluación de los ensayos.

8.4.5. La formación de los correctores

El reglamento de las PAAU no estipula la formación obligatoria del profesorado en técnicas de evaluación para participar en la corrección de las pruebas de Inglés. La distribución del paquete de pruebas a cada uno de los correctores se realiza de forma aleatoria. El número de pruebas que debe evaluar cada corrector oscila entre las ciento cincuenta y las doscientas.

Como ya dijimos (ver capítulo 5, epígrafe 5.5.), las razones por las que no se forma a los correctores o se efectúa una doble corrección en la evaluación de las distintas pruebas se relacionan directamente con la cualidad de la factibilidad: la falta de personal docente, la presión de tiempo, el coste económico, etc. No obstante, si los estudiantes no consideran justa o adecuada la puntuación final que reciben en las pruebas pueden optar por la vía de la reclamación o de la doble corrección[5]. En este caso, se asigna

[5] Por la vía denominada de *reclamación*, un corrector se encargará de verificar que la corrección inicial se ha realizado de acuerdo con los criterios generales y específicos de la materia en cuestión. La elección de esta vía no originará el descenso de la nota inicial excepto en el caso de detectarse un error material de suma de calificaciones parciales. De ser así, el error material se reparará con el descenso o subida de la calificación correspondiente. Si se escoge esta vía, se excluye la posibilidad de optar por la vía denominada de *doble corrección*.

un nuevo corrector que se encarga de revisar o de evaluar nuevamente las pruebas.

Siguiendo este último planteamiento, los correctores que participaron en este estudio no recibieron ningún tipo de formación previa en técnicas de evaluación para la corrección de los ensayos ya que nos interesaba investigar la actuación y el comportamiento de los distintos correctores en un contexto similar al de las PAAU.

8.5. Procedimiento

La recopilación del material para este estudio se efectuó a principios de febrero y a principios de mayo de 2001. Todos los correctores participaron en las dos sesiones de recogida de datos preestablecidas que quedaron distribuidas del siguiente modo:

Fig. 8.2. Distribución del tiempo.

a) Una primera sesión a principios de febrero de 2001 (PRE)
b) Una segunda sesión a principios de mayo de 2001 (i.e. transcurridos tres meses después de la realización de la primera) (POST)*

* Las abreviaciones PRE y POST responden a la terminología convencional utilizada en el campo de la evaluación para designar la primera y segunda ocasión en la que se realiza el estudio.

Por la vía denominada *doble corrección*, un nuevo corrector se encargará de corregir de nuevo el ejercicio: se trata de una nueva corrección independiente de la inicial. La nota final es la media aritmética de las dos correcciones y, por lo tanto, la nota inicial puede subir o bajar. Si la diferencia entre las notas de las dos correcciones es mayor de tres puntos, un tercer corrector asigna la nota definitiva. La vía de reclamación pone fin a la vía administrativa.

Las sesiones de recogida de datos se describen a continuación. A partir de ahora utilizaremos las siglas PRE y POST para referirnos a la primera y segunda ocasión en la que se llevó a cabo este estudio:

a. **PRE**: A cada uno de los treinta y dos correctores que participaban en el estudio se le repartió un sobre que incluía la siguiente documentación: un cuestionario que se debía rellenar primeramente de acuerdo con las instrucciones escritas que se adjuntaban (ver Apéndice 1) y un paquete de 10 ensayos que desarrollaban la misma opción temática (ver Apéndice 3). Estos ensayos iban unidos cada uno de ellos a una hoja adicional en donde los correctores debían anotar la valoración holística, las valoraciones analíticas y la especificación de los elementos positivos y negativos de cada uno de los ensayos (ver Apéndice 2).

Como se recordará, la finalidad del cuestionario era obtener información sociométrica de los correctores (i.e. género y lugar de trabajo especialmente) así como información relativa a su perfil docente y algunos aspectos de su personalidad. También se pretendía averiguar la opinión de los correctores sobre los principales problemas que presentan los estudiantes de segundas lenguas en su producción escrita y la reacción de los correctores ante un número de errores comunes cometidos frecuentemente por los estudiantes e insertados en frases individuales y descontextualizadas.

El paquete de ensayos contenía copias de los 10 ensayos originales que se tomaron como muestra para este estudio. Los ensayos originales no

se alteraron ni modificaron. Por el contrario, se intentaron respetar, en la medida de lo posible, las características originales de los textos (i.e. tipo de escritura, inclusión u omisión de tachones, etc.).

Los ensayos se enumeraron del uno al diez de forma visible para los correctores. De acuerdo con las instrucciones escritas que contenía cada uno de los sobres, se les recordó a los correctores (y se subrayó con el uso de mayúsculas en las instrucciones, ver Apéndice 1) la necesidad especial de corregir los ensayos siguiendo el orden preestablecido, el cual era idéntico para cada uno de los correctores. La adopción de esta última medida pretendía eliminar los posibles efectos que pudieran producirse por el orden de presentación de los ensayos y que, probablemente, sería motivo de estudio de otra tesis doctoral.

El reparto de los treinta y dos sobres se efectuó personalmente y, de este forma, además de las instrucciones escritas, los correctores recibieron instrucciones verbales. El contacto personal también favoreció la resolución de posibles dudas que pudieran plantearse a lo largo del proceso de estudio.

Se les pidió a los correctores que devolvieran los sobres corregidos una semana después de la entrega inicial de los mismos. A pesar de que la mayoría de los correctores cumplió este último requisito, algunos de ellos extendieron dicho período a dos o incluso tres semanas por razones de trabajo.

En ningún momento se les comunicó a los correctores el objetivo de este trabajo aunque se les explicó que la información obtenida sería utilizada con fines pedagógicos. También se les pidió que no guardaran ningún

registro o copia de la información suministrada en el primer sobre. A todos los correctores que participaron en este estudio se les aseguró la confidencialidad de sus datos personales en la interpretación de los resultados obtenidos.

b. **POST**: No se volvió a contactar con los correctores nuevamente hasta transcurridos tres meses desde el inicio del experimento, esto es, a principios de junio de 2001. El período de tiempo que se sucedió entre la primera y la segunda sesión nos permitió realizar el estudio de la fiabilidad intra-evaluador o medición de la consistencia de un mismo corrector en ocasiones diferentes.

A cada uno de los treinta y dos correctores se les repartió un nuevo sobre que contenía los diez mismos ensayos que se evaluaron en la primera sesión (PRE). En esta ocasión los ensayos se presentaban en distinto orden para evitar posibles contaminaciones de la corrección anterior. El cuestionario no se volvió a administrar.

El nuevo orden de presentación de los ensayos en esta segunda sesión (POST) se estableció de forma aleatoria. Los diez ensayos quedaron distribuidos del modo que se señala a continuación:

Fig. 3. Orden de los ensayos

PRE: Orden numérico de los ensayos 1)C1* 2)C2 3)C3 4)C4 5)C5 6)C6 7)C7 8)C8 9)C9 10)C10
POST: Orden numérico de los ensayos 1)C10 2)C8 3)C3 4)C7 5)C9 6)C5 7)C2 8)C1 9)C6 10)C4

*Cada uno de los ensayos responde a la abreviatura Cn

Con respecto a este último punto, cabe decir que, sorprendentemente, algunos de los correctores no reconocieron los ensayos. No obstante, ciertos correctores observaron que, efectivamente, los ensayos en la primera y la segunda sesión (PRE y POST) eran los mismos. Este último hecho incomodó a alguno de ellos aunque, en general, los correctores respondieron de forma positiva. Debido al período de tiempo transcurrido entre la primera (PRE) y la segunda sesión (POST) y al hecho de que los correctores no guardaron registro alguno de las puntuaciones de los ensayos correspondientes a la primera sesión (PRE), pensamos que los resultados obtenidos en esta segunda sesión (POST) no fueron contaminados por los primeros. Asimismo, el orden distinto de presentación de los ensayos ayudó a contrarrestar el recuerdo de las primeras correcciones realizadas.

Por lo que se refiere al resto del proceso, en esta segunda sesión (POST) se procedió del mismo modo que en la primera (PRE). Los ensayos se entregaron numerados en el nuevo orden establecido. Se les volvió a

recordar a los correctores la importancia de corregir los ensayos respetando el orden prescrito. Cada uno de los ensayos se evaluó de forma holística y analítica. También se anotaron los aspectos positivos y negativos de cada uno de los ensayos. Las instrucciones escritas que se les entregó a los correctores fueron exactamente las mismas que las que se suministraron en la primera sesión (*PRE*) excepto que, esta vez, no se incluyó el primer punto que hacía referencia al desarrollo del cuestionario.

Los sobres con los ensayos corregidos se recogieron personalmente o se enviaron por correo al cabo de una o dos semanas de su entrega inicial a los correctores. A todos los participantes se les agradeció su colaboración. Muchos de ellos se mostraron interesados en conocer los resultados finales del estudio y se les comunicó que así se haría una vez elaborado y sometido a juicio el trabajo final.

8.6. Tratamiento de datos

La investigación de los perfiles docente, actitudinales y evaluadores de nuestros correctores se realizó a través del análisis factorial y de tablas de frecuencias.

El estudio de la fiabilidad inter-corrector se llevó a cabo a través de la implementación de diversas técnicas estadísticas.

Se aplicaron las correlaciones de Pearson y Spearman a las evaluaciones holísticas y a las analíticas tanto a nivel general como en el plano de lo concreto. Asimismo, se realizó el análisis de la fiabilidad a través

del cálculo del coeficiente de correlación intra-clase, que valora consistencia y acuerdo absoluto.

Posteriormente, se llevó a cabo el análisis de la varianza (ANOVA) y las pruebas de significación (*T* tests) sobre los datos obtenidos. Estas últimas se aplicaron también al estudio de los errores contenidos en los ensayos y en frases descontextualizadas.

El cálculo de frecuencias y la técnica de regresión múltiple nos permitió identificar las categorías que contribuyen en mayor medida a la puntuación holística final de los ensayos. La relevancia de cada uno de los componentes analíticos se examinó a través de tablas de frecuencia y gráficos.

Finalmente, se procedió al análisis de la fiabilidad intra-corrector y al estudio de los casos de correctores extremos o marginales a través de diagramas de dispersión.

La técnica estadística utilizada para el análisis de los datos fue el *Statistical Package for Social Sciencies* (SPSS 11.0.1).

CAPÍTULO 9. ANÁLISIS DE MATERIALES: EL CUESTIONARIO

9.1. Introducción

En este capítulo se analizarán los diversos perfiles docentes, actitudinales y evaluadores de los correctores que participan en este estudio. Los resultados que se presentan provienen de la sistematización de la información recogida en la II parte del cuestionario (ver Apéndice1) que se facilitó a los correctores al inicio de este estudio (PRE).

Los datos obtenidos van a permitirnos contrastar la actuación teórica de los correctores con su posterior actuación práctica. De esta forma, conseguiremos discriminar entre la tendencia general de los sujetos a buscar la *aceptabilidad social* en las respuestas que se dan a los planteamientos teóricos que se formulan en los cuestionarios y la realidad, esto es, su comportamiento en la evaluación.

El cuestionario se ha configurado en tres partes fundamentales que fueron brevemente comentadas en el capítulo 8 (epígrafe 8.3) pero que aquí se describirán más detalladamente:

I) Parte: datos sociométricos de los correctores

II) Parte: la figura del corrector. Esta parte se halla subdividida en cuatro apartados:

A) Perfil docente: estilos de enseñanza;

B) Rasgos de la personalidad;

C) Autoevaluación del corrector e

D) Identificación de los principales aspectos problemáticos de los ensayos

III) Parte: Evaluación de los errores en frases individuales

Como vemos, en la I Parte del cuestionario se recogen los datos sociométricos de los correctores. La obtención de esta información nos permitirá analizar la influencia de las variables género y situación laboral en las puntuaciones holísticas y analíticas de los ensayos, que se examinarán detenidamente en el capítulo 10 (epígrafe 10.2). El estudio más ambicioso del resto de las variables (i.e. edad, lengua nativa, etc) no pudo llevarse a cabo debido al reducido tamaño de los grupos obtenidos en la muestra. La investigación de dichas variables queda, sin embargo, abierta a estudios posteriores.

La III Parte del cuestionario relativa a la evaluación de los errores contenidos en frases descontextualizadas se abordará también en el capítulo 10 (subepígrafe 10.2.3.3). No obstante, en aras de la unidad y con la finalidad de respetar la visión global del cuestionario se hará una breve introducción al estudio de los errores en este capítulo dada su relevancia en este trabajo.

Así pues, el interés principal de este capítulo se centra en los resultados obtenidos en la II Parte del cuestionario.

Los datos obtenidos se presentan siguiendo el orden de los apartados establecidos en la II Parte del cuestionario.

9.2. Perfil docente de los correctores

El estudio del perfil de los estilos docentes de los correctores se llevará acabo mediante la técnica estadística del análisis factorial que comentamos en el siguiente subepígrafe.

9.2.1. Análisis factorial de los ítems que definen el perfil docente: estilos de enseñanza (sección A)

Basándonos en un estudio inicial de Dreyer (1998), elaboramos 11 enunciados representativos de los estilos de enseñanza que se detallan a continuación:

- Global (A1)[1] y analítico (A2)
- Preferencias sensoriales: visual (A3), auditivo (A4) y manipulativo (A5)
- Extraversión (A7) e introversión (A6)
- Intuitivo (A8) y concreto-secuencial (A10)
- Cerrado (A9) y Abierto (A11)

[1] Los diversos estilos de enseñanza contenidos en la sección A del cuestionario se representan con la abreviatura A*n*

La técnica del análisis factorial nos permite averiguar las relaciones que se establecen entre los diversos enunciados que definen los estilos docentes de los correctores (ver sección A del cuestionario en Apéndice 1)

La Tabla 9.1 nos muestra el valor que representa cada uno de los grupos de componentes que han quedado configurados en la interpretación de los resultados. Como vemos en esta Tabla, la formación de los cuatro grupos explica el 69,484 de la varianza, es decir casi el 70% de su valor.

Tabla 9.1. Varianza total aplicada

Varianza total explicada

	Autovalores iniciales			mas de las saturaciones al cuadra de la extracción			ma de las saturaciones al cuadra de la rotación		
Component	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2,953	26,842	26,842	2,953	26,842	26,842	2,720	24,725	24,725
2	2,164	19,676	46,518	2,164	19,676	46,518	1,742	15,839	40,564
3	1,465	13,317	59,835	1,465	13,317	59,835	1,722	15,657	56,222
4	1,061	9,648	69,484	1,061	9,648	69,484	1,459	13,262	69,484
5	,934	8,491	77,975						
6	,774	7,034	85,009						
7	,479	4,353	89,361						
8	,420	3,814	93,176						
9	,318	2,892	96,067						
10	,261	2,373	98,440						
11	,172	1,560	100,000						

Método de extracción: Análisis de Componentes principales.

La Tabla 9.2 nos aporta información detallada sobre los cuatro factores principales que perfilan los estilos de enseñanza de los correctores de nuestro estudio. Teóricamente, cabría esperar una distribución similar de respuestas en cada uno de ellos. En la Tabla 9.1 se observa que cada uno de estos componentes o factores explica un valor de la varianza distinto:

Tabla 9.1. Matriz de componentes rotados

Matriz de componentes rotados [a]

	Componente			
	1	2	3	4
A1. Prefiero actividades como la conversación a aquellas que conlleven análisis y aplicación de reglas.	,123	,785	,195	-8,57E-03
A.2. Considero fundamental la corrección en el uso de la lengua y me molesta mucho cometer errores.	,197	,179	,151	,827
A3. Normalmente escribo los elementos clave en la pizarra y ofrezco a los estudiantes ejemplos gráficos que les ayuden a entender los nuevos conceptos.	-,139	-,215	,873	6,790E-02
A4. Hago poco uso de la pizarra y de los medios audiovisuales; prefiero la conferencia, el debate y el discurso oral para impartir los conocimientos	-,116	-,250	-,734	-3,74E-03
A5. Me gusta desplazarme mientras explico ofreciendo actividades variadas en la clase	,160	,413	1,031E-02	-,707
A6. Favorezco la planificación, el orden y el trabajo bien hecho; las sorpresas no me agradan	,520	,288	,140	-,318
A7. Prefiero trabajar en compañía más que solo y a menudo me resulta difícil concentrarme durante un largo período de tiempo en una misma tarea.	-6,08E-02	,839	-,249	-8,15E-02
A8. Normalmente intento buscar relaciones entre hechos y datos para llegar a tener una visión global de las cosas.	,429	-,156	,514	,389
A9. Trabajo de forma sitemática el planteamiento de mis lecciones y busco que todo salga de acuerdo al plan establecido.	,871	3,153E-02	1,713E-02	6,915E-02
A10. Me agrada la sistematicidad en la preparación y presentación de las lecciones y raramente me suelo desviar de los objetivos específicos de cada lección.	,869	4,641E-02	9,669E-02	3,141E-02
A11. Prefiero la flexibilidad en el trabajo y me molesta trabajar con fechas que tengan un fin de plazo determinado.	-,797	7,847E-03	6,744E-02	-7,46E-02

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.
a. La rotación ha convergido en 5 iteraciones.

Seguidamente, pasamos a comentar los distintos componentes de forma individual.

Componente 1. Este primer componente explica el 26,842 de la varianza

Los datos del primer y principal grupo de componentes indican una relación positiva entre los enunciados A6 (0,520), A9 (0,871) y A10 (0,869) representativos de los estilos de enseñanza introvertido, cerrado y concreto-secuencial. El rasgo común que comparten estos tres estilos es el deseo de planificación y sistematicidad en la preparación del trabajo y presentación de las lecciones.

Asimismo, los datos revelan una relación negativa de los enunciados anteriores con el enunciado A11 (-0,797) representativo del estilo abierto. Este último estilo favorece la flexibilidad en el trabajo. Según Dreyer (1988), los profesores con un estilo de enseñanza abierto suelen ser personas espontáneas y adaptables a las que no les gusta trabajar bajo la presión de plazos prefijados. De ahí, su relación negativa con el trabajo sistemático y perfectamente planificado que defienden los estilos introvertido, cerrado y concreto-secuencial arriba citados (A6, A9 y A10).

Componente 2. Este segundo componente explica el 19,676 de la varianza

En este segundo grupo de componentes se establece una asociación positiva entre los enunciados A1 (0,785) y A7 (0,839) característicos de los estilos de enseñanza global y extrovertido. El estilo de enseñanza global favorece actividades de contenido personal y social como, por ejemplo, la conversación en el aula (Kinsella, 1995). Este argumento explica la relación del estilo de enseñanza global con el estilo de enseñanza extrovertido dado que las personas extrovertidas necesitan interactuar y trabajar en contacto con los demás.

Dreyer *et al.* (1996) afirman que debido a su proyección social y a su percepción de la gente, las personas con un estilo de aprendizaje global suelen acometer el aprendizaje de los aspectos comunicativos de una segunda lengua con éxito y eficacia.

Los profesores con un estilo de enseñanza global intentan promover la interacción comunicativa entre los estudiantes. Su relación positiva con el estilo de enseñanza extrovertido se explica porque los profesores con este último estilo de enseñanza tienden a requerir la participación activa de los estudiantes, principalmente en las tareas de producción oral. (ver Dreyer, 1998).

Componente 3. Este tercer componente explica el 13,317 de la varianza

En este grupo se relacionan los enunciados A3 (0,873) y A8 (0,514) ilustrativos de los estilos de enseñanza visual e intuitivo respectivamente. Los profesores con un estilo de enseñanza visual favorecen las anotaciones en la pizarra y la utilización de elementos gráficos en sus explicaciones. Este

planteamiento de la enseñanza parece asociarse con un estilo de enseñanza intuitivo que busca establecer relaciones entre hechos y datos para llegar a adquirir una visión global de las cosas (i.e. *big picture*).

Los datos también indican una asociación negativa de ambos estilos con el estilo de enseñanza auditivo (A4) (-0,734). Los profesores con un estilo de enseñanza auditivo suelen hacer poco uso de la pizarra y de los elementos gráficos. Así pues, la relación negativa entre el estilo visual y el estilo auditivo resulta explicable. Según nos aclaran Oxford *et al.* (1992), las personas con un estilo visual necesitan constantes estímulos visuales (i.e. transparencias, videos, etc.). De hecho, las clases teóricas o las discusiones en clase sin ningún tipo de refuerzo visual pueden llegar a ser muy confusas.

Componente 4. Este cuarto componente explica el 9,648 de la varianza

Finalmente, en este último grupo se establece una relación negativa entre los enunciados A2 (0,827) y A5 (-0,707) característicos de los estilos de enseñanza analítico y manipulativo respectivamente.

Los profesores con un estilo analítico favorecen actividades que conllevan el análisis y la atención a los detalles. A estos profesores les molesta mucho cometer errores ya que consideran fundamental la corrección en el uso de la lengua. Por el contrario, los profesores con un estilo de enseñanza manipulativo suelen ser personas activas que introducen constantemente actividades variadas en clase y que promueven

en mayor medida la fluidez y la comunicación de la lengua que no su uso correcto.

En resumen, los datos analizados nos permiten trazar a grandes rasgos los principales perfiles docentes de nuestros correctores:

1. *El profesor sistemático*. El factor principal del primer grupo nos define el perfil del profesor sistemático. Como dijimos, estos profesores suelen plantear de forma minuciosa las lecciones y estructurarlas de acuerdo con un plan establecido. Enfatizan principalmente el orden y la claridad en todos los aspectos relacionados con la enseñanza y el aprendizaje de la lengua (ver Dreyer, 1998).
2. *El profesor comunicativo*. El segundo factor de nuestro estudio perfila al profesor comunicativo. Estos profesores promueven la interacción social y la comunicación en el aula. Según Kinsella (1995), el material utilizado suele ser de contenido social e incluir sentido del humor.
3. *El profesor que favorece el material visual*. Este tipo de profesor también es relevante en la interpretación de los resultados de nuestro estudio. Generalmente, estos profesores ofrecen a sus estudiantes material con un gran refuerzo visual (i.e. mapas, pizarra, pósters, ordenadores, etc.).

4. *El profesor dominante*. El último factor nos define al profesor dominante. Dichos profesores desean controlar y dirigir todo el proceso de enseñanza-aprendizaje. Consideran fundamental la corrección en el uso de la lengua y favorecen los ejercicios mecánicos dirigidos a prevenir cualquier tipo de error lingüístico.

Cabría esperar que la visión teórica que nos aportan estos datos fuese consecuente con la posterior actuación práctica de los correctores, que se abordará en el capítulo siguiente (capítulo 10). No obstante, como se podrá comprobar, el perfil del corrector dominante que ocupa un cuarto lugar en la investigación de la parte teórica pasa a ocupar un primer puesto en el estudio de la parte práctica. Es decir, la visión práctica de los resultados sugiere que nuestro perfil de corrector es el de un profesor principalmente sistemático y dominante.

9.3. Estudio de los ítems que definen el perfil del profesorado (sección B)

Siguiendo con el esquema establecido en la II Parte del cuestionario, y una vez analizados los principales perfiles docentes que presentan los correctores de nuestro estudio, procedemos a investigar algunos rasgos de su personalidad (sección B).

Empezaremos nuestro análisis examinando los resultados de la tabla de frecuencias (Tabla 9.3).

Conviene recordar en este punto que las respuestas de los correctores se midieron para cada uno de los enunciados de acuerdo con la escala de Likert de 5 puntos: 1 (totalmente en desacuerdo) y 5 (totalmente de acuerdo). Consecuentemente, y para el propósito de este estudio, las respuestas se han clasificado de acuerdo con el número de ellas obtenidas expresando desacuerdo (1 ó 2), acuerdo (4 ó 5) e incluyéndose una categoría neutra (3).

Tabla 9.3. Tabla de Frecuencias: personalidad de los correctores

	Desacuerdo (valores 1-2)	**Neutro** (valor 3)	**Acuerdo** (valores 4-5)	**TOTAL**
B1. Soy una persona optimista	1	8	23	32
B2. Me gusta mi trabajo y me parece estimulante	2	6	24	32
B3. Me siento confiado y seguro en el trabajo	1	3	28	32
B4. Tengo una gran autoestima	4	11	17	32
B5. Me siento realizado en mi trabajo y este me resulta enriquecedor	6	10	16	32

Como se vemos en la tabla, el enunciado B3 es el que registra un mayor grado de acuerdo entre los correctores (n = 28). Estos resultados indican que la mayoría de los profesores se sienten confiados y seguros en el trabajo. Se consideran, además, personas optimistas (B1; n = 23) a las que les gusta su trabajo y éste les parece estimulante.

A pesar de ello, algunos profesores afirman no sentirse realizados en el trabajo (B5; n = 6). De hecho, los ítems B4 y B5 son los que registran un mayor grado de desacuerdo entre los correctores (n = 4 y n = 6 respectivamente). Dichos enunciados son los que, además, muestran un mayor número de respuestas neutras (n = 11 y n = 10 respectivamente). Esto es, se tiende hacia la centralidad, a pesar de que el grado de acuerdo supera al resto de los valores.

La disensión que se observa entre los enunciados B2 y B5 probablemente se explique a través de la hipótesis de proyección externa vs. proyección interna de los sujetos. El contenido del enunciado B2 parece proyectarse hacia el exterior. Los sujetos manifiestan que les gusta su trabajo porque éste les confiere prestigio social y se sienten a gusto. Por su parte, el enunciado B5 tiende a proyectarse hacia el interior. Es decir, los sujetos se autoevalúan y por ello se muestran, quizás, más críticos y exigentes en sus opiniones.

Estos datos nos exigen buscar más información. Por tanto, recurrimos al siguiente subepígrafe en el que estudiamos las correlaciones entre los diversos enunciados.

9.3.1. Estudio de las correlaciones

La Tabla 9.4 nos muestra las correlaciones entre los enunciados que definen la personalidad de nuestros correctores.

Tabla 9.4. Correlaciones del perfil del profesorado (i.e. personalidad)

Correlaciones

		B1	B2	B3	B4	B5
B1. Soy una persona optimista.	Correlación de Pearson	1,000	,239	,034	,172	,128
	Sig. (bilateral)	,	,188	,855	,345	,484
	N	32	32	32	32	32
B2. Me gusta mi trabajo y me parece estimulante.	Correlación de Pearson	,239	1,000	,716**	,517**	,754**
	Sig. (bilateral)	,188	,	,000	,002	,000
	N	32	32	32	32	32
B3. Me siento confiado y seguro en el trabajo	Correlación de Pearson	,034	,716**	1,000	,539**	,493**
	Sig. (bilateral)	,855	,000	,	,001	,004
	N	32	32	32	32	32
B4. Tengo una gran autoestima.	Correlación de Pearson	,172	,517**	,539**	1,000	,274
	Sig. (bilateral)	,345	,002	,001	,	,129
	N	32	32	32	32	32
B5. Me siento realizado en mi trabajo y este me resulta enriquecedor.	Correlación de Pearson	,128	,754**	,493**	,274	1,000
	Sig. (bilateral)	,484	,000	,004	,129	,
	N	32	32	32	32	32

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

En esta tabla de correlaciones se aprecia:

1. El mayor nivel de correlación se da en los enunciados B3 y B2. Los datos sugieren que la seguridad laboral y el hecho de disponer de un trabajo estimulante constituyen los principales componentes de la personalidad de los correctores.
2. Las correlaciones entre los enunciados B3, B4 y B5 son altas. Estos datos nos permiten inferir que el nivel de autoestima y de realización en el trabajo correlaciona bien con la seguridad laboral de los correctores.

3. Sorprendentemente, el enunciado B1 no muestra correlación alguna con los demás enunciados. Este dato nos lleva a pensar que el optimismo no constituye un elemento decisivo de la personalidad del corrector.

9.4. Autoevaluación del corrector: perfil del corrector (sección C)

La tercera dimensión planteada en la II parte del cuestionario era la autoevaluación de los correctores en su faceta evaluadora. El perfil evaluador de los correctores se examinará a través de un análisis de frecuencias. Debido a la importancia que los resultados de este apartado tienen para nuestro trabajo, nos detendremos de forma extensa en el comentario de cada uno de los enunciados.

Recordemos que los correctores tenían que evaluar cada uno de los enunciados en la escala de Likert de 5 puntos. Cabe destacar, sin embargo, que los valores 1 y 5 de la escala representan cualidades *positivas* o *negativas* de forma aleatoria. Es decir, los valores 1 ó 5 de la escala no se atribuyen de forma sistemática a cualidades *positivas* o *negativas* ya que la cualidad evaluadora representa los dos polos de un continuo en tres de los cinco casos (enunciados C2, C3 y C4).

En estos casos, y a diferencia de lo que ocurre en una escala ordinal, las respuestas se miden en función de la proximidad a cada polo u extremo (p. e. condescendiente _ duro). La adopción de esta medida tenía como objetivo evitar que los correctores asociaran las cualidades *positivas* o

negativas a un valor determinado de la escala que pudiera condicionar su respuesta.

El diseño y naturaleza de estas escalas nos obliga a comentarlas de forma individual.

Los resultados de la primera escala se recogen en Tabla 9.5a.

Tabla 9.5a.Dominio de la lengua inglesa

	1 (escaso)	**2**	**3**	**4**	**5 (muy bueno)**	**TOTAL**
C.1. Mi dominio de la lengua inglesa es...	0	0	1	16	15	32

Los datos indican que la totalidad de los correctores asegura poseer un dominio de la lengua inglesa bueno o muy bueno. Tan sólo se registra una única respuesta neutra. Asimismo, cabe resaltar que no se observa ningún caso en el que los correctores afirmen tener un dominio de la lengua escaso. Más bien, sorprende la seguridad con la que los correctores evalúan su dominio de la segunda lengua (n = 15, valor 5 de la escala). Así pues, se puede afirmar que la calificación personal de los correctores se rige por un criterio que podríamos calificar como de autocomplaciente.

El estudio de la primera faceta evaluadora comienza en la Tabla 9.4b.

Tabla 9.5b. Grado de exigencia en las evaluaciones

	1 (condescendiente)	**2**	**3**	**4**	**5 (duro / exigente)**	**TOTAL**
C.2. Me considero un corrector...	1	3	17	11	0	32

Los datos indican que la mayoría de los correctores tienden hacia la centralidad (n = 17; valor 3 de la escala) y la admisión de cierto grado de exigencia en sus evaluaciones (n = 11; valor 4 de la escala). No obstante, ninguno de ellos alcanza el valor 5 de la escala que se identifica con la figura del corrector exigente o duro. Tan sólo tres correctores se declaran condescendientes en sus valoraciones del rendimiento de los candidatos. Uno de ellos se declara extremadamente condescendiente (valor 1 de la escala).

A la luz de estos datos se puede afirmar que el mostrarse exigente o estricto en las evaluaciones se considera una cualidad positiva. Por el contrario, la condescendencia se considera una cualidad o atributo negativo.

La siguiente Tabla (Tabla 9.5c) analiza los conceptos de flexibilidad e inflexibilidad en las evaluaciones.

Tabla 9.5c. Grado de flexibilidad en las evaluaciones

	1 (inflexible)	**2**	**3**	**4**	**5 (flexible)**	**TOTAL**
C.3. Me considero un corrector...	0	3	7	19	3	32

En esta cualidad, el punto central no lo ocupan las equidistancias como ocurría con la cualidad anterior. Si trazáramos una hipotética curva de distribución, los resultados mostrarían un sesgo marcadamente negativo.

Los datos demuestran que la mayoría de los correctores alcanza el valor 4 de la escala (n = 19) que se identifica con la cualidad de la flexibilidad. Cabe subrayar que tres correctores juzgan su actuación con el valor 5, que representa un grado de flexibilidad extrema. No obstante, también hay correctores que se califican como inflexibles (n = 3) aunque en ningún caso se alcanza el valor extremo de la escala (i.e. valor 1). Estos datos sugieren que la flexibilidad se juzga como un atributo positivo y la inflexibilidad como una cualidad negativa.

Nuestra siguiente Tabla (Tabla 9.5d) examina el grado de consecuencia que presentan los correctores en sus actuaciones.

Tabla 9.5d. Grado de consecuencia en las evaluaciones

	1 (consecuente)	**2**	**3**	**4**	**5 (inconsecuente)**	**TOTAL**
C.4. Me considero un corrector...	9	9	5	2	1	32

Los resultados tienden esta vez a polarizarse en torno a los valores 1 (n = 9) y 2 (n = 9) de la escala que representan la cualidad de ser consecuentes. En este caso, el polo positivo lo constituye el punto de partida. De ahí que, si trazáramos una hipotética curva de distribución los resultados presentarían un sesgo positivo.

A pesar de ello, algunos correctores afirman ser inconsecuentes en sus valoraciones (valor 4; n = 2). Sorprende el caso de uno de ellos que alcanza el valor extremo de la inconsecuencia (i.e. valor 5 de la escala).

En los tres últimos casos analizados hasta ahora (C2, C3, C4), la actuación de los correctores se valora ante terceros. Es decir, se juzga cómo los correctores ejercen y practican su tarea.

La última cualidad (C5) que se analiza, al igual que ocurría en el primer caso (C1), mide cómo los correctores se juzgan a sí mismos y valoran su preparación. Así pues, se prevé un sesgo negativo en los resultados.

La Tabla 9.5e nos muestra los resultados obtenidos.

Tabla 5e. Calificación de los correctores

	1 (no cualificado)	**2**	**3**	**4**	**5 (experto)**	**TOTAL**
C.5. Me considero un corrector...	0	0	3	20	9	32

Llama la atención el alto grado de acuerdo entre los correctores en cuanto a la valoración experta que hacen de su actuación. Los datos indican que los valores extremos de la escala son muy altos (n = 9; valor 5 de la escala y n = 20; valor 4 de la escala).

Es interesante destacar que las Tabla 9.5e y 9.5a son las únicas que no registran ninguna respuesta en los dos valores extremos de la escala (i.e. valores 1 y 2). Los correctores se muestran confiados y seguros sobre la idoneidad de su perfil en su faceta evaluadora.

En resumen, la actuación de los correctores se manifiesta en dos vertientes:

> 1ª vertiente: los correctores se califican a sí mismos y juzgan su preparación (C1 y C5). En estos casos, su valoración es extremadamente positiva.
>
> 2ª vertiente: los correctores valoran su actuación ante los demás (C2, C3 y C4). En estos casos van en busca de los extremos de la escala que tengan un grado considerable de aceptabilidad social.

Así pues, podemos concluir que el perfil del corrector que participa en nuestro trabajo de acuerdo con sus valoraciones personales es el siguiente:

1. Es un profesor con un dominio de la lengua inglesa muy bueno,
2. Es un profesor mayoritariamente exigente
3. También es flexible en sus valoraciones
4. Admite ser mayoritariamente consecuente
5. Y, por último, se considera un corrector experto.

9.5. Identificación de los aspectos problemáticos (sección D)

Hasta aquí, nuestra aproximación al perfil del corrector se ha basado en su valoración personal. Nos falta definir ahora qué aspectos concretos de la evaluación de los ensayos centran su atención. De ahí, que nuestro siguiente paso lo constituya el estudio de los principales aspectos

problemáticos que destacan los correctores en las producciones escritas de los estudiantes.

Siguiendo el esquema de Johns (1991), los correctores se pronuncian sobre aquellos aspectos de los ensayos que más les preocupa. Recordemos que los correctores tenían que elegir tres problemas básicos de entre los seis que se les ofrecía y ordenarlos según criterios de relevancia. Las Tablas de frecuencias 9.6a, 9.6b y 9.6c que se presentan a continuación recogen las opciones primera, segunda y tercera respectivamente.

La Tabla 9.6a nos muestra los datos obtenidos en la valoración del problema prioritario.

Tabla 9.6a. Principal aspecto problemático (D1)

Principal aspecto problemático	**TOTAL**
D.1. Carencia de conocimientos generales previos	13
D.2. Identificación de los objetivos del texto	5
D.3. Falta de planificación	7
D.4. Problemas para establecer inferencias	2
D.5. Carencia de vocabulario básico o esencial	5
D.6. Subjetividad	0

Como puede apreciarse, los datos indican que la *carencia de conocimientos generales previos* (n = 13) es el principal aspecto problemático que observan los correctores en la evaluación de los ensayos.

Estos resultados sugieren que el aspecto del *contenido* parece primarse por encima de los aspectos discursivos y léxicos de la producción escrita.

A continuación, pasamos a analizar la segunda dificultad principal que acusan los correctores en las producciones escritas de los estudiantes (Tabla 9.6b). En esta tabla vemos que la *carencia de vocabulario básico o esencial* constituye, esta vez, la preocupación principal de los correctores (n = 10).

Tabla 9.6b. Segundo aspecto problemático (D2)

Segundo aspecto problemático	**TOTAL**
D.1. Carencia de conocimientos generales previos	5
D.2. Identificación de los objetivos del texto	2
D.3. Falta de planificación	7
D.4. Problemas para establecer inferencias	6
D.5. Carencia de vocabulario básico o esencial	10
D.6. Subjetividad	2

Así pues, los resultados obtenidos en las dos primeras tablas (Tablas 9.6a y 9.6b) indican que el interés fundamental de los correctores se centra en el *contenido* del ensayo en primer lugar y en el *vocabulario* o aspectos léxicos del ensayo en segundo lugar.

Por último, la Tabla 9.6c nos muestra el tercer problema principal que denuncian los correctores en los ensayos de los estudiantes.

Tabla 9.5c. Tercer aspecto problemático (D3)

Tercer aspecto problemático	TOTAL
D.1. Carencia de conocimientos generales previos	6
D.2. Identificación de los objetivos del texto	5
D.3. Falta de planificación	7
D.4. Problemas para establecer inferencias	6
D.5. Carencia de vocabulario básico o esencial	7
D.6. Subjetividad	1

En esta tabla (Tabla 9.5c), la distribución de las respuestas para cada uno de los enunciados es similar. No se observan aspectos marcados. Como se ve, esta vez son dos los problemas que se identifican como prioritarios: la *falta de planificación* (n = 7) y, de nuevo, la *carencia de vocabulario básico o esencial* (n = 7). También es interesante observar que la *carencia de conocimientos generales previos* vuelve a ser representativa (n = 6). La repetición de estos dos últimos aspectos en las valoraciones de los correctores señala su preocupación por los aspecto léxicos y semánticos de los ensayos.

Cabe mencionar que los criterios establecidos por Johns (1991) carecen de aspectos evaluativos relacionados con la *forma* del texto escrito. No obstante, es interesante observar que el interés principal de los correctores se centra en los aspectos de *contenido* y de *vocabulario*. Los aspectos discursivos como son la *falta de planificación* también se

consideran relevantes aunque estos últimos parecen despertar un menor interés en los correctores.

9.6. Evaluación de los errores contenidos en frases descontextualizadas

A pesar de que el análisis de los resultados de esta última parte del cuestionario se realizará en el siguiente capítulo (capítulo 10), realizaremos ahora una breve aproximación a los datos.

En este apartado se pretende contrastar la teoría que hemos visto hasta ahora con la práctica. Las frases individuales y descontextualizadas que los correctores evalúan nos permitirán observar la relevancia de los criterios establecidos por Johns (1991) a la hora de identificar los aspectos problemáticos de los ensayos. Dado que estos criterios se centran en aspectos relacionados principalmente con el contenido de los ensayos, consideramos que la evaluación de los errores nos permitirá observar la aplicación de otros criterios de carácter formal que centran el interés de gran parte de los correctores en el ámbito de la evaluación de segundas lenguas

Como se recordará, cada una de las 14 frases que los correctores debían evaluar contenía una muestra de error relacionada con una de las siete categorías analíticas desarrolladas en este estudio (ver Apéndice 1, III parte del cuestionario). En total, se incluyeron dos muestras de error representativas de cada una de las categorías siguientes: *contenido* (enunciados E1 y E10), *organización* (enunciados E6 y E9), *gramática* (enunciados E3 y E9), *vocabulario* (enunciados E5 y E8), *registro*

(enunciados E7 y E11), *mecánica* (enunciados E2 y E4) y *presentación* (enunciados E13 y E14). Como se observa, a diferencia de los criterios establecidos por Johns (1991), las categorías analíticas citadas incluyen aspectos relacionados tanto con la *forma* como con el *contenido* de los ensayos.

La gravedad o seriedad de los errores se debían juzgar de acuerdo con la escala de Likert de 5 puntos: 1 (sin importancia) y 5 (muy importante). Los errores se señalaron en letra cursiva para facilitar su localización (ver Apéndice 1, III parte del cuestionario).

Una primera aproximación a los datos nos lleva a analizar los estadísticos descriptivos. La Tabla 9.7 nos muestra las medias y la desviación típica de cada una de las muestras de error asociadas con las categorías analíticas.

En la tabla, las puntuaciones medias se han dispuesto en orden ascendente para facilitar la lectura de los datos. Los errores se presentan en orden de gravedad, de menos grave a más grave según la media global obtenida. Conviene recordar en este punto que las puntuaciones medias más altas indican una penalización mayor del error de acuerdo con la escala utilizada.

Tabla 9.7. Gravedad de los errores contenidos en frases aisladas

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
E2. Thank you, I will accept your INVITACION	32	1	5	2,44	1,16
E1. I want to go to London to see THE STATUE OF LIBERTY	32	1	5	2,66	1,33
E4. I like speaking eNGLISH very much	32	1	5	2,91	1,28
E7. I have been to London and I have visited a few shops. Perhaps you think I'm stupid but don't worry about that (style)	32	1	5	2,94	1,22
E11. A: Are you sure it would not be a problem for your cousin to put me up? B: No, silly (style)	32	1	5	3,06	1,19
E14. Sample of student's handwriting	31	1	5	3,32	1,17
E10. I don' t like London and I think it's a horrible island (content)	32	1	5	3,34	1,15
E6. I would visit all the museums, FURTHERMORE Big Ben	32	2	5	3,38	,87
E5. I find London very interesting but the food is WRONG	31	2	5	3,42	,89
E8. One of my main HOPENESS is to visit London	32	2	5	3,44	,72
E9. I like London. HOWEVER, my parents and I will visit it next year	32	1	5	3,59	,95
E13. Sample of student's handwriting	31	1	5	3,61	1,09
E12. I would visit the most IMPORTANTS places in London	32	2	5	4,13	,87
E3. If I went to London I would BOUGHT a lot of things	32	3	5	4,25	,67
N válido (según lista)	30				

Los datos nos indican que los elementos del ensayo que se penalizan de forma más severa son: la *gramática* ($\bar{\times}$ = 4,13 (E12) y $\bar{\times}$ = 4,25 (E3)), la *organización* ($\bar{\times}$ = 3,38 (E6) y $\bar{\times}$ = 3,59 (E9)) y el *vocabulario* ($\bar{\times}$ = 3,42 (E5) y $\bar{\times}$ = 3,44 (E8)). De ahí, se deduce que la atención principal de los correctores se centra en los errores gramaticales, discursivos y léxicos de los ensayos, en contraposición al perfil que intentan dar como profesores comunicativos (ver subepígrafe 9.2.1.)

Asimismo, es interesante observar que las menores desviaciones típicas se registran de nuevo en las categorías analíticas anteriormente citadas, esto es: la *gramática* (DT = 0,87 (E12) y DT = 0,672 (E3)), el *vocabulario* (DT = 0,89 (E5) y DT = 0,72 (E8)) y la *organización* (DT = 0,95 (E9) y DT = 0,87 (E6)). Estos datos sugieren que hay poca dispersión en las puntuaciones alrededor de la media en la evaluación de estos errores. En otras palabras, los correctores se muestran más consistentes si bien más estrictos, en la evaluación de los errores gramaticales, discursivos y léxicos del ensayo.

Por otro lado, llama la atención la importancia que se le concede a los errores de *presentación* ($\bar{\times}$ = 3,61 (E13) y $\bar{\times}$ = 3,32 (E14)) que son juzgados globalmente de forma más estricta que los errores de *contenido* ($\bar{\times}$ = 2,66 (E1) y $\bar{\times}$ = 3,34 (E10)). Los errores de *ortografía* y de *mecánica* parecen ser los que preocupan a los correctores en menor medida.

En resumen, los datos sugieren que el interés principal de los correctores se centra en los aspectos formales, léxicos y discursivos del ensayo. Por el contrario, el *contenido* ocupa un lugar menos relevante en la

evaluación de los ensayos. Estos resultados contradicen los resultados obtenidos en la identificación de los principales aspectos problemáticos de los ensayos (seccion D, II parte del cuestionario (epígrafe 9.5)) ya que en opinión de los correctores la *carencia de conocimientos generales previos* constituía su principal preocupación. Así pues, en teoría, los correctores afirman primar y enfatizar el *contenido* de los ensayos. En la práctica, en cambio, los correctores le conceden un peso mayor a los aspectos relacionados con la *forma* del ensayo.

CAPÍTULO 10. ANÁLISIS DE MATERIALES: LOS ENSAYOS

10.1. Introducción

En este capítulo nos proponemos realizar un análisis tanto de la fiabilidad inter-corrector como de la fiabilidad intra-corrector en la evaluación de los ensayos de la prueba de Inglés de Selectividad.

La exposición del análisis de los resultados se hará en el siguiente orden. Empezaremos estudiando el comportamiento de los correctores en la evaluación holística y analítica de los ensayos a nivel general. A continuación, agruparemos a los correctores de acuerdo con las variables género y situación laboral con el fin de determinar la influencia que ejercen estas variables en las puntuaciones de los correctores. Seguidamente, profundizaremos en el estudio de las evaluaciones globales y analíticas específicas. Nos detendremos, posteriormente, en la evaluación de los errores hallados fuera y dentro del discurso. En este último caso, se intentará discriminar entre las evaluaciones de los errores contextualizados y descontextualizados. Pasaremos a examinar, a continuación, la influencia que ejercen los distintos elementos analíticos en la evaluación holística de los ensayos. Finalmente, acometemos el estudio de la fiabilidad intra-corrector.

Los resultados se analizaran a través de técnicas estadísticas diversas utilizando el método *Statistical Package for the Social Sciencies* 11.0.1 (SPSS 0.11). El protocolo establecido es el siguiente:

1. Estudio de las correlaciones (Pearson y Spearman, según se trate de variables cuantitativas o cualitativas) entre las evaluaciones holísticas y analíticas, primero a nivel global y luego individualmente. A nivel más concreto, se examinarán las correlaciones entre los ensayos según su nivel de dominio lingüístico.
2. Se llevará a cabo el análisis de la varianza (ANOVA) sobre los datos obtenidos. Los resultados se presentarán a través de análisis gráficos y descriptivos.
3. Se examinarán las pruebas de significación en las evaluaciones holísticas y en las analíticas consideradas globalmente y en sus distintos componentes analíticos. A continuación, se aplicarán estas pruebas de significación a la evaluación de los errores en frases individuales (i.e. descontextualizadas) y se contrastarán con los resultados obtenidos en la evaluación de los errores contenidos en el ensayo (i.e. errores contextualizados).
4. Identificaremos los componentes analíticos que determinan las puntuaciones holísticas a través de la recta de regresión.
5. Se analizará el peso o relevancia que tiene cada uno de los componentes analíticos en los comentarios, tanto positivos como negativos, que realizan los correctores sobre los distintos ensayos.

Estos datos se presentarán a través de tablas de frecuencias y gráficos (i.e. histogramas).

6. Finalmente, se llevará a cabo el estudio de la fiabilidad intra-corrector a través de descriptivos y diagramas de dispersión que nos permitirán identificar los casos extremos o marginales.

10.1.1. Evaluación holística (PRE y POST)

Nuestra aproximación a los datos tendrá un carácter progresivo. Comenzaremos analizando las evaluaciones de carácter general para luego centrarnos en las de carácter más concreto, i.e. primero las puntuaciones globales u holísticas y, a continuación, las analíticas.

Queremos empezar comentando los estadísticos descriptivos de la puntuación holística de los diez ensayos calificados por el grupo de correctores en la primera y en la segunda parte del estudio (PRE y POST) (Tabla 10.1). La diferencia entre las medias de esta evaluación holística alcanza casi los tres puntos (2,9) siendo superior la media de los resultados globales en el PRE que en el POST. Las diferencias, no obstante, aplicando el estadístico de medias relacionadas, no llegan a ser significativas: $t = 1,83$ y $p = 0,075$.

Sin embargo, el hecho de que estos valores estén en el umbral de la significación ($p = \leq 0,05$), nos indica que los correctores, en general, se mostraron más parcos en la evaluación de los ensayos en la segunda parte del estudio (POST). La reducción de las puntuaciones, que puede llevar a

interpretaciones de mayor severidad, nos lleva a subrayar como dato positivo el mayor nivel de homogeneidad. No sólo la desviación típica, que nos explica la dispersión de las puntuaciones, pasó de 1,0803 a 0,6720 sino también la amplitud del rango se redujo y pasó de 4,1 a 2,8 puntos en la segunda evaluación (POST).

Tabla 10.1. Estadísticos descriptivos de las evaluaciones holísticas en el PRE y en el POST

Estadísticos descriptivos

	N	Rango	Mínimo	Máximo	Media	Desv. típ.
HG1	32	4,10	2,90	7,00	4,8563	1,0803
HG2	32	2,80	2,80	5,60	4,5594	,6720
N válido (según lista)	32					

HG1 y HG2 se refieren a las puntuaciones holísticas globales en el PRE y en el POST respectivamente.

Esta primera tabla nos muestra las puntuaciones medias de todos los correctores en todos los ensayos. La información que nos proporciona es más bien limitada. Se requiere, por tanto, examinar con más detalle las puntuaciones de cada corrector en cada uno de los ensayos (Tablas 10.2a y 10.2b).

Tabla 10.2a. Evaluaciones holísticas (PRE)

Raters	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10
R1	3	4	5	5	7	7	6	9	9	9
R2	3	3	2	4	5	6	4	5	7	8
R3	2	6	4	6	6	7	6	9	9	10
R4	4	8	5	6	6	5	3	8	7	9
R5	1	2	2	3	3	4	2	4	5	6
R6	2	5	3	6	5	7	4	7	9	9
R7	1	1	1	1	5	5	4	6	6	5
R8	2	2	3	3	3	4	4	5	5	5
R9	2	3	3	3	4	4	5	7	8	8
R10	3	3	4	3	5	5	4	6	7	6
R11	1	3	2	3	5	5	4	4	6	5
R12	2	4	3	3	4	4	3	6	7	6
R10	2	3	3	2	2	3	2	3	4	5
R14	2	6	3	2	4	6	6	7	8	7
R15	1	3	3	3	4	4	4	7	8	8
R16	2	4	3	2	5	6	5	7	8	9
R17	3	2	2	4	4	5	4	6	6	7
R18	3	4	3	3	6	8	4	8	9	10
R19	3	4	4	5	5	6	4	7	8	8
R20	6	7	6	5	4	5	6	7	8	9
R21	4	6	3	5	7	7	5	8	8	9
R22	1	3	2	3	5	4	2	7	6	8
R23	2	3	2	2	3	4	3	3	5	5
R24	1	2	1	3	3	4	3	6	8	9
R25	3	3	2	3	3	4	3	4	5	8
R26	2	2	3	2	6	6	6	6	7	8
R27	4	6	5	5	8	9	4	9	10	10
R28	2	3	2	3	2	5	4	7	7	8
R29	3	3	2	2	5	7	7	8	7	8
R30	2	2	3	6	2	7	3	8	9	9
R31	3	4	4	3	6	7	5	9	8	9
R32	4	5	4	4	4	6	4	8	7	9

La abreviatura hn hace referencia a los diferentes ensayos de la primera ocasión (PRE)

Tabla 2b. Evaluaciones holísticas (POST)

Raters	iih1	iih2	iih3	iih4	iih5	iih6	iih7	iih8	iih9	iih10
R1	2	3	3	3	5	4	4	6	7	9
R2	2	3	1	3	4	7	3	5	6	8
R3	2	5	2	4	4	6	3	7	4	9
R4	3	4	5	3	4	5	5	8	8	9
R5	1	3	1	4	2	4	3	5	6	6
R6	3	4	3	4	3	7	2	5	8	9
R7	3	5	3	6	5	6	4	7	8	8
R8	2	3	1	3	3	5	3	7	6	7
R9	1	2	3	3	2	5	3	6	8	8
R10	3	3	4	3	4	5	4	5	5	6
R11	1	4	2	4	3	5	4	5	6	9
R12	2	3	2	2	3	3	3	5	6	6
R10	2	2	2	3	2	3	3	3	4	4
R14	2	5	2	3	4	5	3	8	7	6
R15	3	5	3	4	4	8	4	6	8	8
R16	2	5	3	3	5	7	4	7	8	9
R17	3	4	3	4	2	4	3	6	8	9
R18	1	4	1	2	4	5	2	4	9	9
R19	3	4	4	3	4	4	3	6	8	7
R20	4	5	3	4	5	5	3	5	6	7
R21	2	5	3	4	6	6	3	7	6	7
R22	2	4	3	4	5	7	4	6	8	8
R23	1	3	1	1	3	6	2	4	7	8
R24	1	3	2	2	2	5	2	5	8	9
R25	2	3	3	3	5	5	3	4	6	7
R26	2	4	3	3	6	6	4	8	7	9
R27	2	3	2	3	5	6	3	8	9	10
R28	2	4	3	3	4	6	3	7	6	7
R29	4	4	4	4	5	6	6	7	7	7
R30	2	3	3	4	4	5	3	7	8	9
R31	4	3	3	3	6	5	4	8	6	9
R32	4	6	2	3	6	6	5	7	8	9

La abreviatura iihn hace referencia a los diferentes ensayos en la segunda ocasión (POST)

Si se observan detenidamente los distintos valores de estas tablas se comprueban algunas fluctuaciones de cierta relevancia tanto en el ámbito de la fiabilidad inter-corrector como en el de la fiabilidad intra-corrector.

Estas tablas nos suministran una información particularizada que no se encuentra en la Tabla 10.1, cuyo objetivo es darnos una perspectiva general de los datos obtenidos. En dicha tabla, el interés se centra en valorar el comportamiento del grupo de correctores en cuanto tal. Al tomar como referente el grupo y no el individuo las medias tienden a neutralizar la dispersión de las puntuaciones que se observa en las tablas (10.2a y 10.2b).

Ambas tablas traducen el comportamiento de cada corrector en cada uno de los ensayos, así como las puntuaciones que se le otorgan a cada ensayo tanto en la primera ocasión como en la segunda (PRE y POST).

Dado que es difícil extraer conclusiones si no se trazan parámetros específicos para estudiar estos datos, hemos optado por centrar nuestra atención en los ensayos: 1(h1), 5 (h5) y 10 (h10) del PRE. Nuestra elección se ha basado en su carácter representativo de los niveles de dominio de la lengua bajo, medio y alto respectivamente. Dicha categorización se ha inferido a partir de las medias de las puntuaciones holísticas asignadas en el PRE: h1 ($\bar{\times}$ = 2,47); h5 ($\bar{\times}$ = 4,56); h10 ($\bar{\times}$ = 7,78).

Si tomamos como punto de referencia el rango de las puntuaciones, se observa que si bien es semejante en todos los casos: 5, 6 y 5 puntos (en los ensayos h1, h5 y h10 respectivamente), su amplitud es considerable, ya que la puntuación mínima es el 1 y la máxima es el 10.

Cabe notar, también, la frecuencia de los valores mínimos y máximos en estos ensayos. En los casos en los que el dominio lingüístico es bajo (h1) o alto (h10), la frecuencia de los valores mínimos es superior a la que se recoge en el ensayo con dominio lingüístico medio (h5). Por el contrario, las frecuencias de los valores máximos sólo se ven incrementados en el ensayo de nivel superior (h10), como se observa en la siguiente tabla que se ha extraído de la tabla 10.2a para facilitar la lectura de estos estadísticos:

Tabla 10.3 Frecuencia de las puntuaciones mínima y máximas de los ensayos h1, h5 y h10

Puntuaciones	Mínimas / Frecuencias		Máximas / Frecuencia	
(h1)	1	5	6	1
(h5)	2	3	8	1
(h10)	5	5	10	3

Estos datos ponen de manifiesto que la valoración del rendimiento del candidato se ve afectada por el nivel del dominio lingüístico del candidato y por los criterios personales de cada corrector.

Hasta aquí se ha analizado la fiabilidad desde la perspectiva del inter-corrector. Para ello, nos hemos centrado en el rango y las frecuencias de mínimos y máximos del PRE (Tabla 10.1), ya que los valores del POST tienden a ser similares.

Una comparación de las tablas (10.2a y 10.2b) nos permite hacer un análisis desde la perspectiva de la fiabilidad intra-corrector. Si la fiabilidad

inter-corrector, que acabamos de comentar, queda en entredicho, no ocurre lo mismo con la fiabilidad intra-corrector. En el PRE y POST se observa una tendencia en cada corrector a mantener valores similares. De los treinta y dos correctores las calificaciones de veintiocho correctores oscilan entre +/- 1. Tan sólo las diferencias entre las calificaciones en ambas ocasiones pueden valorarse como menos consistentes en tres casos (correctores R8, R11 y R29) y, considerablemente menos consistente, en el corrector R23.

Las diferencias de estos últimos casos tienen su importancia, puesto que ya no se trata de la fiabilidad inter-corrector sino de la distinta lectura que hace un corrector según el momento de la evaluación. Como vemos, a pesar del alto grado de consecuencia que declaran poseer los correctores de nuestro estudio en la evaluación de los ensayos (ver capítulo 9, epígrafe 9.4), los datos revelan que las circunstancias personales o estados anímicos de algunos correctores pueden tener su incidencia en la valoración del candidato.

10.2. Fiabilidad inter-corrector

10.2.1. Estudio de las correlaciones

En los siguientes subepígrafes, analizaremos las correlaciones entre las evaluaciones holísticas (PRE y POST), las correlaciones entre las evaluaciones analíticas (PRE y POST), las correlaciones globales entre las evaluaciones holísticas y analíticas (PRE y POST) y las correlaciones entre los ensayos según el nivel de dominio lingüístico.

10.2.1.1. Correlaciones entre las evaluaciones holísticas (PRE y POST)

Hasta aquí se ha estudiado el comportamiento de los correctores recurriendo a estadísticos de rangos y frecuencias de mínimos y máximos. Procede, por tanto, que se aplique el estadístico de la correlación de Pearson para valorar la fiabilidad intra-corrector.

En la Tabla 10.4 se presenta la fiabilidad desde los estadísticos del nivel de consistencia y del acuerdo entre el PRE y el POST de las puntuaciones holísticas de los 10 ensayos.

Tabla 10. 4. Análisis de la fiabilidad: Puntuaciones holísticas (PRE y POST)

Nivel de consistencia Media de la correlación intra-clase= , 6556** 95,00% C.I.: Inferior= ,2944 Superior= ,8319 F= 2,9033 DF= (31, 31, 0) Sig.= , 0020 (Valor del test= ,0000)
Coeficiente de fiabilidad Nº de casos= 32,0 Nº de ítems= 2 α = ,6556
Acuerdo absoluto Media de la correlación intra-clase= , 6390** 95,00% C.I.: Inferior= ,2728 Superior= ,8223 F= 2,9033 DF= (31, 31, 0) Sig.= , 0020 (Valor del test= ,0000)

* Las abreviaturas HG1 y HG2 corresponden a las puntuaciones holísticas en el PRE y en el POST respectivamente.

El nivel de consistencia, según nos indica el coeficiente de correlación intra-clase (CCI) y el coeficiente de fiabilidad α de las puntuaciones holísticas es de 0,6556. Estadísticamente, se entiende que es muy significativo, pero su nivel de consistencia tan sólo se puede calificar de **regular - buena** siguiendo el criterio de Fleiss (1986)[1]. Por su parte, el nivel

[1]La interpretación de los valores obtenidos con el CCI es hasta cierto punto arbitraria, si bien existe un cierto consenso al aceptar el criterio de Fleiss (1986) que establece: valor del CCI = <0,40 (concordancia baja); 0,41-0,75 (concordancia regular-buena); 0,76-1,00 (concordancia muy buena).

de acuerdo absoluto (0,6390), se encuentra en unos valores muy semejantes a los de los niveles de consistencia.

Tras el análisis de la correlación holística global de todos los ensayos procede la presentación de las correlaciones de todos y cada uno de los ensayos entre el PRE y el POST (Tabla 10.5).

Tabla 10.5. Correlaciones entre las evaluaciones holísticas (PRE y POST)

Correlaciones

		2hol.1	2hol.2	2hol.3	2hol.4	2hol.5	2hol.6	2hol.7	2hol.8	2hol.9	2hol.10
hol.1	Correlación de Pearson	,516**	,183	,264	-,103	,441*	-,114	,137	,126	-,001	,061
	Sig. (bilateral)	,003	,316	,145	,575	,011	,536	,455	,492	,994	,740
	N	32	32	32	32	32	32	32	32	32	32
hol.2	Correlación de Pearson	,286	,381*	,226	-,098	,315	,057	,073	,265	-,008	,115
	Sig. (bilateral)	,112	,031	,213	,592	,079	,757	,692	,142	,963	,531
	N	32	32	32	32	32	32	32	32	32	32
hol.3	Correlación de Pearson	,407*	,119	,291	-,101	,348	-,151	,170	,259	-,041	,125
	Sig. (bilateral)	,021	,518	,106	,582	,051	,410	,352	,152	,824	,494
	N	32	32	32	32	32	32	32	32	32	32
hol.4	Correlación de Pearson	,181	,088	,235	,110	,064	-,008	-,137	,152	,073	,366*
	Sig. (bilateral)	,321	,630	,196	,548	,728	,967	,453	,407	,692	,039
	N	32	32	32	32	32	32	32	32	32	32
hol.5	Correlación de Pearson	,161	,241	,226	,119	,537**	,227	,267	,407*	,218	,486**
	Sig. (bilateral)	,378	,184	,213	,515	,002	,211	,140	,021	,231	,005
	N	32	32	32	32	32	32	32	32	32	32
hol.6	Correlación de Pearson	,207	,265	,080	,088	,527**	,233	,074	,460**	,295	,537**
	Sig. (bilateral)	,257	,143	,662	,630	,002	,200	,686	,008	,101	,002
	N	32	32	32	32	32	32	32	32	32	32
hol.7	Correlación de Pearson	,345	,392*	,234	,160	,461**	,200	,311	,476**	-,053	,201
	Sig. (bilateral)	,053	,026	,198	,383	,008	,273	,083	,006	,775	,269
	N	32	32	32	32	32	32	32	32	32	32
hol.8	Correlación de Pearson	,403*	,373*	,401*	,163	,534**	,227	,290	,650**	,355*	,501**
	Sig. (bilateral)	,022	,036	,023	,372	,002	,211	,107	,000	,046	,003
	N	32	32	32	32	32	32	32	32	32	32
hol.9	Correlación de Pearson	,121	,240	,200	,001	,314	,257	-,089	,399*	,402*	,545**
	Sig. (bilateral)	,510	,186	,271	,997	,080	,156	,629	,024	,022	,001
	N	32	32	32	32	32	32	32	32	32	32
hol.10	Correlación de Pearson	,196	,278	,250	-,046	,452**	,324	,014	,346	,367*	,574**
	Sig. (bilateral)	,283	,123	,167	,802	,009	,070	,940	,053	,039	,001
	N	32	32	32	32	32	32	32	32	32	32

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

*. La correlación es significante al nivel 0,05 (bilateral).

En una primera lectura se observa:

1. Ausencia de correlación en los ensayos (h4, h6, y h7), ensayos, que están a caballo entre dos categorías en el dominio de la lengua. El ensayo h4 ($\bar{\times}$ = 3,53) estaría próximo a los etiquetados como de nivel bajo y no lejos de los etiquetados como de nivel medio. Por el contrario, los ensayos h6 ($\bar{\times}$ = 5,50) y h7 ($\bar{\times}$ = 4,16) estarían próximos a los catalogados como de nivel medio y a una distancia considerable de los categorizados como de nivel alto.
1. Presencia de correlaciones significativas del ensayo etiquetado como de nivel medio (h5) del PRE con 2 ensayos de los que se consideran de mayor dominio lingüístico del POST. Se observa una situación similar si se toma el ensayo (h5) del POST. En este último caso, se da una correlación significativa o muy significativa con 2 de los ensayos que se consideran de mayor dominio lingüístico del PRE. Un nivel similar de correlaciones significativas se observa también en el ensayo h10.
2. Tendencia al incremento de correlaciones significativas entre las puntuaciones holísticas individuales, en los ensayos que muestran un dominio de la lengua alto, incluido el h5: h8 ($\bar{\times}$ =6,59), h9 ($\bar{\times}$ = 7,22) y h10 ($\bar{\times}$ = 7,78) tanto se trate del PRE como del POST.

Bajo nivel de correlación entre los ensayos que presentan un nivel de dominio lingüístico medio-bajo (h3, h4, h6 y h7). De hecho, estos ensayos son los únicos que no correlacionan consigo mismo.

La lectura de estos datos nos permiten inferir que los ensayos que no pertenecen a una de las categorías definidas y los que tipificaríamos como ensayos de nivel de dominio lingüístico medio-bajo son los que presentan mayor problema de correlación en los correctores (ver Connor-Linton 1994; Vaughan 1991).

10.2.1.2. Correlaciones entre las evaluaciones analíticas (PRE y POST)

En una segunda fase, tras haber estudiado la correlación PRE / POST en la evaluación holística, nos planteamos aplicar los mismos estadísticos a la evaluación analítica global y a cada uno de los componentes que configuran la evaluación analítica.

Empezamos examinando la consistencia y acuerdo en la evaluación analítica global.

Los valores que se observan en la Tabla 10.6, indican que el nivel de consistencia que nos facilita el coeficiente de correlación intra-clase y el coeficiente de fiabilidad α de las puntuaciones analíticas en las dos ocasiones (PRE y POST) es algo superior en las evaluaciones analíticas que en las holísticas ya que se alcanza el 0,6967. Sin embargo, el nivel de acuerdo absoluto entre los correctores en las puntuaciones analíticas es inferior al que se encuentra en las puntuaciones holísticas (0,5993).

Tabla 10.6. Análisis de la fiabilidad: puntuaciones analíticas (PRE y POST)

Nivel de consistencia Media de la correlación intra-clase= , 6967** 95,00% C.I.: Inferior= ,3710 Superior= ,8538 F= 3,2972 DF= (30, 30, 0) Sig.= , 0008 (Valor del test= ,0000)
Coeficiente de fiabilidad Nº de casos= 31,0 Nº de ítems= 2 α = ,6967
Acuerdo absoluto Media de la correlación intra-clase= , 5993** 95,00% C.I.: Inferior= -,0392 Superior= ,8289 F= 3,2972 DF= (30, 30, 0) Sig.= , 0008 (Valor del test= ,0000)

* Las abreviaturas AG1 y AG2 se refieren a las puntuaciones analíticas en el PRE y en el POST respectivamente.

Pasamos a analizar ahora las correlaciones de Spearman entre los componentes de la evaluación analítica

La Tabla 10.7 de correlaciones entre los componentes de la evaluación analítica global entre el PRE y el POST nos muestra:

1. Los valores de las correlaciones de los distintos componentes en el PRE / POST que se muestran en la diagonal son bajos.
2. Los niveles de correlación de dos de los componentes de la diagonal de la matriz: *contenido* y *gramática*, no son significativos.
3. Las correlaciones del resto de componentes de la diagonal: *organización* (0,574) y *registro* (0,545) PRE / POST son bajas, pero muy significativas, mientras que: *vocabulario* (0,415), *mecánica* (0,401) y *presentación* (0,388) apenas alcanzan el umbral de la significación.
4. Los valores adyacentes que alcanzan el nivel de significación son: el *vocabulario* del PRE, que correlaciona bien con la *organización*

(0,483) y la *mecánica* (0,402), y la *mecánica* del POST, que tiene niveles de correlación significativa con el *contenido* (0,364), el *vocabulario* (0,402) y el *registro* (0,397).

Tabla 10.7. Correlaciones entre las evaluaciones analíticas (PRE y POST).

Correlaciones

			Analítico contenido2	Analítico organización2	Analítico gramática2	Analítico vocabulario2	Analítico registro2	Analítico mecánica2	Analítico presentación2
Rho de Spearman	Analítico contenido1	Coeficiente de correlación	,339	,347	,137	,318	,267	,364*	,075
		Sig. (bilateral)	,058	,052	,453	,076	,147	,041	,682
		N	32	32	32	32	31	32	32
	Analítico organización1	Coeficiente de correlación	,229	,574**	,184	,291	,137	,222	,320
		Sig. (bilateral)	,207	,001	,314	,107	,463	,221	,074
		N	32	32	32	32	31	32	32
	Analítico gramática1	Coeficiente de correlación	,178	,324	,285	,250	,139	,338	,292
		Sig. (bilateral)	,329	,070	,114	,167	,455	,058	,105
		N	32	32	32	32	31	32	32
	Analítico vocabulario1	Coeficiente de correlación	,341	,483**	,324	,415*	,227	,402*	,129
		Sig. (bilateral)	,056	,005	,071	,018	,219	,023	,481
		N	32	32	32	32	31	32	32
	Analítico registro1	Coeficiente de correlación	,207	,329	,174	,049	,545**	,397*	,312
		Sig. (bilateral)	,256	,066	,341	,790	,002	,025	,082
		N	32	32	32	32	31	32	32
	Analítico mecánica1	Coeficiente de correlación	,243	,384*	,218	,223	,099	,401*	,089
		Sig. (bilateral)	,180	,030	,230	,219	,596	,023	,629
		N	32	32	32	32	31	32	32
	Analítico presentación1	Coeficiente de correlación	,103	,251	,186	,203	,263	,264	,388*
		Sig. (bilateral)	,576	,166	,309	,266	,153	,145	,028
		N	32	32	32	32	31	32	32

*. La correlación es significativa al nivel 0,05 (bilateral).

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

Los números 1 y 2 hacen referencia a las evaluaciones analíticas del PRE y del POST respectivamente

Tras este análisis, cabe subrayar que las correlaciones señaladas entre los componentes: *vocabulario*, *organización*, *contenido* y *mecánica* tienen su justificación desde una perspectiva mayoritariamente léxico-semántica. Llama la atención la ausencia de correlación de la *gramática* consigo mismo en el PRE / POST y con los demás componentes si se tiene en cuenta que la *gramática* es un punto de referencia en la mayor parte de los parámetros estudiados.

10.2.1.3. Correlaciones globales entre las evaluaciones holísticas y analíticas (PRE y POST)

Una vez que se han analizado tanto global como individualmente o por componentes las evaluaciones globales y las analíticas procede hacer un estudio cruzado al menos en las evaluaciones de carácter global (Tabla 10.8).

Tabla 10.8. Correlaciones entre las evaluaciones holísticas y las evaluaciones analíticas (PRE y POST)

Correlaciones

			Holístico Global 1	Holístico Global 2	Analítica Global1	Analítica Global2
Rho de Spearman	Holístico Global 1	Coeficiente de correlación	1,000	,508**	,684**	,495**
		Sig. (unilateral)	,	,001	,000	,002
		N	32	32	32	31
	Holístico Global 2	Coeficiente de correlación	,508**	1,000	,243	,706**
		Sig. (unilateral)	,001	,	,090	,000
		N	32	32	32	31
	Analítica Global1	Coeficiente de correlación	,684**	,243	1,000	,475**
		Sig. (unilateral)	,000	,090	,	,003
		N	32	32	32	31
	Analítica Global2	Coeficiente de correlación	,495**	,706**	,475**	1,000
		Sig. (unilateral)	,002	,000	,003	,
		N	31	31	31	31

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

Los valores PRE y POST los hemos comentado anteriormente (Tablas 10.4 y 10.6). En ambos casos las correlaciones son muy significativas, aunque los valores son muy bajos: 0, 508 para las holísticas y 0,475 para las analíticas. En la correlación cruzada el nivel de significación se mantiene con un $p<0,01$ y los valores se ven incrementados en HG1 y AG1[2] (0,684) y en HG2 y AG2 (0,706) en relación a las correlaciones HG1 y HG2 y AG1 y AG2. Los valores son superiores a los presentados en los apartados anteriores ya que en estos pares cruzados lo que se altera es la calificación del ensayo en el mismo momento ya sea desde una perspectiva global o analítica.

Si bien las diferencias de los valores de las correlaciones son mínimas, estas se incrementan ligeramente en la correlación POST.

Cuando en los cruces se introduce el factor PRE / POST los valores cambian. Sigue siendo muy significativo el cruce HG1 con AG2: 0,495. Sorprende la ausencia de correlación en el cruce HG2 y AG1: 0,243. Es decir, las puntuaciones holísticas que otorgaron los correctores a los ensayos en la segunda ocasión no guardan relación alguna con las puntuaciones analíticas que asignaron los correctores en la primera ocasión. Estos datos revelan de nuevo la discrepancia que se observa entre la actuación teórica de los correctores, que se declaran mayoritariamente consecuente en sus evaluaciones, y su posterior actuación práctica (ver

[2]A partir de ahora utilizaremos las abreviaturas HG1 y HG2 para referirnos a las evaluaciones holísticas del PRE y del POST respectivamente.De igual modo, las abreviaturas AG1 y AG2 se utilizarán para referirnos a las evaluaciones analíticas del PRE y del POST respectivamente.

capítulo 9, epígrafe 9.4). Esta última evidencia actuaciones inconsecuentes y arbitrarias.

La explicación, quizás, pueda encontrarse en la tendencia observada en las puntuaciones; en el PRE los correctores puntuaron más alto en la evaluación analítica y en el POST tanto la media de las puntuaciones holísticas como analíticas bajaron. La similitud de las puntuaciones puede explicar la asociación entre las puntuaciones holísticas del PRE y las puntuaciones analíticas del POST, mientras que la disparidad de las puntuaciones analíticas del PRE y las puntuaciones holísticas del POST puede explicar la no correlación entre AG1 y HG2.

10.2.1.4. Correlaciones entre los ensayos según el nivel de dominio lingüístico

Tras haber estudiado la fiabilidad desde perspectivas globales tanto holísticas como analíticas procede observar el grado de fiabilidad en el ámbito de lo concreto. Estudiar la fiabilidad en los diez ensayos generaría, probablemente, información, en ocasiones, redundante. De ahí que, en aras de la efectividad y del rigor, planteemos aplicar los estadísticos que valoran la fiabilidad en los ensayos que venimos estudiando: h1, h5, y h10, por su carácter representativo del dominio que se tiene de la lengua: bajo, medio, alto. Como ya avanzamos (ver epígrafe 10.1.1), dicha categorización se infiere a partir de las medias de las puntuaciones holísticas asignadas a los ensayos en el PRE: h1 ($\bar{\times}$ = 2,47); h5 ($\bar{\times}$ = 4,56) y h10 ($\bar{\times}$ = 7,78).

Con este análisis se pretende:

1. Observar si la fiabilidad entre las puntuaciones holísticas otorgadas a cada uno de estos ensayos y las asignadas a cada uno de los componentes que configuran la evaluación analítica es superior o inferior a las globales estudiadas anteriormente.
2. Comparar los valores obtenidos en cada uno de los componentes con la puntuación holística
3. Valorar el nivel de correlación en función del nivel del ensayo.

En la tabla 10.9 encontramos respuesta a las preguntas formuladas. El ensayo 1 (h1), con un nivel de dominio de la lengua bajo muestra:

Tabla 10.9. Correlaciones entre las puntuaciones holísticas y las analíticas del ensayo h1 (PRE)

Correlaciones

			contenido1	organización1	gramática1	vocabulario1	registro1	mecánica1	presentación1
Rho de Spearman	hol.1	Coeficiente de correlación	,682**	,675**	,392*	,359*	,534**	,485**	,363*
		Sig. (bilateral)	,000	,000	,027	,044	,002	,005	,041
		N	32	32	32	32	31	32	32

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

*. La correlación es significativa al nivel 0,05 (bilateral).

1. La correlación HG1 y AG1 (0,684) (Tabla 10.8) es superior a las que presentan cada uno de los componentes de la evaluación analítica
2. La correlación es significativa en todos su componentes, pero su valor más alto se encuentra en el *contenido* (0,682) y los más bajos

en la *gramática* (0,392) y en el *vocabulario* (0,359). Este resultado podría ser alentador en cuanto que sugiere que el *contenido* de los ensayos parece jugar un papel relevante en las puntuaciones de los ensayos con un nivel lingüístico bajo, superando a otros aspectos relacionados con la *forma* como es la *gramática* (ver Freedman 1979; Freedman y Pringle 1980; Santos 1988; Song y Caruso 1996)

El ensayo 5 (h5): con un nivel de dominio de la lengua medio presenta:

Tabla 10.10. Correlaciones entre las puntuaciones holísticas y las analíticas del ensayo h5 (PRE))

Correlaciones

		contenido1	organización1	gramática1	vocabulario1	registro1	mecánica1	presentación1
Rho de Spearman	hol.5 Coeficiente de correlación	,636**	,570**	,656**	,726**	,731**	,711**	,572**
	Sig. (bilateral)	,000	,001	,000	,000	,000	,000	,001
	N	32	32	32	32	31	32	32

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

1. Unos valores de correlación en los componentes de: *vocabulario* (0,726), *registro* (0,731) y *mecánica* (0,711) de la evaluación analítica superiores a la correlación HG1 y AG1 (0,684) anteriormente citada (ver Tabla 8).
2. Una correlación muy significativa en todos sus componentes. Si bien hay pocas diferencias, ya que en todos los componentes el nivel de significación es <0,01. Estos datos nos permiten inferir que los componentes que podríamos etiquetar de carácter principalmente

léxico-semántico tienen mayor nivel de asociación con la puntuación holística.

El ensayo 10 (h10), que representa el nivel de dominio de la lengua alto, muestra:

Tabla 10.11. Correlaciones entre las puntuaciones holísticas y las analíticas del ensayo h10 (PRE)

Correlaciones

			contenido1	organización1	gramática1	vocabulario1	registro1	mecánica1	presentación1
Rho de Spearman	hol.10	Coeficiente de correlación	,557**	,447*	,606**	,417*	,430*	,413*	,376*
		Sig. (bilateral)	,001	,010	,000	,018	,016	,019	,034
		N	32	32	32	32	31	32	32

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

*. La correlación es significativa al nivel 0,05 (bilateral).

1. Correlaciones inferiores al valor (0,684) de la HG1 y AG1.
2. Una correlación significativa en todos sus componentes y muy significativa en *contenido* (0,557) y *gramática* (0,606). Las correlaciones son algo más bajas que las que se observaban en los dos ensayos anteriores (Tablas 10.9 y 10.10). Si en el ensayo de un nivel bajo (h1) la correlación más alta se encuentra en el *contenido* (0,682) y en la de tipo medio (h5) en el *vocabulario* (0,726), en este (h10), categorizado como de nivel alto, la correlación más alta se da en el componente *gramática* (0,606). Estos datos muestran la relevancia de cada uno de estos componentes según el tipo de

ensayo: el *contenido* en el ensayo de nivel bajo, el *léxico* en el ensayo de nivel medio y la *gramática* en el de nivel alto.

3. Estos resultados manifiestan que las correlaciones más altas (h1, h5 y h10) se dan en el ensayo de nivel medio. Probablemente, debido al peso del componente léxico.

Este análisis de las correlaciones requiere tomar como referente los ensayos que se acaban de estudiar y compararlos con la correlación que muestran en el POST.

Seguimos el protocolo de análisis utilizado en las correlaciones PRE de estos ensayos.

Ensayo 1 (h1):

Tabla 10.12. Correlaciones entre las puntuaciones holísticas y las analíticas del ensayo h1 (POST)

Correlaciones

			contenido2	organización2	gramática2	vocabulario2	registro2	mecánica2	presentación2
Rho de Spearman	2hol.1	Coeficiente de correlación	,480**	,467**	,523**	,639**	,410*	,619**	,326
		Sig. (bilateral)	,005	,007	,002	,000	,022	,000	,073
		N	32	32	32	32	31	32	31

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

*. La correlación es significativa al nivel 0,05 (bilateral).

1. La correlación HG2 y AG2 (0,706) (Tabla 10.8) es superior a la correlación que muestra cualquiera de los componentes de la evaluación analítica.

2. Las correlaciones del los componentes analíticos son significativas, salvo en la categoría *presentación*. Paradójicamente, en esta segunda ocasión, son las categorías de *vocabulario* (0,639), *mecánica* (0,619) y *gramática* (0,523), todas ellas relacionadas con el aspecto de la *forma*, las que muestran una correlación más alta con las puntuaciones holísticas. Estos datos, demuestran la inconsistencia en el énfasis que los correctores aplican a las distintas categorías analíticas en el PRE y en el POST. Asimismo, ponen en evidencia la relación que se establece entre la *forma* del ensayo y las puntuaciones globales cuando el nivel de dominio lingüístico es bajo.

Ensayo 5 (h5):

Tabla 10.13. Correlaciones entre las puntuaciones holísticas y las analíticas del ensayo h5 (POST)

Correlaciones

			contenido2	organización2	gramática2	vocabulario2	registro2	mecánica2	presentación2
Rho de Spearman	2hol.5	Coeficiente de correlación	,771**	,777**	,419*	,714**	,392*	,446*	,319
		Sig. (bilateral)	,000	,000	,017	,000	,029	,011	,076
		N	31	31	32	32	31	32	32

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

*. La correlación es significativa al nivel 0,05 (bilateral).

1. La correlación HG2 y AG2 (0,706) (Tabla 10.8) es inferior en los componentes: *contenido* (0,771), *organización* (0,777) y *vocabulario* (0,714).
2. En relación al PRE, cabe decir que se mantiene la relevancia del *vocabulario* (0,714) y se incrementan los valores de correlación de los

otros factores que podemos asociar con el aspecto semántico. Por el contrario, el aspecto gramatical (0,419) baja considerablemente su nivel de correlación.

El ensayo 10 (h10):

Tabla 10.14. Correlaciones entre las puntuaciones holísticas y las analíticas del ensayo h10 (POST)

Correlaciones

			contenido2	organización2	gramática2	vocabulario2	registro2	mecánica2	presentación2
Rho de Spearman	2hol.10	Coeficiente de correlación	,574**	,773**	,666**	,640**	,765**	,619**	,646**
		Sig. (bilateral	,001	,000	,000	,000	,000	,000	,000
		N	32	32	32	32	31	32	32

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

1. La correlación HG2 y AG2 (0,706) (Tabla 10.8) es inferior en todos los componentes salvo en *organización* (0,773) y *registro* (0,765).
2. La correlación se incrementa en todos los componentes. La *gramática* (0,666) es en el único ensayo en el que mantiene su relevancia.

Estos datos evidencian la discrepancia entre los correctores en las evaluaciones del PRE y del POST cuando recurrimos a la casuística concreta de los ensayos. Se mantiene la importancia del *contenido* y de la *forma* (i.e. la gramática) en los ensayos con un dominio de la lengua medio y alto (h5 y h10). En el ensayo con un nivel de dominio de lengua bajo se observa una discrepancia mayor, ya que los correctores se muestran

dubitativos sobre los aspectos de la *forma* o del *contenido* que han de subrayar.

Finalmente, si se aplica el análisis de la fiabilidad a estos ensayos se observa que el mayor grado de consistencia se da en el ensayo con un dominio de la lengua alto (0,7220), que junto con la correlación intraclase de acuerdo absoluto (0,7274), según los parámetros de Fleiss (1986), se acerca a los valores óptimos de concordancia.

Por su parte, el ensayo con un nivel lingüístico bajo (h1) registra el menor nivel de consistencia (0,6710) y una correlación intraclase de acuerdo absoluto similar (0,6708), considerada como de *regular-buena* siguiendo los parámetros anteriores.

Por último, el ensayo con un nivel lingüístico medio (h5), si bien registra mayor consistencia que el ensayo (h1), alcanza un grado de acuerdo absoluto menor (0,6663). Estas pequeñas diferencias se podría explicar en función del mayor rango de variabilidad de las puntuaciones que se otorgan a un ensayo de tipo medio.

10.2.2. Estudio de la fiabilidad inter-corrector en las evaluaciones holísticas y analíticas de acuerdo con las variables género y situación laboral.

Una vez analizadas las correlaciones entre los diversos correctores a nivel general, nos interesa averiguar la influencia específica que ejercen las

variables género (i.e. hombre / mujer) y situación laboral (secundaria / universidad) de los correctores en las valoraciones del rendimiento de los candidatos.

El estudio de la fiabilidad inter-corrector entre las evaluaciones holísticas y las analíticas PRE y POST se realizará a través de la implementación de la técnica de análisis de la varianza (ANOVA). El análisis de la varianza (ANOVA) es una técnica estadística general, desarrollada por Fisher, que permite estudiar el efecto de uno o más factores (i.e. variables independientes categóricas) sobre una respuesta cuantitativa.

En nuestro caso, hemos estudiado el efecto de las variables categóricas de género y situación laboral en los distintos tipos de evaluación. Esto nos va a permitir calcular la variabilidad de las puntuaciones tanto dentro de un mismo grupo como entre los distintos grupos y determinar si las diferencias que se observan son significativas.

El protocolo de estudio establecido es el siguiente:

a) análisis de los estadísticos descriptivos;
b) estudios de los gráficos de perfil y
c) análisis de las pruebas de contraste intra-sujetos.

Este análisis se aplicará a los distintos subgrupos en función de dos variables:

a) *factor tiempo*. Aquí se examinarán las siguientes puntuaciones.
   - puntuaciones holísticas en el PRE y en el POST (HG1 y HG2 respectivamente)
   - puntuaciones analíticas en el PRE y en el POST (AG1 y AG2 respectivamente)

b) *método de corrección*. Se analizarán las siguientes puntuaciones:
   - puntuaciones holísticas (HG1) y analíticas en el PRE (AG1)
   - puntuaciones holísticas y puntuaciones analíticas en el POST (HG2 y AG2 respectivamente)

Los resultados obtenidos se detallan en los siguientes subepígrafes:

10.2.2.1. Puntuaciones holísticas en el PRE y en el POST (HG1 y HG2 respectivamente)

Empezaremos nuestro estudio analizando los estadísticos descriptivos de las puntuaciones holísticas en el PRE y en el POST en los distintos grupos de correctores establecidos de acuerdo con las variables género y situación laboral (Tabla 10.15). Se examinará, en primer lugar, la actuación de los grupos a nivel general.

Tabla 10.15. Estadísticos descriptivos: evaluaciones holísticas en el PRE (HG1) y en el POST (HG2)

Estadísticos descriptivos

	sexo	profesión	Media	Desv. típ.	N
Holístico Global 1	Hombre	secundaria	49,1250	11,5936	8
		universidad	51,8750	9,7312	8
		Total	50,5000	10,4371	16
	Mujer	secundaria	49,6250	13,8558	8
		universidad	43,6250	7,3278	8
		Total	46,6250	11,1467	16
	Total	secundaria	49,3750	12,3444	16
		universidad	47,7500	9,3488	16
		Total	48,5625	10,8030	32
Holístico Global 2	Hombre	secundaria	44,3750	5,1807	8
		universidad	49,7500	4,8917	8
		Total	47,0625	5,6032	16
	Mujer	secundaria	45,7500	6,7771	8
		universidad	42,5000	8,4177	8
		Total	44,1250	7,5708	16
	Total	secundaria	45,0625	5,8705	16
		universidad	46,1250	7,6322	16
		Total	45,5938	6,7195	32

Como vemos, la media de las puntuaciones holísticas globales de los distintos grupos es superior en el PRE ($\bar{\times}$ = 48,5625) que en el POST ($\bar{\times}$ = 45,5938). Estos datos sugieren que los correctores se mostraron más exigentes en la evaluación holística de los ensayos en el POST. Por otra parte, la media de los distintos grupos en el PRE y en el POST señala algunas tendencias en la actuación de estos grupos que se mantiene en ambas ocasiones (PRE y POST).

Así, el estudio específico de las variables género y situación laboral indica que el grupo de hombres se muestra más indulgente que el grupo de mujeres corrigiendo los ensayos de forma holística tanto en el PRE ($\bar{\times}$ = 50,5000 vs. $\bar{\times}$ = 46,6250 hombres y mujeres respectivamente) como en el POST ($\bar{\times}$ = 47,0625 vs. $\bar{\times}$ = 44,1250 hombres y mujeres respectivamente).

Los datos también demuestran que el grupo de hombres de universidad es el que registra las puntuaciones medias más altas en ambas ocasiones ($\bar{\times}$ = 51,8750 en el PRE) y ($\bar{\times}$ = 49,7500 en el POST). Por tanto, se deduce que este grupo es el que presenta un perfil evaluador de mayor condescendencia (ver capítulo 9, epígrafe 9.4).

Por el contrario, las puntuaciones medias del grupo de mujeres de universidad en el PRE ($\bar{\times}$ = 43,6250) y en el POST ($\bar{\times}$ = 42,5000) indican que este grupo de correctores es el más exigente en la evaluación de los ensayos.

Asimismo, y de acuerdo con este planteamiento, los datos sugieren que el grupo de mujeres de secundaria es más indulgente que el grupo de hombres de secundaria en la evaluación holística de los ensayos en ambas ocasiones ($\bar{\times}$ = 49,6250 vs. $\bar{\times}$ = 44,3750 mujeres y hombres de secundaria respectivamente en el PRE) y ($\bar{\times}$ = 45,7500 vs. $\bar{\times}$ = 44,3750 mujeres y hombres de secundaria respectivamente en el POST).

Es interesante resaltar la gran desviación típica que presenta el profesorado de secundaria en su conjunto en la evaluación holística de los ensayos en el PRE (DT = 12,3444). Este resultado indica que hay una gran dispersión en las puntuaciones holísticas que asignan dichos profesores a los ensayos. Esta dispersión de puntuaciones se ve sensiblemente reducida en el POST, tal y como nos indica la menor desviación típica (DT = 5,8705) que se registra en esta ocasión. De ahí, se deduce que las puntuaciones holísticas que otorgó el profesorado de secundaria a los ensayos en el PRE

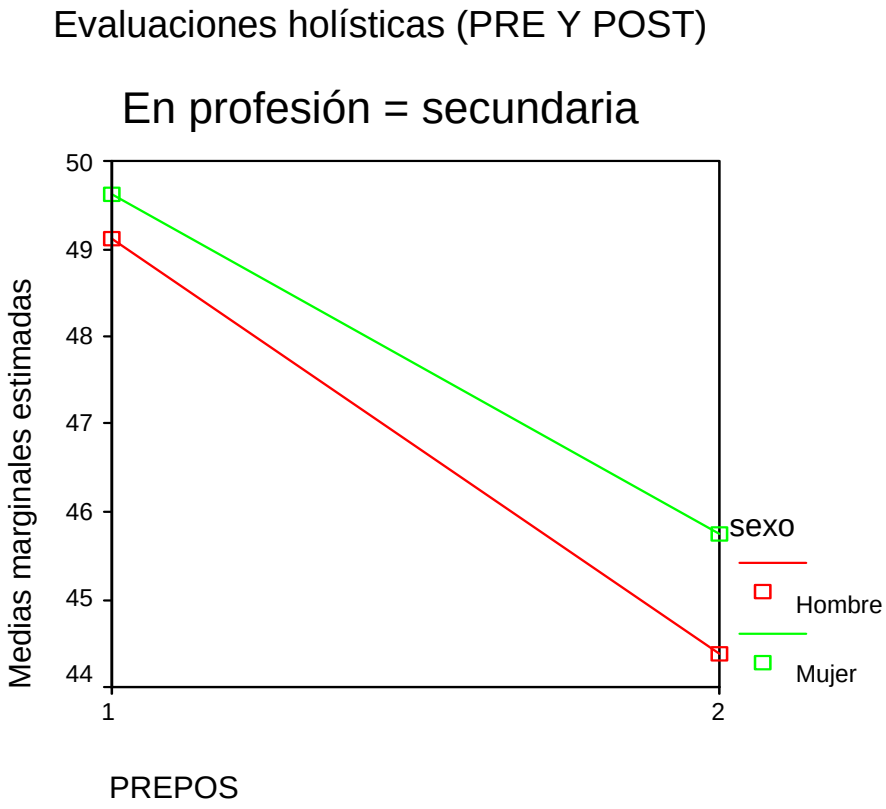
fueron mucho menos consistentes que las que asignaron a los ensayos en el POST.

Los gráficos que se presentan a continuación tratarán de facilitar la comprensión de la información suministrada por los estadísticos descriptivos (Figs. 10.1a y 10.1b). Estos gráficos reflejan de forma plástica las tendencias que se observan entre los distintos grupos de correctores en las evaluaciones holísticas de los ensayos (PRE y POST).

El primer gráfico (Fig. 10.1a) representa la evaluación holística de los ensayos (PRE y POST) del grupo de hombres y de mujeres de enseñanza secundaria. Como se puede apreciar, la dirección descendente de las líneas de actuación nos indica que las puntuaciones holísticas de ambos grupos son más altas en el PRE que en el POST. Las puntuaciones holísticas en el POST experimentan un notable descenso en ambos grupos de correctores.

Asimismo, la distinta altura de las líneas de actuación en el gráfico nos permite inferir que el grupo de hombres de secundaria se muestra más exigente que el grupo de mujeres de secundaria en la evaluación holística de los ensayos en ambas ocasiones (PRE y POST).

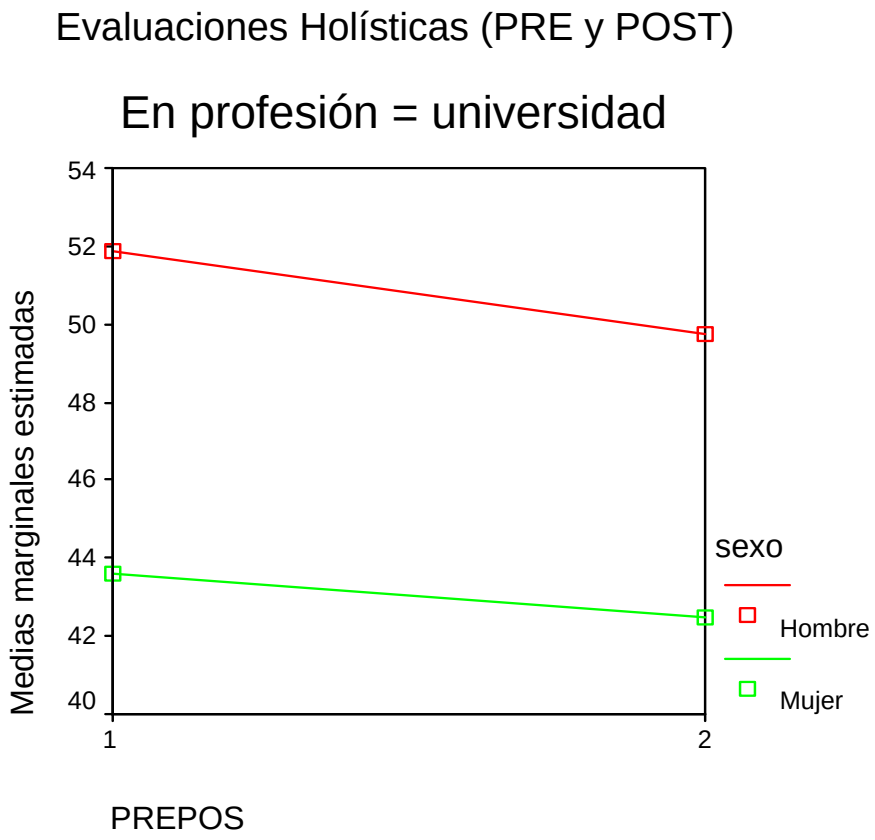
Fig.10.1a. Evaluaciones holísticas (PRE y POST) de acuerdo con el género y situación laboral (i.e. enseñanza secundaria) de los correctores



* La abreviatura PREPOS hace referencia al conjunto de las evaluaciones holísticas en el PRE y en el POST respectivamente.

La Fig. 10.1b representa la actuación del grupo de hombres y de mujeres de universidad en las evaluaciones holísticas de los ensayos PRE y POST. Como vemos, las líneas de actuación siguen una dirección descendente. Esto indica que, de nuevo, las puntuaciones holísticas de ambos grupos de correctores son más altas en el PRE que en el POST.

Fig.10.2b. Evaluaciones holísticas (PRE y POST) de acuerdo con el género y profesión (i.e. enseñanza universitaria) de los correctores



* La abreviatura PREPOS hace referencia al conjunto de las evaluaciones holísticas en el PRE y en el POST respectivamente.

No obstante, la inclinación de las líneas de actuación en este gráfico es mucho menor que la que se observa en el gráfico anterior (Fig. 10.1a). Este dato nos permite inferir que las diferencias entre las puntuaciones holísticas PRE y POST de los correctores de universidad fueron menos notables y, por consiguiente, más consistentes que la de los correctores de secundaria.

Asimismo, la distancia que separa las líneas de actuación del grupo de hombres y mujeres de universidad refleja la disparidad de criterios que aplican ambos grupos a la evaluación de los ensayos. En cualquier caso, el grupo de hombres de universidad se muestra mucho más indulgente que el de las mujeres en ambas ocasiones (PRE y POST).

Por último, se llevará a cabo el análisis de las pruebas de contrastes intra-sujetos. Los resultados de este análisis nos van a permitir averiguar si las diferencias que hemos observado hasta ahora en la actuación de los correctores son significativas desde un punto de vista estadístico.

Los resultados de las pruebas de contrastes intra-sujetos (Tabla 10.16) nos indican que no hay diferencias significativas entre los correctores en la evaluación holística de los ensayos (PRE y POST) de acuerdo con las variables género y situación laboral ($F_{1,31}$ = 3,151; p = 0,087). En otras palabras, la variabilidad que se observa en las evaluaciones holísticas PRE y POST no es significativa y no puede atribuirse ni al género ni a la situación laboral de los correctores. Se requiere una muestra más amplia de sujetos para poder confirmar estas tendencias y predecir las futuras actuaciones de los correctores.

Tabla 10.16. Pruebas de contrastes intra-sujetos

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	PREPOS	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
PREPOS	Lineal	141,016	1	141,016	3,151	,087
PREPOS * SX	Lineal	3,516	1	3,516	,079	,781
PREPOS * PROF	Lineal	28,891	1	28,891	,646	,428
PREPOS * SX * PROF	Lineal	1,563E-02	1	1,563E-02	,000	,985
Error(PREPOS)	Lineal	1253,062	28	44,752		

* La abreviatura PREPOS se refiere al conjunto de las evaluaciones holísticas del PRE y del POST. Las abreviaturas SX y PROF se refieren a las variables género y situación laboral respectivamente.

10.2.2.2. Puntuaciones analíticas en el PRE y en el POST

Siguiendo con el protocolo establecido, iniciaremos el estudio de las puntuaciones analíticas PRE y POST examinando los estadísticos descriptivos de los diversos grupos de correctores (Tabla 10.17).

Tabla 10.17. Estadísticos descriptivos: evaluaciones analíticas en el PRE (AG1) y en el POST (AG2)

Estadísticos descriptivos

	sexo	profesión	Media	Desv. típ.	N
Analítica Global1	Hombre	secundaria	59,0357	7,4316	8
		universidad	60,7143	9,5825	8
		Total	59,8750	8,3293	16
	Mujer	secundaria	60,7143	11,3918	8
		universidad	57,1837	4,0116	7
		Total	59,0667	8,6665	15
	Total	secundaria	59,8750	9,3320	16
		universidad	59,0667	7,4922	15
		Total	59,4839	8,3610	31
Analítica Global2	Hombre	secundaria	54,2500	4,4432	8
		universidad	56,2143	7,1596	8
		Total	55,2321	5,8449	16
	Mujer	secundaria	52,9643	7,5577	8
		universidad	53,1837	4,8218	7
		Total	53,0667	6,2077	15
	Total	secundaria	53,6071	6,0257	16
		universidad	54,8000	6,1679	15
		Total	54,1843	6,0229	31

Se analizará, en primer lugar, la actuación de los correctores a nivel general. Como se observa, y al igual que ocurría en las evaluaciones holísticas, las puntuaciones medias del conjunto total de correctores en las evaluaciones analíticas de los ensayos son más altas en el PRE ($\bar{\times}$ = 59,4839) que en el POST ($\bar{\times}$ = 54,1843). Estos resultados nos permiten inferir que los correctores se mostraron más estrictos en la evaluación analítica de los ensayos en el POST.

Asimismo, los datos específicos suministrados por las variables género y situación laboral parecen repetir las tendencias que se habían observado anteriormente en el estudio de las evaluaciones holísticas PRE y POST. Las puntuaciones medias sugieren que el grupo de hombres se muestra, en general, ligeramente más indulgente que el grupo de mujeres corrigiendo los ensayos tanto en el PRE ($\bar{\times}$ = 59,8750 vs. $\bar{\times}$ = 59, 0667 hombres y mujeres respectivamente) como en el POST ($\bar{\times}$ = 55,2321 vs. $\bar{\times}$ = 53,0667 hombres y mujeres respectivamente).

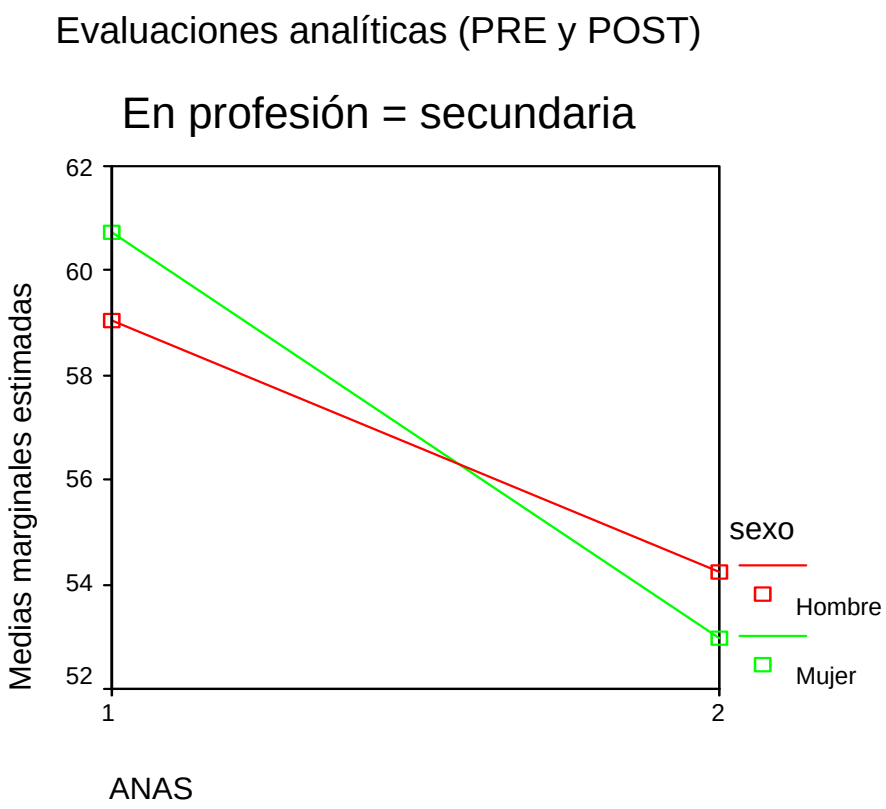
Al igual que observábamos en el análisis de las puntuaciones holísticas PRE y POST, el grupo de mujeres de secundaria ($\bar{\times}$ = 60,7143) asigna a los ensayos puntuaciones analíticas más altas que el grupo de hombres de secundaria ($\bar{\times}$ = 59,0357) en el PRE. No obstante, esta vez, los valores se invierten en el POST ya que el grupo de hombres de

secundaria ($\bar{\times}$ = 54,2500) otorga puntuaciones analíticas más altas que el de mujeres de secundaria ($\bar{\times}$ = 53, 6071) en esta ocasión.

Por otro lado, conviene destacar la gran dispersión de puntuaciones que se observa en el grupo de mujeres de secundaria en el PRE (DT = 11,3918). Estos resultados, junto con el de las evaluaciones holísticas del PRE analizadas anteriormente (ver Tabla 10.15), indican que el grupo de mujeres de secundaria se mostró poco consistente y asignó puntuaciones muy heterogéneas tanto en las evaluaciones holísticas como en las analíticas del PRE.

La información suministrada por los estadísticos descriptivos se representa de forma visual en los gráficos que se incluyen a continuación (Fig. 10.2a y 10.2b). El primer gráfico (Fig. 10.2a) nos muestra la actuación del grupo de hombres y de mujeres de enseñanza secundaria en las evaluaciones analíticas (PRE y POST).

Fig. 10.2a. Evaluaciones analíticas en el PRE y en el POST de acuerdo con las variables género y situación laboral de los correctores (i.e. enseñanza secundaria)



*La abreviatura ANAS se refiere a las evaluaciones analíticas PRE y POST.

Como vemos, se confirma la tendencia que se observaba en las evaluaciones holísticas (PRE y POST). La dirección descendente de las líneas de actuación nos lleva a inferir que tanto el grupo de hombres como el de mujeres se muestra más estricto en la evaluación analítica de los ensayos en el POST.

Cabe hacer notar que la línea de actuación del grupo de mujeres desciende de forma más abrupta que la del grupo de hombres. Este dato

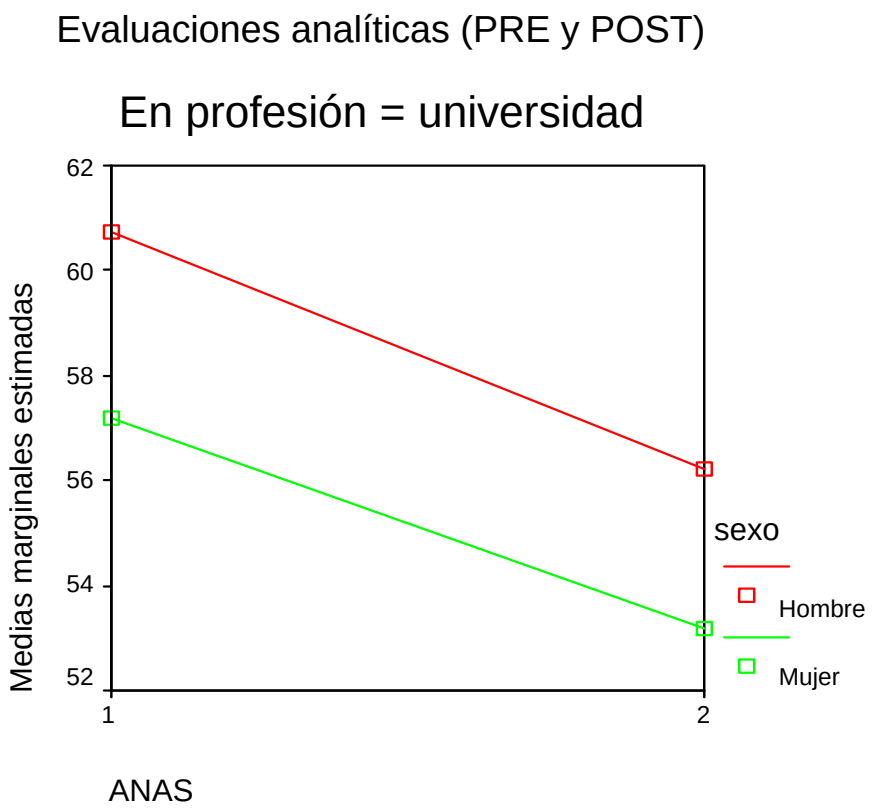
revela que la discrepancia entre las puntuaciones analíticas del PRE y del POST es superior en este primer grupo.

Asimismo, el gráfico confirma que el grupo de mujeres puntúa ligeramente más alto que el grupo de hombres en la evaluación analítica del PRE, si bien, tal y como se aprecia, los valores se invierten en el POST. En este sentido, la caída abrupta de las líneas de actuación del grupo de mujeres de secundaria es indicativa del alto grado de exigencia que demuestra este grupo en las evaluaciones analíticas de los ensayos en el POST. De hecho, la media de las puntuaciones analíticas del grupo de mujeres de secundaria ($\bar{\times}$ = 52,9643) (ver Tabla 10.17) es la más baja que se registra a nivel general.

Por último, es interesante destacar el cruce de las líneas de actuación que se observa en el gráfico (Fig.10.2a). Esto nos indica que hay interacción en la actuación de los correctores de secundaria de acuerdo con la variable género.

El gráfico que se incluye a continuación (Fig. 10.2b) representa de forma plástica la actuación de los correctores de enseñanza universitaria.

Fig. 10.2b. Evaluaciones analíticas en la primera y segunda ocasión (AG1 y AG2) de acuerdo con las variables género y situación laboral de los correctores (i.e. universidad)



*La abreviatura ANAS se refiere a las evaluaciones analíticas PRE y POST.

Como podemos comprobar, se repiten las tendencias anteriores. La dirección descendente de las líneas de actuación nos indica que tanto el grupo de hombres como el de mujeres asigna puntuaciones analíticas más altas en el PRE que en el POST.

En el gráfico también se aprecia que el grupo de hombres de universidad se muestra más indulgente que el grupo de mujeres de universidad en las evaluaciones analíticas de los ensayos en ambas ocasiones.

Las líneas de actuación de ambos grupos siguen direcciones más bien paralelas. Este dato confirma que no se da ningún tipo de interacción en la actuación del profesorado de universidad de acuerdo con la variable género. Así pues, ambos grupos de profesores mantienen su independencia tanto en las evaluaciones holísticas (PRE y POST) de los ensayos (ver Fig. 10.1b) como en las analíticas (PRE y POST).

Por último, se hace necesario comprobar el grado de significación estadística que presentan las diferencias que hemos observado entre los distintos grupos de correctores. Procedemos, por consiguiente, a realizar el análisis de las pruebas de contraste intra-sujetos.

Los resultados de la Tabla 10.18 nos indican que las puntuaciones analíticas del PRE y del POST muestran diferencias significativas. En otras palabras, los correctores no son consistentes en las distintas evaluaciones analíticas de los ensayos (PRE y POST). El alto grado de significación obtenido ($F_{1,31} = 16,299$; $p = 0,000$) sugiere que existe una gran probabilidad de que dicha inconsistencia se repita en ocasiones futuras.

No obstante, como vemos, la discrepancia que se observa en el comportamiento de los correctores en las evaluaciones analíticas PRE y POST no pueden atribuirse al género o a la situación laboral de los correctores dado que los resultados del análisis de estas variables no muestra diferencias significativas.

Tabla 10.18. Pruebas de contrastes intra-sujetos (PRE y POST).

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	ANAS	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
ANAS	Lineal	427,243	1	427,243	16,299	,000
ANAS * SX	Lineal	5,863	1	5,863	,224	,640
ANAS * PROF	Lineal	15,725	1	15,725	,600	,445
ANAS * SX * PROF	Lineal	11,587	1	11,587	,442	,512
Error(ANAS)	Lineal	707,750	27	26,213		

* La abreviatura ANAS representa el conjunto de las evaluaciones analíticas del PRE y del POST. Las abreviaturas SX y PROF se refieren a las variables género y situación laboral respectivamente.

10.2.2.3. Puntuaciones holísticas y analíticas en el PRE

Una vez estudiado el comportamiento de los correctores en las evaluaciones holísticas y analíticas en función del factor tiempo (PRE y POST), pasamos a analizar dichas evaluaciones en función del método de evaluación. Para ello, vamos a cruzar las puntuaciones holísticas y analíticas e investigar la relación que se establece entre ellas.

Iniciamos nuestro estudio examinando las puntuaciones holísticas y analíticas en el PRE. La primera aproximación a los datos es el estudio de los estadísticos descriptivos que se presentan en la Tabla 10.19.

Tabla 10.19. Puntuaciones holísticas (HG1) y analíticas (AG1) en el PRE de acuerdo con la variable género y situación laboral.

Estadísticos descriptivos

	sexo	profesión	Media	Desv. típ.	N
Holístico Global 1	Hombre	secundaria	49,1250	11,5936	8
		universidad	51,8750	9,7312	8
		Total	50,5000	10,4371	16
	Mujer	secundaria	49,6250	13,8558	8
		universidad	43,6250	7,3278	8
		Total	46,6250	11,1467	16
	Total	secundaria	49,3750	12,3444	16
		universidad	47,7500	9,3488	16
		Total	48,5625	10,8030	32
Analítico Global1	Hombre	secundaria	59,0357	7,4316	8
		universidad	60,7143	9,5825	8
		Total	59,8750	8,3293	16
	Mujer	secundaria	60,7143	11,3918	8
		universidad	56,0357	4,9332	8
		Total	58,3750	8,8179	16
	Total	secundaria	59,8750	9,3320	16
		universidad	58,3750	7,7489	16
		Total	59,1250	8,4719	32

La lectura global de los resultados nos indica que, en general, los correctores asignan puntuaciones más altas en las evaluaciones analíticas ($\bar{\times}$ = 59,1250) que en las evaluaciones holísticas ($\bar{\times}$ = 48,5625) de los ensayos en el PRE.

El estudio específico de las variables género y situación laboral confirma las tendencias hasta ahora observadas. Así, el grupo total de hombres tiende a puntuar más alto que el grupo total de mujeres tanto en las evaluaciones holísticas ($\bar{\times}$ = 50,5000 vs. $\bar{\times}$ = 46,6250 hombres y mujeres respectivamente) como en las evaluaciones analíticas ($\bar{\times}$ = 59,8750 vs. $\bar{\times}$ = 58, 3750 hombres y mujeres respectivamente).

Como vemos, las mayores diferencias entre el grupo de hombres y el grupo de mujeres se registran en las puntuaciones holísticas. Estos datos sugieren que el método de evaluación analítico produce resultados más consistente, y, por tanto, más fiables (ver Song y Caruso 1996).

Asimismo, las puntuaciones medias de los diversos grupos de correctores corroboran que el grupo de hombres de universidad es el grupo de correctores más indulgente a nivel general tanto en la evaluación holística ($\bar{\times}$ = 51,8750) como en la evaluación analítica ($\bar{\times}$ = 60,7143) de los ensayos. Por su parte, el grupo de mujeres de universidad es el más estricto en sus valoraciones del rendimiento de los candidatos ($\bar{\times}$ = 43,6250 y $\bar{\times}$ = 56,0357 en las evaluaciones holísticas y analíticas respectivamente).

Como se viene observando, la desviación típica de las puntuaciones totales del grupo del profesorado de secundaria (DT = 12,3444 y DT = 9,3320 en la evaluación holística y analítica respectivamente) tiende a ser muy elevada si se compara con la del grupo del profesorado de universidad (DT = 9,3488 y DT = 7,7489 en la evaluación holística y analítica respectivamente). Estos datos indican que las puntuaciones holísticas y analíticas del grupo del profesorado de secundaria son muy dispersas. Es decir, las puntuaciones de este último grupo, especialmente las puntuaciones holísticas, son poco homogéneas y consistentes.

Conviene resaltar que el grupo de hombres de secundaria reduce considerablemente la dispersión de sus puntuaciones holísticas (DT = 11,5936) en las evaluaciones analíticas (DT = 7,4316). Este resultado era, en cierta medida, previsible ya que la escala holística de puntuaciones

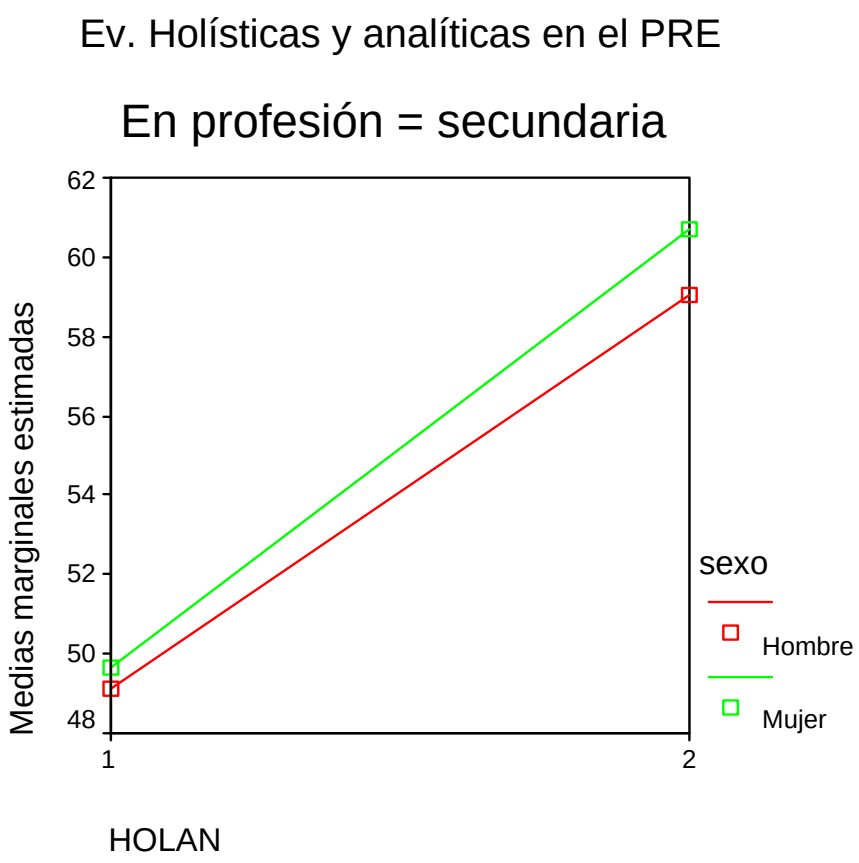
consta de 9 valores (valores del 1 al 10) mientras que la escala analítica consta únicamente de 5 valores (valores del 1 al 5). Esto es, el margen de actuación de los correctores se ve considerablemente reducido en la escala analítica. Este hecho favorece que las puntuaciones se agrupen y sean, por tanto, más homogéneas.

No obstante, llama la atención que el grupo de mujeres de secundaria mantenga el alto grado de dispersión de sus puntuaciones en ambos tipos de evaluación: evaluación holística (DT = 10,8558) y evaluación analítica (DT = 11,3918). Estos datos, unidos a los resultados obtenidos hasta ahora, perfilan al grupo de mujeres de secundaria como el grupo de correctores más inconsistente.

La información facilitada por los estadísticos descriptivos se representa de forma plástica en los gráficos siguientes (Fgs. 10.3a y 10.3b).

La Fig. 10.3a nos muestra la actuación del grupo de hombres y de mujeres de enseñanza secundaria en la evaluación holística y analítica de los ensayos en el PRE. Como se ve, a diferencia de lo que ocurría en el análisis de las puntuaciones holísticas (PRE y POST) y analíticas (PRE y POST), la dirección de las líneas de actuación de los correctores muestra, esta vez, una inclinación ascendente. De ahí, se deduce que las puntuaciones analíticas fueron más altas que las puntuaciones holísticas.

Fig. 10.3a. Evaluaciones holísticas y analíticas en el PRE de acuerdo con el género y profesión (i.e. secundaria)de los correctores.



La abreviatura HOLAN responde a las evaluaciones holísticas y analíticas en el PRE

Asimismo, la altura final que alcanzan dichas líneas de actuación revela una gran disparidad en los criterios holísticos y analíticos que aplican ambos grupos de correctores. La discrepancia entre las puntuaciones holísticas y analíticas del grupo de hombres ($\bar{\times}$ = 49,1250 vs. $\bar{\times}$ = 59,0357 puntuaciones holísticas y analíticas respectivamente) y del grupo de mujeres ($\bar{\times}$ = 49,6250 vs. $\bar{\times}$ = 60,7143 puntuaciones holísticas y analíticas

respectivamente) es muy notable. Como dijimos, la evaluación analítica es la que muestra valores más altos en ambos grupos.

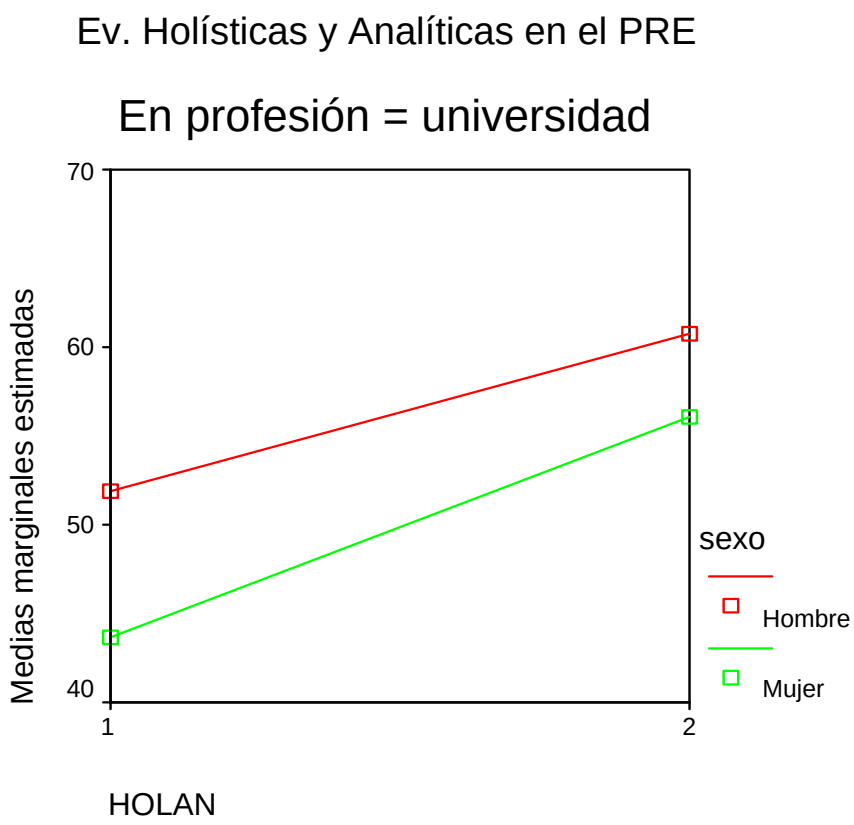
Por último, los datos sugieren que el grupo de mujeres de secundaria es ligeramente menos estricto que el grupo de hombres de secundaria tanto en las puntuaciones holísticas ($\bar{\times}$ = 49,6250 vs. $\bar{\times}$ = 49,1250 respectivamente) como en las puntuaciones analíticas ($\bar{\times}$ = 60,7143 vs. $\bar{\times}$ = 59,0357 respectivamente).

La Fig. 10.3b refleja la actuación del grupo de hombres y de mujeres de enseñanza universitaria. Como se comprueba, la dirección ascendente de las líneas de actuación corrobora la tendencia general de los correctores a asignar puntuaciones más altas en las evaluaciones analíticas que en las holísticas.

Como era previsible, y a diferencia del profesorado de secundaria, la lectura del gráfico indica que la actuación del grupo de hombres y de mujeres de universidad es más dispar en la evaluación holística que en la evaluación analítica de los ensayos.

Finalmente, se reiteran las tendencias que sugieren que el grupo de hombres de universidad es más indulgente que el grupo de mujeres de universidad tanto en las asignación de puntuaciones holísticas como en la de puntuaciones analíticas.

Fig. 10.3b. Gráfico de perfil de las evaluaciones holísticas y analíticas en el PRE de acuerdo al género y profesión (i.e. universidad)de los correctores



La abreviatura HOLAN responde a las evaluaciones holísticas

Nuestro siguiente paso consiste en examinar los resultados de las pruebas de contrastes intra-sujetos (Tabla 10.20). Este análisis nos permitirá averiguar el grado de significación de las diferencias que se observan en las actuaciones de los correctores.

Tabla 10.20. Pruebas de contrastes intra-sujetos

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	HOLAN	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
HOLAN	Lineal	1785,062	1	1785,062	57,514	,000
HOLAN * SX	Lineal	22,563	1	22,563	,727	,401
HOLAN * PROF	Lineal	6,250E-02	1	6,250E-02	,002	,965
HOLAN * SX * PROF	Lineal	5,726	1	5,726	,184	,671
Error(HOLAN)	Lineal	869,036	28	31,037		

* La abreviatura HOLAN corresponde al conjunto de las evaluaciones holísticas y analíticas en el PRE. Las abreviaturas SX y PROF se refieren a las variables género y situación laboral respectivamente.

Como se puede apreciar, los resultados indican que la discrepancia entre las evaluaciones holísticas y analíticas es significativa ($F_{1,31}$ = 57,514; p = 0,000). Estos datos nos permite inferir que el método de evaluación que se elija para la corrección de los ensayos afectará las puntuaciones de los correctores (ver Hamp-Lyons, 1991). No obstante, los datos indican que las diferencias entre las puntuaciones holísticas y las analíticas no son atribuibles al género o a la situación laboral de los correctores.

10.2.2.4. Puntuaciones holísticas y analíticas en el POST

El último grupo de estudio es el de las puntuaciones holísticas y analíticas en el POST. Siguiendo con el la línea de actuación establecida, vamos a empezar centrando nuestra atención en los estadísticos descriptivos que de la Tabla 10.21.

Tabla 10.21. Estadísticos descriptivos: evaluaciones holísticas (HG2) y evaluaciones analíticas (AG2) en el POST

Estadísticos descriptivos

	sexo	profesión	Media	Desv. típ.	N
Holístico Global 2	Hombre	secundaria	44,3750	5,1807	8
		universidad	49,7500	4,8917	8
		Total	47,0625	5,6032	16
	Mujer	secundaria	45,7500	6,7771	8
		universidad	41,0000	7,8528	7
		Total	43,5333	7,4438	15
	Total	secundaria	45,0625	5,8705	16
		universidad	45,6667	7,6687	15
		Total	45,3548	6,6910	31
Analítico Global 2	Hombre	secundaria	54,2500	4,4432	8
		universidad	56,2143	7,1596	8
		Total	55,2321	5,8449	16
	Mujer	secundaria	52,9643	7,5577	8
		universidad	53,1837	4,8218	7
		Total	53,0667	6,2077	15
	Total	secundaria	53,6071	6,0257	16
		universidad	54,8000	6,1679	15
		Total	54,1843	6,0229	31

Una primera lectura global de los resultados confirma que las puntuaciones medias de los distintos grupos de correctores repiten las tendencias que se observaban en las evaluaciones holísticas y analíticas en el PRE. Así, las puntuaciones analíticas totales ($\bar{\times}$ = 54,1843) son superiores a las puntuaciones holísticas totales ($\bar{\times}$ = 45,3548) en el POST. De ahí se deduce que los correctores se muestran más estrictos corrigiendo los ensayos de forma holística que de forma analítica.

El estudio específico de las variables género y situación laboral indica que el grupo de hombres asigna, en general, puntuaciones más altas a los ensayos que el grupo de mujeres. Este fenómeno se observa tanto en las

evaluaciones holísticas ($\bar{\times}$ = 47,0625 vs. $\bar{\times}$ = 43,5333 hombres y mujeres respectivamente) como en las evaluaciones analíticas ($\bar{\times}$ = 55,2321 vs. $\bar{\times}$ = 53,0667 hombres y mujeres respectivamente).

Como vemos, el grupo de hombres de universidad mantiene la misma línea de actuación que en ocasiones anteriores y se perfila como el grupo de correctores más indulgente en ambos tipos de evaluación (ver Herrera, 2000 / 2001).

Cabe destacar que, por primera vez, el grupo de mujeres de universidad ($\bar{\times}$ = 53,1837) registra puntuaciones más altas que las del grupo de mujeres de secundaria ($\bar{\times}$= 52,9643) en las evaluaciones analíticas del POST. No obstante, el grupo de mujeres de universidad ($\bar{\times}$ = 41,0000) sigue mostrándose más exigente que el grupo de mujeres de secundaria ($\bar{\times}$ = 45,7500) en las evaluaciones holísticas del POST.

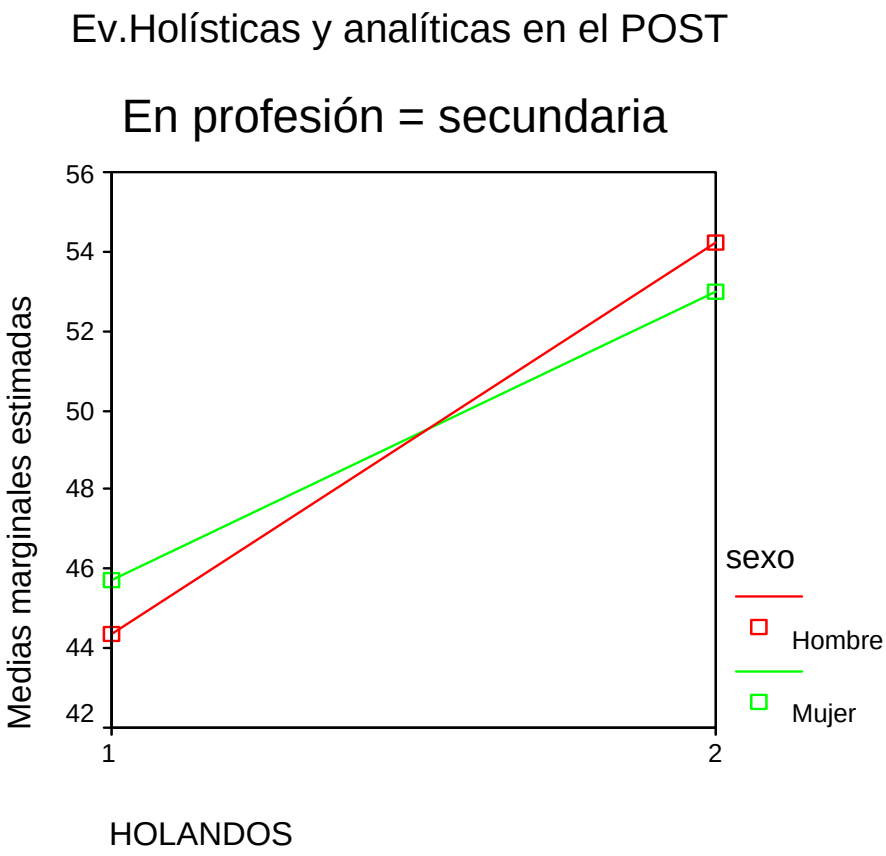
A diferencia de lo que ocurría en las evaluaciones holísticas y analíticas del PRE, el grado de similitud que se observa en las desviaciones típicas de los grupos de correctores en el POST resulta muy llamativo. La dispersión de las puntuaciones holísticas y analíticas se ve notablemente reducida en el POST. Esto significa que las puntuaciones de los correctores son, inexplicablemente, mucho más homogéneas y consistentes en el POST. Paradójicamente, el grupo de mujeres de secundaria, que viene registrando la mayor dispersión de puntuaciones en todos los casos analizados hasta ahora (ver Tablas 10.15, 10.17 y 10.19), revela, en esta ocasión, una desviación típica menor que la del grupo de mujeres de

universidad en las evaluaciones holísticas del POST ($\bar{\times}$ = 6,7771 vs. $\bar{\times}$ = 7,8528 mujeres de secundaria y universidad respectivamente).

La información suministrada por los estadísticos descriptivos se reproduce visualmente en los gráficos siguientes (Figs. 10.4a y 10.4b).

La Fig. 10.4a representa la actuación del grupo de hombres y de mujeres de secundaria.

Fig. 10.4a. Evaluaciones holísticas y evaluaciones analíticas en el POST de acuerdo con las variables género y situación laboral de los correctores (i.e. secundaria)



*La abreviatura HOLANDOS se refiere a las evaluaciones holísticas y analíticas del POST

Como vemos, se mantiene la línea de actuación ascendente en las puntuaciones analíticas dado que éstas son más altas que las puntuaciones holísticas.

Se reitera la tendencia del grupo de mujeres de secundaria a asignar puntuaciones holísticas más altas a los ensayos que el grupo de hombres de secundaria.

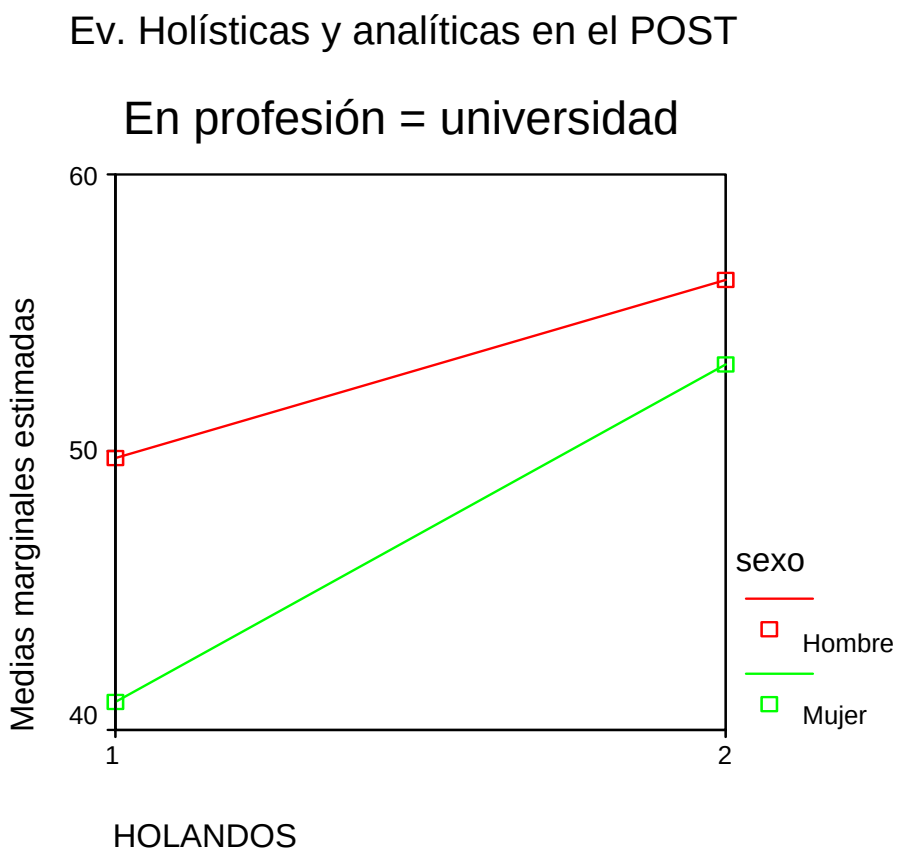
No obstante, el punto final de las líneas de actuación de ambos grupos de correctores indica que el grupo de mujeres termina asignando puntuaciones analíticas más bajas que el grupo de hombres, tal y como ocurría en las evaluaciones analíticas PRE y POST. Al igual que en esta última ocasión, el gráfico indica que hay interacción en la actuación de los profesores de secundaria con respecto al género (ver Tabla 10.2a)

La Fig. 10.4b representa la actuación del grupo de hombres y mujeres de enseñanza universitaria. La dirección ascendente de las líneas de actuación de los correctores indica que las puntuaciones analíticas fueron más altas que las puntuaciones holísticas en ambos grupos.

Asimismo, el gráfico revela una gran discrepancia en las evaluaciones holísticas de ambos grupos; el grupo de mujeres se muestra mucho más exigente que el grupo de hombres corrigiendo los ensayos. Al igual que ocurría en las evaluaciones analíticas del PRE, la discrepancia entre los dos grupos se reduce considerablemente en las evaluaciones analíticas del POST, dada la mayor proximidad de las líneas de actuación de ambos

grupos en estas últimas evaluaciones. Estos resultados sugieren que las puntuaciones analíticas de los correctores tienden a ser más consistentes que las puntuaciones holísticas.

Fig. 10.4b. Evaluaciones holísticas y evaluaciones analíticas en el POST de acuerdo con las variables género y situación laboral de los correctores (i.e. universidad)



*La abreviatura HOLANDOS se refiere a las evaluaciones holísticas y analíticas del POST

Seguidamente, examinaremos el grado de significación estadística de las tendencias anteriormente apuntadas a través del análisis de las pruebas de contraste intra-sujetos (ver Tabla 10.22).

Tabla 10.22. Pruebas de contrastes intra-sujetos: evaluaciones holísticas y analíticas en el POST de acuerdo con el género y situación laboral de los correctores

Pruebas de contrastes intra-sujetos

Medida: MEASURE_1

Fuente	HOLANDOS	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
HOLANDOS	Lineal	1233,111	1	1233,111	106,908	,000
HOLANDOS * SX	Lineal	9,033	1	9,033	,783	,384
HOLANDOS * PROF	Lineal	2,346	1	2,346	,203	,656
HOLANDOS * SX * PROF	Lineal	67,805	1	67,805	5,879	,022
Error(HOLANDOS)	Lineal	311,427	27	11,534		

La abreviatura HOLANDOS representa las puntuaciones holísticas y analíticas en el POST. Las abreviaturas SX y PROF se refieren a las variables género y situación laboral respectivamente.

Los datos indican que la diferencia entre las puntuaciones holísticas y analíticas del POST son significativas ($F_{1,31}$ = 106,908; p = 0,000). Es decir, las puntuaciones holísticas y analíticas difieren de forma significativa, lo que nos lleva a inferir que el método de corrección que se utilice para evaluar los ensayos afectará notablemente las puntuaciones finales que se obtengan (ver Hamp-Lyons, 1991).

Asimismo, cabe resaltar que, en esta ocasión, las diferencias que se observan son significativas en cuanto al género y situación laboral de los correctores ($F_{1,31}$ = 5,879; p = 0,022). Por consiguiente, estas diferencias no son fortuitas sino que se atribuyen a estas dos últimas variables.

Así pues, estos datos confirman que hay interacción entre las variables género y situación laboral de los correctores y las diferencias que se observan en las evaluaciones holísticas y analíticas del POST.

10.2.3. Pruebas de significación

En los siguientes subepígrafes examinaremos las pruebas de significación en: las evaluaciones globales: holísticas y analíticas (PRE y POST), las evaluaciones analíticas específicas (PRE y POST) y en los errores contenidos en el discurso y en frases descontextualizadas.

10.2.3.1. Evaluaciones globales: holísticas y analíticas (PRE y POST)

El análisis de la varianza (ANOVA) nos ha permitido estudiar la influencia de las variables género y situación laboral de los correctores en los distintos tipos de evaluación. Los resultados obtenidos con ANOVA confirman la interacción entre las evaluaciones holísticas y las analíticas del POST. Procede, por tanto, que entremos a examinar las pruebas de significación a través del *T-test* en los distintos pares de muestras relacionadas que puedan aumentar la información estadística.

Se presenta, en primer lugar, la tabla de muestras relacionadas de las evaluaciones globales que es un resumen de los análisis de varianza (ANOVA) llevados a cabo hasta ahora (Tabla 10.23). En esta tabla vemos de una manera más plástica el rango de diferencias entre los distintos pares: HG1-HG2 ($\bar{\times}$ = 2,9688); AG1-AG2 ($\bar{\times}$ = 5,2995); HG1-AG1 ($\bar{\times}$ =-10,5625) y HG2-AG2 ($\bar{\times}$ = -8,8295).

Tabla 10.23. Diferencia de puntuaciones entre las puntuaciones holísticas y analíticas (PRE y POST)

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	HG1 - HG2	2,9688	9,10684	1,60988	-,3146	6,2521	1,844	31	,075
Par 2	AG1 - AG2	5,2995	7,02990	1,26261	2,7210	7,8781	4,197	30	,000
Par 3	HG1 - AG1	-10,5625	7,60894	1,34508	-13,3058	-7,8192	-7,853	31	,000
Par 4	HG2 -AG2	-8,8295	5,08585	,91345	-10,6950	-6,9640	-9,666	30	,000

*Las abreviaturas HG1 y HG2 se refieren a las evaluaciones holísticas en la primera y en la segunda ocasión respectivamente. De igual modo, las abreviaturas AG1 y AG2 se refieren a las evaluaciones analíticas en la primera y en la segunda ocasión respectivamente.

Llama la atención el hecho de que el par HG1 – HG2 (evaluaciones holísticas PRE y POST respectivamente), con la desviación típica más alta (DT = 9,1068), presente la diferencia de medias más baja: 2,9688. Además, este par de muestras es el único que no alcanza el nivel de significación (t = 1,844 y p = 0,075).

Estos datos indican que, a pesar de la gran dispersión de puntuaciones que se observa, las diferencias entre las evaluaciones holísticas del PRE (HG1) y del POST (HG2) no son significativas.

10.2.3.2. Evaluación específica: evaluaciones analíticas PRE y POST

Una vez examinadas las pruebas de significación (*T-tests*) en las evaluaciones globales, vamos a llevar a cabo los *T-tests* para las muestras relacionadas de las categorías analíticas específicas.

La Tabla 10.24 nos describe los estadísticos descriptivos de las distintas categorías analíticas en el PRE y en el POST.

Tabla 10.33. Estadísticos descriptivos de las evaluaciones analíticas (PRE y POST)

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	AGContenido1	30,6875	32	5,59630	,98930
	AGContenido2	28,0313	32	4,56130	,80633
Par 2	AGOrganización1	30,1875	32	5,56161	,98316
	AGOrganización2	27,5313	32	4,33280	,76594
Par 3	AGGramática1	27,3750	32	4,36075	,77088
	AGGramática2	23,3438	32	3,46046	,61173
Par 4	AGVocabulario1	28,5938	32	4,46390	,78911
	AGVocabulario2	25,7813	32	3,73073	,65951
Par 5	AGRegistro1	30,8065	31	4,28501	,76961
	AGRegistro2	28,8387	31	3,13153	,56244
Par 6	AGMecánica1	29,3750	32	4,93016	,87154
	AGMecánica2	26,9063	32	3,78758	,66956
Par 7	AGPresentación1	30,7500	32	5,03536	,89013
	AGPresentación2	29,2500	32	4,75191	,84003

La abreviatura AG delante de las distintas categorías analíticas indica la evaluación analítica de dichas categorías en el PRE (1) y en el POST (2).

Como vemos, las medias de las puntuaciones analíticas son todas ellas más bajas en las categorías analíticas evaluadas en el POST. Este resultado confirma las tendencias hasta ahora observadas que indican que

el conjunto de correctores se muestra más severo en la evaluación de las categorías analíticas de los ensayos en el POST que en el PRE.

Conviene destacar que la *gramática* es la categoría que registra las medias más bajas en el PRE y en el POST, lo que nos lleva a inferir que dicha categoría es la que mayormente se penaliza. Estos datos demuestran que el perfil docente de los correctores no es el del profesor comunicativo tal y como aseguran los correctores en sus planteamientos teóricos (ver capítulo 9, subepígrafe 9.2.1.). Los datos sugieren que los correctores muestran un perfil docente más bien sistemático y dominante en su actuación práctica.

Las pruebas de significación nos facilitan, además de los estadísticos descriptivos, las correlaciones entre las distintas muestras relacionadas que se comentan a continuación (Tabla 10.25).

Tabla 10.25. Correlaciones de las evaluaciones analíticas (PRE Y POST)

Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	AGConte1 y AGConte2	32	,409	,020
Par 2	AGOrgan1 y AGOrgan2	32	,567	,001
Par 3	AGGram1 y AGGram2	32	,282	,118
Par 4	AGVocab1 y AGVocab2	32	,595	,000
Par 5	AGRegis1 y AGRegis2	31	,621	,000
Par 6	AGMecan1 y AGMecan2	32	,529	,002
Par 7	AGPrese1 y AGPrese2	32	,491	,004

La abreviatura AG delante de las distintas categorías analíticas indica la evaluación analítica de dichas categorías en el PRE (1) y en el POST (2)

Los resultados de la Tabla 10.25 indican que la relación entre las categorías analíticas es significativa en todos los casos salvo en el caso de

la categoría *gramática* (p= 0,118). Estos datos sugieren que la evaluación de la *gramática* registra una mayor variabilidad de puntuaciones que el resto de las categorías analíticas. De ahí se deduce que los correctores aplican los criterios más arbitrarios y más estrictos a la evaluación de esta última categoría. De nuevo se contradicen los planteamientos teóricos que defienden los correctores en el cuestionario dado que estos últimos describen su faceta evaluadora como de experta y consecuente (ver capítulo 9, epígrafe 9.4).

Por último, las pruebas de significación (*T*-tests) examinan las diferencias entre las medias de las distintas muestras relacionadas. Esta información se presenta en la Tabla 10.26.

Tabla 10.26. Diferencia de puntuaciones entre las puntuaciones analíticas (PRE y POST)

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	AGContenido1 - AGContenido2	2,6563	5,59152	,98845	,6403	4,6722	2,687	31	,011
Par 2	AGOrganización1 - AGOrganización2	2,6563	4,72884	,83595	,9513	4,3612	3,178	31	,003
Par 3	AGGramática1 - AGGramática2	4,0313	4,74161	,83821	2,3217	5,7408	4,809	31	,000
Par 4	AGVocabulario1- AGVocabulario2	2,8125	3,74543	,66210	1,4621	4,1629	4,248	31	,000
Par 5	AGRegistro1 - AGRegistro2	1,9677	3,39101	,60904	,7239	3,2116	3,231	30	,003
Par 6	AGMecánica1- AGMecánica2	2,4688	4,34767	,76857	,9012	4,0363	3,212	31	,003
Par 7	AGPresentación1 - AGPresentación2	1,5000	4,94486	,87414	-,2828	3,2828	1,716	31	,096

La abreviatura AG delante de las distintas categorías analíticas indica la evaluación analítica de dichas categorías en el PRE (1) y en el POST (2)

Como se observa, las diferencias entre las puntuaciones de las categorías analíticas en el PRE y en el POST son significativas en todos los casos salvo en el caso de la categoría *presentación*, que se halla en el umbral de la significación (p= 0,096).

Cabe resaltar que la categoría *gramática* presenta el nivel de significación mayor (t= 4,809; p= 0,000), a pesar de ser la única categoría que no correlaciona consigo misma.

10.2.3.3. Evaluación de los errores en el discurso y en frases descontextualizadas

Nuestro siguiente paso consiste en la aplicación de las pruebas de significación (*T-tests*) a la evaluación de los errores. Con ello se pretende averiguar si las diferencias que se observan entre las evaluaciones de los distintos elementos del ensayo que se hallan en el discurso (i.e. dentro de un contexto) y la evaluación de los errores contenidos en frases descontextualizads[3] son significativas.

Antes de iniciar el análisis de los datos, conviene recordar que las escalas de valoración que se aplicaron a las categorías analíticas contenidas en el discurso (i.e. los ensayos) y a los errores relacionados con dichas categorías y contenidos en 14 frases descontextualizadas sigue un orden inverso. Esto es, en una escala de Likert de 5 puntos, el valor 5 representa la máxima puntuación (i.e. puntuación muy buena) en la

[3] Ver cuestionario (Apéndice), sección IV para examinar los errores contenidos en frases descontextualizadas.

evaluación de las categorías analíticas contenidas en los ensayos. Por el contrario, el valor 5 de la escala de Likert indica un error clasificado como muy grave en la evaluación de los errores contenidos en las frases descontextualizadas.

El cambio de escala pretendía evitar que los correctores asociaran las evaluaciones de las categorías analíticas con la de los errores contenidos en las frases descontextualizadas. No obstante, con el fin de facilitar la lectura de los datos y evitar confusiones en la interpretación de los resultados, en este análisis, hemos optado por transformar la escala de valoración de los errores contenidos en las frases descontextualizadas para adaptarla a la escala utilizada en el análisis de las categorías analíticas.

Siguiendo con el protocolo establecido en el estudio de las pruebas de significación (*T-tests*), examinaremos, en primer lugar, los estadísticos descriptivos de las distintas muestras relacionadas (ver Tabla 10.27).

Tabla 10.27. Estadísticos descriptivos entre las categorías contenidas en el discurso y los errores contenidos en frases aisladas

Estadísticos de muestras relacionadas

		Media	N	Desviación típ.	Error típ. de la media
Par 1	AGContendido	3,0688	32	,55963	,09893
	T.Contenido	3,0000	32	1,06256	,18784
Par 2	AGOrganización	3,0188	32	,55616	,09832
	T.Organización	2,5156	32	,78786	,13927
Par 3	AGGramática	2,7375	32	,43607	,07709
	T.Gramática	1,8125	32	,61892	,10941
Par 4	AGVocabulario	2,8677	31	,45121	,08104
	T.Vocabulario	2,5645	31	,70406	,12645
Par 5	AGRegistro	2,9969	32	,63423	,11212
	T.Registro	3,0000	32	1,11442	,19700
Par 6	AGMecánica	2,9375	32	,49302	,08715
	T.Mecánica	3,3281	32	1,11159	,19650
Par 7	AGPresentación	3,0774	31	,51167	,09190
	T.Presentación	2,5323	31	1,07963	,19391

Las abreviaturas AG y T que se observan al principio de cada una de las categorías analíticas representan las categorías analíticas evaluadas dentro del discurso y las categorías analíticas evaluadas en las frases descontextualizadas respectivamente

Los datos indican que las medias de las categorías analíticas contenidas en el discurso (i.e. los ensayos) son superiores a las medias de los errores representativos de las categorías analíticas contenidas en las frases descontextualizadas. Esto es así en todos los casos salvo en las categorías de *registro* ($\bar{\times}$ = 2,9969 vs. $\bar{\times}$ = 3,0000 en errores contextualizados y descontextualizados respectivamente) y *mecánica* ($\bar{\times}$ = 2,9375 vs. $\bar{\times}$ = 3,3281 en errores contextualizados y descontextualizados respectivamente).

Estos datos nos permiten inferir que, en general, los errores descontextualizados se penalizan de forma más estricta que los errores contenidos en el discurso (i.e. los ensayos).

Asimismo, se observa que las puntuaciones medias de las categorías analíticas de los ensayos tienden a centralizarse en torno al valor neutro (i.e. valor 3) de la escala de Likert[4]. Las categorías de *vocabulario* ($\bar{\times}$ = 2,8677) y *gramática* ($\bar{\times}$= 2,7375) son las que registran las puntuaciones medias más bajas.

La *gramática* es, sin duda, la categoría que se evalúa de forma más estricta tanto dentro como fuera del discurso y, por consiguiente, es la que muestra las puntuaciones medias más bajas ($\bar{\times}$ = 2,7375 y $\bar{\times}$= 1,8125 en errores contextualizados y descontextualizados respectivamente). Por el contrario, el *contenido* ($\bar{\times}$ = 3,0688) es la categoría analítica que obtiene una media más alta después de la *presentación* ($\bar{\times}$ = 3,0774).

A la luz de estos resultados se puede inferir que el *contenido* se evalúa de forma positiva, especialmente dentro del discurso. Por su parte, la *gramática* se juzga como un elemento negativo tanto dentro como fuera del discurso.

A continuación, se examinarán los resultados de las correlaciones de las muestras relacionadas (Tabla 10.28).

[4] Los valores de la escala de Likert no están definidos cuantitativamente sino que representan distintos intervalos que definen estimaciones personales en una escala de valores que va de malo a muy bueno, etc..

Tabla 10.28. Correlaciones entre las categorías analíticas en el discurso y los errores contenidos en frases individuales

Correlaciones de muestras relacionadas

		N	Correlación	Sig.
Par 1	AGContenido y T.Contenido	32	,255	,159
Par 2	AGOrganización y T.Organización	32	,058	,752
Par 3	AGGramática y T.Gramática	32	,158	,387
Par 4	AGVocabulario y T.Vocabulario	31	,017	,927
Par 5	AGRegistro y T.Registro	32	,404	,022
Par 6	AGMecánica y T.Mecánica	32	,404	,022
Par 7	AGPresentación y T.Presentación	31	-,128	,491

Las abreviaturas AG y T que se observan al principio de cada una de las categorías analíticas representan las categorías analíticas evaluadas dentro del discurso y las categorías analíticas evaluadas en las frases descontextualizadas respectivamente

Los resultados de la Tabla 10.28 indican que la relación entre las puntuaciones de las categorías analíticas incluidas en el discurso y las puntuaciones de los errores contenidos en las frases descontextualizadas no es significativa, salvo en las categorías de *registr*o (p = 0,022) y *mecánica* (p = 0,022). Estos datos sugieren que las puntuaciones de las categorías analíticas contenidas dentro y fuera del discurso son independientes.

Finalmente, las pruebas de significación (*T-tests*) nos permiten averiguar el grado de significación entre las medias de las muestras relacionadas (Tabla 10.29).

Tabla 10.29. Diferencia entre las puntuaciones de las categorías analíticas contenidas en el discurso y los errores contenidos en frases individuales

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	AGContenido - T.Contenido	,0688	1,06724	,18866	-,3160	,4535	,364	31	,718
Par 2	AGOrganización - T.Organización	,5031	,93756	,16574	,1651	,8412	3,036	31	,005
Par 3	AGGramática - T.Gramática	,9250	,69839	,12346	,6732	1,1768	7,492	31	,000
Par 4	AGVocabulario - T.Vocabulario	,3032	,82965	,14901	-,0011	,6075	2,035	30	,051
Par 5	AGRegistro - T.Registro	-,0031	1,03596	,18313	-,3766	,3704	-,017	31	,986
Par 6	AGMecánica - T.Mecánica	-,3906	1,01802	,17996	-,7577	-,0236	-2,17	31	,038
Par 7	AGPresentación- T. Presentación	,5452	1,25269	,22499	,0857	1,0047	2,423	30	,022

Las abreviaturas AG y T que se observan al principio de cada una de las categorías analíticas representan las categorías analíticas evaluadas dentro del discurso y las categorías analíticas evaluadas en las frases descontextualizadas respectivamente

Como vemos, los resultados de la Tabla 10.29 indican diferencias significativas en la evaluación de los elementos de *mecánica* (p = 0,038), *presentación* (p = 0,022) y, especialmente, *organización* (p = 0,005) y *gramática* (p = 0,000). Este último elemento es el que registra un mayor grado de significación (t = 7,492). Dicho resultado coincide con las diferencias observadas en la evaluación de los elementos analíticos específicos que indican que la *gramática* presenta el mayor nivel de significación (ver Tabla 10.26). Por consiguiente, se deduce que este elemento es el que se juzga de forma más arbitraria. Conviene también

señalar que el *vocabulario* se halla en el umbral de la significación (p = 0,051).

A la luz de estos datos se puede afirmar que el contexto ayuda a minimizar la importancia que se les concede a los errores evaluados en frases descontextualizadas (ver Ludwig 1982; Davies 1983; Santos 1988; Dordick 1996). Como se ha podido comprobar, los errores descontextualizados tienden a penalizarse con mayor rigor, especialmente los errores de *gramática* y *organización*.

10.2.4. Variables que determinan las puntuaciones holísticas

La técnica estadística de la regresión múltiple nos permite determinar las variables individuales o combinación de variables que explican una respuesta concreta y predicen su valor.

En nuestro estudio, la técnica de la regresión múltiple va a posibilitarnos la identificación de los elementos de los ensayos que contribuyen de forma significativa a la asignación de las puntuaciones holísticas según su nivel de dominio lingüístico (i.e. bajo, medio y alto).

Como se recordará, los ensayos 1 (h1), 5 (h5) y 10 (h10) se escogieron como ensayos representativos de los dominios de la lengua bajo, medio y alto respectivamente. En los ensayos de nivel medio (h5) y alto (h10) no se encontraron rectas de regresión significativas, por esa razón no se han incluido las Tablas de estos resultados. Tan sólo en el ensayo con un dominio de la lengua bajo (h1) los resultados fueron significativos (ver

Tablas 10.30 y 10.31). Este último caso es el que comentamos a continuación.

La Tabla 10.30 nos indica el coeficiente de determinación de R^2 (0,733): i.e. la varianza de la variable dependiente que explican las variables predictoras.

Tabla 10.30. Resumen del modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,856[a]	,733	,652	,661

a. Variables predictoras: (Constante), presentación, organización, gramática, registro, vocabulario, mecánica, contenido

La Tabla 10.31, por otro lado, nos muestra que la recta de regresión es significativa (F = 9,038; p = 0,000).

Tabla 10.31.Anova

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	27,679	7	3,954	9,038	,000[a]
	Residual	10,063	23	,438		
	Total	37,742	30			

a. Variables predictoras: (Constante), presentación, organización, gramática, registro, vocabulario, mecánica, contenido

b. Variable dependiente: hol.1

Por último, la Tabla 10.32 identifica las categorías de *organización* (p = 0,012), *vocabulario* (p = 0,041) y *gramática* (p = 0,064), esta última se halla en el umbral de la significación, como las variables explicativas de las evaluaciones holísticas del ensayo 1 (h1).

Tabla 10.32.Coeficientes.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	,270	,531		,509	,616
	contenido	3,792E-02	,204	,033	,186	,854
	organización	,686	,252	,564	2,728	,012
	gramática	,658	,338	,292	1,945	,064
	vocabulario	-,664	,307	-,358	-2,160	,041
	registro	,162	,181	,122	,894	,381
	mecánica	,315	,211	,256	1,491	,149
	presentación	,127	,175	,098	,727	,474

a. Variable dependiente: hol.1

Estos datos nos permiten inferir que en la evaluación de los ensayos con un nivel de dominio de la lengua bajo los correctores dirigen su atención principalmente a los aspectos formales, discursivos y léxicos del texto. Estos aspectos se relacionan, principalmente, con la *forma* del ensayo (ver Homburg 1984; Ziv N. 1984; Zamel 1985; Amengual y Herrera 2000). Por su parte, el *contenido* y la expresión de ideas apenas se consideran.

10.2.5. Análisis de los Comentarios positivos y negativos de los ensayos

Para completar nuestro estudio sobre la fiabilidad inter-corrector, decidimos analizar los comentarios positivos y negativos que los correctores destacan en los ensayos. El análisis de estos comentarios va a permitirnos ampliar la información obtenida a través de las técnicas estadísticas aplicadas hasta el momento para identificar los componentes específicos del ensayo que determinan las puntuaciones holísticas finales.

En primer lugar, se examinarán los comentarios positivos y negativos realizados por los correctores a nivel general. Posteriormente se analizarán los comentarios positivos y negativos según el nivel de dominio lingüístico de los ensayos.

10.2.5.1. Comentarios positivos y negativos de los ensayos a nivel general

El estudio de los comentarios positivos y negativos de los ensayos se ha realizado a través de la utilización de tablas de frecuencia. Estas tablas nos permiten comparar la frecuencia o el número total de comentarios positivos y negativos relacionados con las distintas categorías analíticas[5] que se destacan en la evaluación holística PRE y POST (ver Tablas 10.33 y 10.34).

Tabla 10. 33. Frecuencia de comentarios positivos y negativos en la evaluación holística (PRE)

	Comentarios Positivos	Comentarios Negativos	TOTAL
Contenido	22	15	37
Organización	86	24	110
Gramática	114	159	273
Vocabulario	21	25	46
Registro	9	5	14
Mecánica	0	19	19
Presentación	2	9	11
TOTAL	254	256	510

[5] Los comentarios de los correctores se relacionan con las siete categorías analíticas diseñadas en este estudio: *contenido*, *organización*, *gramática*, *vocabulario*, *registro*, *mecánica* y *presentación*

Tabla10.34.Frecuencia de comentarios positivos y negativos en la evaluación holística (POST)

	Comentarios Positivos	Comentarios Negativos	TOTAL
Contenido	22	9	31
Organización	74	26	100
Gramática	106	190	296
Vocabulario	14	16	30
Registro	5	5	10
Mecánica	1	15	16
Presentación	8	7	15
TOTAL	230	268	498

Como se aprecia, las Tabla 10.33 y 10.34 muestran, en general, resultados similares en cuanto a las categorías positivas y negativas de los ensayos que los correctores destacan en las evaluaciones holísticas PRE y POST.

Los datos indican que las dos categorías principales que centran el interés de los correctores son la *gramática* (n = 273 y n = 296 en el recuento total obtenido en el PRE y el POST respectivamente) y la *organización* (n = 110 y n = 100 en el recuento total obtenido en el PRE y el POST respectivamente).

Es interesante subrayar que la *organización* se juzga como un elemento fundamentalmente positivo tanto en el PRE (n = 86 vs. n = 24 comentarios positivos y negativos respectivamente) como en el POST (n = 74 vs. n = 26 comentarios positivos y negativos respectivamente). Por el contrario, los datos indican que la *gramática* se identifica como un elemento clave tanto en la formación de juicios positivos (n = 114 y n = 106 en el PRE

y en el POST respectivamente) como negativos (n = 159 y n = 190 en el PRE y en el POST respectivamente), especialmente en este último caso. De ahí se deduce que, la *gramática* afecta a los correctores de forma decisiva en la evaluación holística de los ensayos. Como vemos, se premia la utilización de una buena *gramática* y se penaliza duramente la utilización de una *gramática* deficiente.

El tercer elemento que los correctores destacan es el *vocabulario*. Este elemento se evalúa tanto de forma positiva (n = 21 y n = 14 en el PRE y POST respectivamente) como de forma negativa (n = 25 y n = 16 en el PRE y POST respectivamente) en los ensayos, si bien su influencia en las evaluaciones holísticas es muy inferior a la que ejerce la *gramática* o la *organización*.

Por lo que respecta al *contenido* (n = 22 vs. n = 15 comentarios positivos y negativos respectivamente en el PRE; n = 22 vs. n = 9 comentarios positivos y negativos respectivamente en el POST), cabe decir que éste se considera un aspecto fundamentalmente positivo en la evaluación de los ensayos. No obstante, como se puede comprobar, su importancia es relativa.

Es interesante observar la reacción negativa que despiertan los errores de *mecánica* en los correctores. La *mecánica* se valora exclusivamente de forma negativa en el PRE y en el POST (n = 0 vs. n = 19 comentarios positivos y negativos respectivamente en el PRE; n = 1 vs. n = 15 comentarios positivos y negativos respectivamente en el POST). Estos datos sugieren que la presencia de una *mecánica* deficiente constituye un

elemento sobresaliente que llama la atención de los correctores en la emisión de juicios negativos del ensayo.

Por último, cabe señalar que, si bien la frecuencia de comentarios positivos (n = 254) y negativos (n = 256) que se registra en el PRE es similar, en el POST, se observa una mayor disparidad. Por una parte, disminuye el número de comentarios positivos y, por otra, se incrementa el número de comentarios negativos (n = 230 vs. n = 268 en el PRE y POST respectivamente). Estos datos coinciden con el mayor grado de exigencia que demuestran los correctores en las evaluaciones holísticas del POST.

10.2.5.2. Comentarios positivos y negativos según el dominio lingüístico de los ensayos

Nuestro siguiente paso, consiste en analizar los diferentes comentarios positivos y negativos que los correctores destacan en los ensayos representativos de los tres niveles diferentes de dominio lingüístico: bajo (h1), medio (h5) y alto (h10). Con ello queremos averiguar si la frecuencia de comentarios positivos y negativos sobre cada una de las categorías analíticas varía y se adecua al nivel lingüístico de los ensayos.

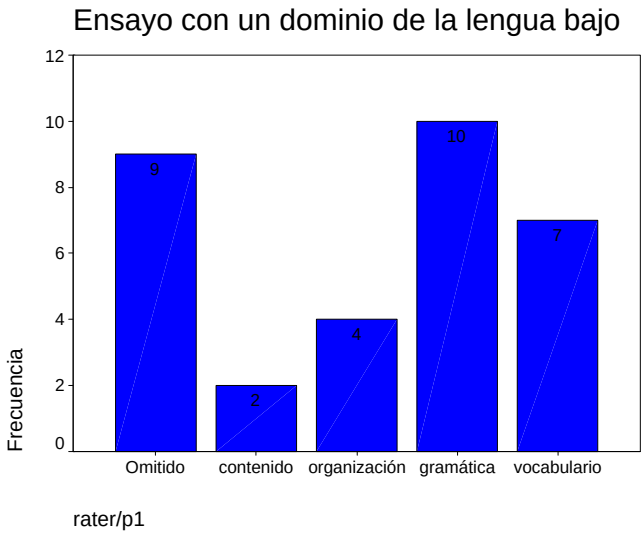
Conviene mencionar en este punto que, además de las siete categorías analíticas analizadas en nuestro estudio (i.e. *contenido*, *organización*, *gramática*, *vocabulario*, *registro*, *mecánica* y *presentación*), se incluye, en esta ocasión, una nueva categoría denominada *omitido*. Esta categoría recoge los comentarios de los correctores que indican que no

encuentran ningún elemento positivo o negativo que destacar en el ensayo en cuestión. La inclusión de dicha categoría puede resultar interesante ya que la omisión de comentarios positivos o negativos se considera relevante en la interpretación de los resultados (Gamaroff, 2000).

La técnica estadística que se utilizará para medir la frecuencia de los comentarios es el histograma. De este modo, obtendremos una representación gráfica de los resultados obtenidos en los ensayos con los dominios de la lengua: bajo (h1), medio (h5) y alto (h10).

Empezaremos nuestro análisis examinando los histogramas del ensayo con un dominio de la lengua bajo (h1). La Fig. 10.5a muestra la frecuencia de los comentarios positivos para cada una de las categorías analíticas que se destacan en este ensayo.

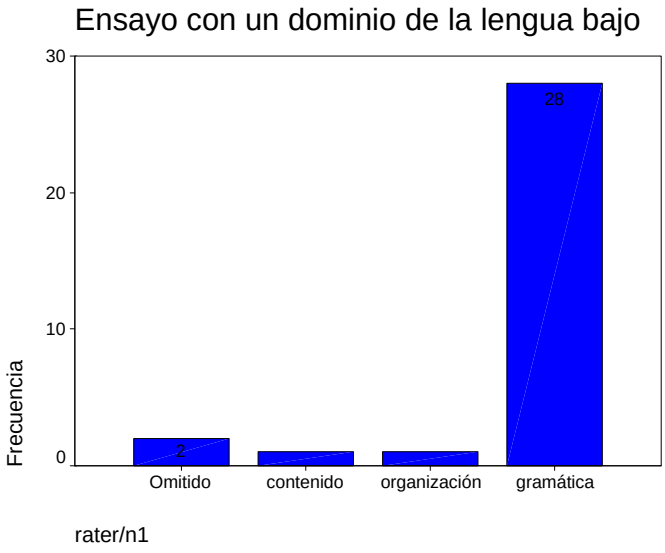
Fig. 10.5a. Comentarios positivos del ensayo con un nivel de dominio lingüístico bajo (h1)



Como vemos en el gráfico, los elementos positivos que se subrayan son la *gramática* (n = 10) y el *vocabulario* (n = 7). El dominio de estos dos últimos aspectos se evalúa de forma positiva. Por su parte, la categoría *omitido* registra una frecuencia elevada (n = 9), lo que significa que un gran número de correctores no observó ninguna cualidad positiva en el ensayo digna de mención. El *contenido* (n = 2) también se evalúa positivamente si bien su importancia es muy baja.

La Fig. 10.5b representa las distintas frecuencias de los comentarios negativos que se realizan sobre el ensayo con un nivel lingüístico bajo (h1).

Fig. 10.5b. Comentarios negativos del ensayo con un nivel de dominio lingüístico bajo (h1)



Los datos muestran claramente que la *gramática* (n = 28) polariza la mayoría de las respuestas de los correctores. De ahí se deduce que la *gramática* juega un papel esencial en la evaluación negativa del ensayo.

Como era previsible, la categoría *omitido* (n = 2) experimenta un notable descenso. Este último dato indica que tan sólo dos correctores alegaron no encontrar ningún aspecto negativo digno de mención en este ensayo.

A continuación, pasamos a comentar los histogramas del ensayo con un dominio de la lengua medio (h5). El primer histograma (Fig.10.6a) nos muestra la frecuencia de los comentarios positivos del ensayo (h5).

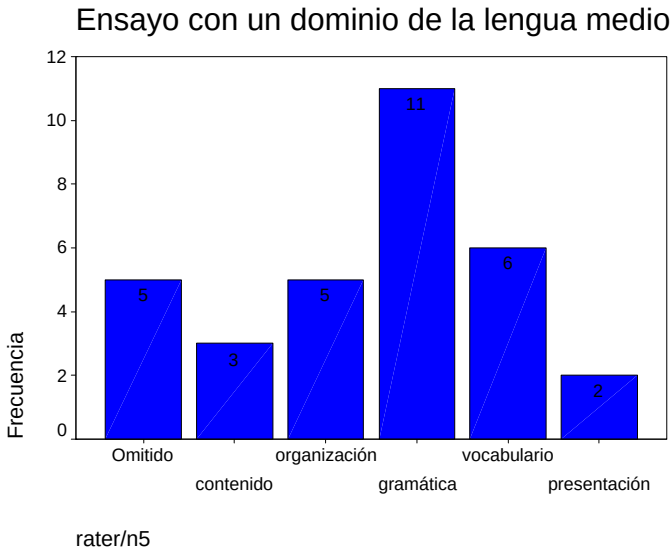
Fig. 10.6a. Comentarios positivos del ensayo con un nivel de dominio lingüístico medio (h5)

En esta ocasión, la *organización* (n = 17) es la categoría que centra el interés principal de los correctores. Esta categoría se evalúa de forma muy positiva. La *gramática* (n = 6) se sitúa en segundo lugar en orden de importancia junto con la categoría *omitido* (n = 6). Esta última categoría registra un número de respuestas inferior al que registraba el ensayo con un dominio lingüístico bajo (h1) ya que los correctores encuentran un número

mayor de cualidades positivas a destacar en el ensayo con un dominio de la lengua medio. Como se ve, el *contenido* (n = 3), si bien se contempla, vuelve a despertar un interés muy moderado.

El siguiente histograma (Fig. 10.6b) representa la frecuencia de los comentarios negativos del ensayo con un dominio de la lengua medio.

Fig. 10.6b. Comentarios negativos del ensayo con un nivel de dominio lingüístico medio (h5)



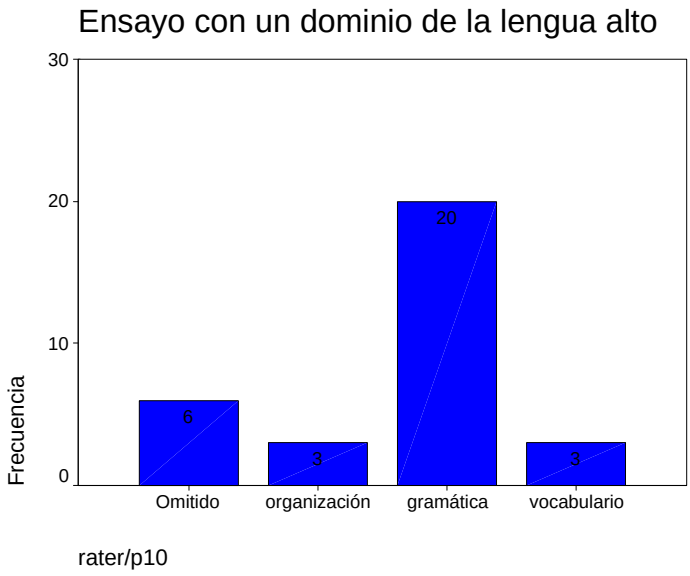
Como podemos comprobar, en este ensayo, a diferencia de lo que ocurre en el ensayo con un dominio lingüístico bajo en el que se destaca casi de forma exclusiva la *gramática*, los correctores atienden a un mayor número de categorías. No obstante, la *gramática* (n = 11) vuelve a ser el elemento negativo principal que se subraya. A este elemento, le siguen el *vocabulario* (n = 6) y la *organización* (n = 5) en orden de importancia. Cabe destacar, la incorporación de la *presentación* (n = 2) como elemento

negativo. Esta categoría recibe una atención similar a la de *contenido* (n = 3), que como vimos no tiende a evaluarse negativamente.

Por último, pasamos a examinar los comentarios positivos y negativos del ensayo con un dominio de la lengua alto (h10).

La Fig. 10.7a. nos muestra la frecuencia de los comentarios positivos realizados sobre el ensayo con un dominio lingüístico alto.

Fig. 10.7a. Comentarios positivos del ensayo con un nivel de dominio lingüístico alto (h10)

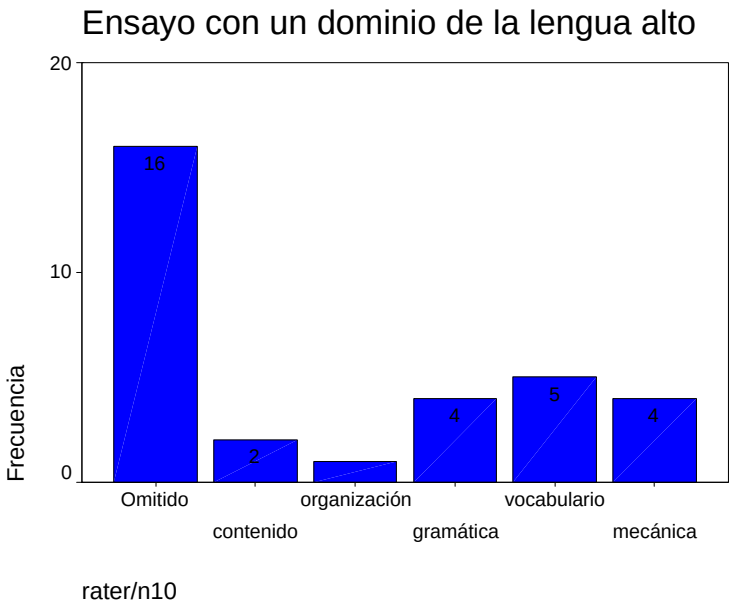


Los resultados indican que la *gramática* (n = 20) continúa polarizando las respuestas positivas de los correctores. Probablemente, esto se deba a que este tipo de ensayos suele incluir estructuras gramaticales de cierta complejidad que los correctores evalúan de forma positiva. El resto de las categorías recibe una importancia mucho menor: *vocabulario* (n = 3) y *organización* (n = 3).

Llama la atención que el *contenido* no se considere en la evaluación de este ensayo. Asimismo, resulta algo sorprendente el número de respuestas que registra la categoría *omitido* (n = 6), ya que, paradójicamente, esto significa que los correctores no encuentran ninguna cualidad positiva digna de mención en este ensayo. Estos datos sugieren que los correctores se muestran reticentes a destacar los aspectos positivos de los ensayos aún cuando estos últimos se consideran *buenos* ensayos.

Por último, la Fig. 10.7b nos muestra los comentarios negativos del ensayo con un dominio de la lengua alto.

Fig. 10.7b. Comentarios negativos del ensayo con un dominio lingüístico alto (h10)



Si comparamos las respuestas de esta figura (Fig. 10.7b) con las de la figura anterior (Fig.10.7a), veremos que el número de categorías

negativas que se incluye es superior al de categorías positivas, tal y como ocurría en el ensayo con un dominio lingüístico medio. Así pues, es interesante observar que a medida que aumenta el dominio lingüístico del ensayo, aumenta el número de categorías que se incluyen en la evaluación negativa.

Como era de prever, la categoría *omitido* (n = 16) es la que registra una frecuencia mayor de respuestas ya que los correctores no observan cualidades negativas que deban destacarse en este ensayo.

El resto de las categorías que se incluyen registran una frecuencia mucho menor que la que se observaba en los ensayos con un dominio lingüístico bajo y medio. El *vocabulario* (n = 5) es el elemento que se penaliza en mayor medida seguido muy de cerca por la *gramática* (n = 4) y la *mecánica* (n = 4) en orden de importancia.

Cabe resaltar que la *mecánica* aparece por primera vez como un elemento negativo. Este dato es interesante ya que corrobora la idea que apuntábamos anteriormente en el análisis de los comentarios generales de los ensayos (ver epígrafe 10.2.5.1.). Es decir, los errores de *mecánica* son elemento sobresalientes, especialmente en los ensayos con un dominio lingüístico alto y, por ello, se tienden a penalizar.

Probablemente, la importancia de los errores de *mecánica* se minimiza en los ensayos con un nivel lingüístico inferior ya que entonces los correctores centran su atención en aquellos aspectos del ensayo que se consideran más importantes desde el punto de vista comunicativo como, por

ejemplo, una mala *gramática* o una mala *organización* que puedan dificultar la comprensión del texto.

En resumen, los resultados indican que los correctores, en general, parecen adecuar la provisión de *feedback* al nivel lingüístico de los ensayos. No obstante, se observan ciertas pautas comunes. Así, la *gramática* polariza la mayor parte de las respuestas de los correctores y se considera un elemento clave tanto en la emisión de juicios positivos como negativos en la evaluación de los ensayos independientemente de su nivel de dominio lingüístico.

Por su parte, la *organización* destaca como un elemento fundamentalmente positivo, especialmente en los ensayos con un dominio de la lengua medio. La *mecánica*, sin embargo, se penaliza casi de forma exclusiva en los ensayos con un dominio de la lengua alto. Este elemento tiende a destacar o sobresalir en la evaluación negativa de los ensayos.

Cabe señalar la poca importancia que se le concede, en general, al *contenido* de los ensayos. Como vimos, este elemento se evalúa de forma positiva en los ensayos con un dominio lingüístico bajo y medio y de forma negativa en aquellos ensayos con un dominio de la lengua alto. Estos datos sugieren que el *contenido* de los ensayos afecta en menor medida las evaluaciones holísticas de los ensayos que los aspectos relacionados con la *forma* (ver Applebee 1981; Perkins 1983; Zamel 1985; Sweedler-Brown 1993).

Los resultados anteriores contradicen la opinión expresada por los correctores en la sección de evaluación de los aspectos problemáticos de los ensayos incluida en el cuestionario (ver capítulo 9, epígrafe 9.5). Como se recordará, la mayor objeción que los correctores destacan en la evaluación de los ensayos es la "carencia de conocimientos generales previos" por parte de los candidatos. Este aspecto se relaciona directamente con el *contenido* del ensayo. No obstante, en la práctica, los correctores parecen atender principalmente a los aspectos *formales* de los ensayos. Estos datos ponen una vez más de manifiesto las discrepancias que se observan entre los planteamientos teóricos y las actuaciones prácticas de los correctores.

Un último aspecto interesante que cabe mencionar, es que los correctores se muestran reacios a destacar las cualidades positivas de los ensayos, especialmente cuando los ensayos muestran un buen dominio de la lengua. Este comportamiento se constata en la siguiente Tabla (Tabla 10.35) extraída a partir de la información contenida en los histogramas analizados en el epígrafe anterior (10.2.5.2.). Puede que la visión teórica que poseen los correctores sobre su faceta evaluadora se relacione exclusivamente con el aspecto de localización y penalización de los errores y no con el de provisión de un sistema de retroalimentación positivo. Esta hipótesis explicaría los resultados del cuestionario que indican que los correctores de nuestro estudio se consideran expertos en la evaluación de los ensayos (capítulo 9, epígrafe 9.4).

La Tabla 10.35 incluye el cómputo total de los comentarios positivos y negativos de los ensayos representativos de los diferentes niveles de dominio lingüístico: bajo (h1), medio (h5) y alto (h10). La categoría *omitido* no se incluye esta vez en el recuento.

Tabla 10.35. Frecuencia de comentarios positivos y negativos según los distintos niveles de dominio de la lengua

Dominio de la lengua	**Comentarios positivos**	**comentarios negativos**
Bajo	23	32
Medio	26	27
Alto	26	16
TOTAL	75	75

Como era de prever, el número de comentarios negativos aumenta a medida que disminuye el dominio lingüístico de los ensayos: bajo (n =32), medio (n = 27) y alto (n =16). No obstante, la frecuencia de los comentarios positivos es muy similar en los tres ensayos (h1, h5 y h10).

Una lectura positiva que se desprende de estos datos es que los correctores tienden a subrayar los aspectos positivos de los ensayos con un dominio de la lengua bajo. Esto puede resultar muy alentador para los estudiantes menos preparados deseosos de encontrar algún aspecto positivo en sus ensayos que les anime a elaborar mejores producciones escritas.

No obstante, los correctores muestran cierta reticencia a la hora de destacar las cualidades positivas de un *buen* ensayo. Así, el reconocimiento

de las cualidades positivas de estos ensayos se evidencia fundamentalmente por la disminución del número de comentarios negativos.

10.3. Fiabilidad intra-corrector

Una vez finalizado el estudio de la fiabilidad inter-corrector, nuestro siguiente paso será analizar la fiabilidad intra-corrector o fiabilidad de las puntuaciones de un mismo corrector en diferentes ocasiones. El estudio de la fiabilidad intra-corrector se centrará en el análisis de las puntuaciones holísticas que asignaron cada uno de los correctores a los diez ensayos en las ocasiones PRE y POST. En un último subepígrafe (10.3.1) se comentarán los casos de correctores extremos (i.e. *extreme cases*) o marginales.

La fiabilidad intra-corrector se abordará, en primer lugar, desde los estadísticos de la correlación, del nivel de consistencia y del acuerdo entre las puntuaciones holísticas de los treinta y dos correctores en el PRE y el POST. La Tabla 10.36 nos resume los resultados obtenidos en este análisis.

Como vemos en la Tabla, los valores de la correlación entre un mismo corrector en las dos ocasiones (PRE y POST) son muy altos ya que la mayoría de ellos muestra una correlación $\geq$ 0,90. Cabe subrayar que estos valores son muy superiores a los que registran los correctores a nivel de grupo (0,5436).

Tabla 10.36. Análisis de la fiabilidad de cada uno de los correctores en las evaluaciones holísticas PRE y POST

RATERS	Correlación	Consistencia	Acuerdo absoluto
R1	0,9104	0,9531	0,8109
R2	0,9566	0,9702	0,9587
R3	0,8285	0,9045	0,7703
R4	0,6051	0,7502	0,7416
R5	0,8961	0,9378	0,9359
R6	0,8929	0,9434	0,9109
R7	0,6922	0,8095	0,6391
R8	0,8531	0,8407	0,8407
R9	0,9243	0,9566	0,9449
R10	0,8878	0,9147	0,8973
R11	0,7323	0,8151	0,8107
R12	0,9547	0,9767	0,9332
R10	0,6799	0,7967	0,8101
R14	0,8489	0,9176	0,9055
R15	0,7865	0,8756	0,8550
R16	0,9654	0,9822	0,9822
R17	0,7934	0,8617	0,8684
R18	0,9043	0,9487	0,8708
R19	0,9014	0,9481	0,9046
R20	0,5881	0,7333	0,5238
R21	0,9371	0,9661	0,8607
R22	0,8690	0,9267	0,8845
R23	0,9747	0,8350	0,8390
R24	0,9504	0,9745	0,9767
R25	0,8343	0,9089	0,9089
R26	0,8726	0,9319	0,9308
R27	0,9380	0,9579	0,8543
R28	0,9007	0,9350	0,9388
R29	0,9670	0,8957	0,9026
R30	0,9163	0,9438	0,9461
R31	0,9147	0,9532	0,9337
R32	0,8351	0,9058	0,9108

Asimismo, el coeficiente de correlación intraclase (CCI) que mide el nivel de consistencia y de acuerdo absoluto muestra valores semejantes e incluso superiores a los que registra el nivel de correlación ($\geq$ 0,90). De acuerdo con el criterio de Fleiss (1986) estos valores señalan un grado de concordancia que se define como *muy bueno*. A la luz de estos resultados se puede afirmar que, en general, los correctores muestran un alto grado de consistencia y fiabilidad en sus distintas actuaciones individuales.

No obstante, también se observan algunas excepciones. Así, el corrector 20 (R20) muestra un nivel de correlación (0,5881) y de acuerdo absoluto muy bajo (0,5238). Los correctores R4 (0,6051), R7 (0,6922) y R10 (0,6799) también muestran una consistencia moderada que se aleja de la fiabilidad perfecta que se halla en el 1.

Estos datos son ciertamente preocupantes ya que sugieren que la actuación de un mismo corrector se ve afectada por variables irrelevantes desde el punto de vista de la evaluación como son: las circunstancias personales (ver capítulo 9, epígrafe 9.3), los distintos estados anímicos (i.e. fatiga, cansancio, etc), etc. Dichos aspectos no se contemplan en los planteamientos teóricos pero sin duda tienen su relevancia en la práctica.

Nuestra siguiente etapa en la investigación de la fiabilidad intra-corrector será examinar las diferencias que se observan en el rango de las puntuaciones holísticas PRE y POST de cada uno de los treinta y dos correctores. La Tabla 10.37 nos muestra las diferencias entre las puntuaciones máximas y mínimas de los diversos correctores en las evaluaciones holísticas PRE y POST.

Tabla 10.37. Diferencia entre las puntuaciones máximas y mínimas (PRE y POST)

Puntos	**0**	**1**	**2**	**3**	**4**	**Diferencia Entre Medias** (PRE-POST)

Correctores						
R1[6]		1				1.8
R2		1				0.5
R3		1				1.9
R4	1					0.7
R5	1					-0.3*
R6	1					0.9
R7	1					-2
R8				1		-0.4
R9		1				0.6
R10		1				0.4
R11				1		-0.5
R12		1				0.7
R10		1				0.1
R14	1					0.6
R15			1			-0.8
R16	1					-0.2
R17			1			-0.3
R18		1				1.7
R19	1					0.8
R20		1				1.6
R21		1				1.3
R22		1				-1
R23					1	-0.4
R24	1					0.1
R25		1				-0.3
R26		1				-0.4
R27			1			1.9
R28		1				-0.2
R29				1		-0.2
R30	1					0.3
R31	1					0.7
R32			1			-0.1
TOTAL	10	14	4	3	1	23.7

*El signo negativo se debe a que los correctores fueron más estrictos en la segunda ocasión y, por tanto, la sustracción entre las medias PRE-POST resulta negativa

Como se observa, la diferencia entre las puntuaciones máximas y mínimas del PRE y POST de la mayoría de los correctores (n = 24) oscila entre los 0 y 1 punto de diferencia. Este dato nos permite inferir que los correctores son, en general, consecuentes consigo mismo en la evaluación

[6]La abreviatura Rn se utiliza para referirnos a los distintos correctores

holística de los ensayos en el PRE y en el POST. Por otra parte, la diferencia global entre las medias de las evaluaciones holísticas PRE y POST de los correctores es inferior a 1 punto (0,7 puntos), lo que corrobora el alto grado de consistencia de las actuaciones individuales de los correctores, a nivel general.

No obstante, también se observan actuaciones arbitrarias en el comportamiento de algunos correctores. Así, en la Tabla 10.37 se observan cuatro casos de correctores que registran 2 puntos de diferencia entre las puntuaciones máximas y mínimas en el PRE y en el POST y cuatro casos de correctores que muestran unas diferencias que podrían considerarse *extremas* entre las puntuaciones holísticas PRE y POST dado que alcanzan los 3 y 4 puntos de diferencia. La Tabla 10.37 nos permite identificar estos últimos casos: R8, R11, R29 y R23, siendo el corrector (R23) el que demuestra ser menos consistente.

Si volvemos a las Tablas 10.2a. y 10.2b que comentábamos al principio de este capítulo (epígrafe 10.1) obtenemos una información particularizada del comportamiento de estos cuatro últimos casos citados. A modo ilustrativo, podemos señalar que el corrector 23 (R23) asigna 5 puntos al ensayo 10 (h10) en la evaluación holística del PRE mientras que en la evaluación holística del POST dicho corrector le concede 8 puntos a ese mismo ensayo (h10).

Como venimos subrayando, la diferencia de puntos que se observa en la evaluación holística de los ensayos en el PRE y POST cobra una relevancia especial dentro del contexto de las PAAU ya que las décimas y

centésimas de punto condicionar la elección de la futura carrera universitaria de los candidatos.

En este sentido, conviene destacar que, existen técnicas estadísticas como las de la asociación lineal que nos permiten identificar el comportamiento *extremo* de algunos correctores (i.e. *outliers* o *extreme cases*) con respecto a la media. Estos correctores destacan tanto por su extremada indulgencia como por su excesiva exigencia.

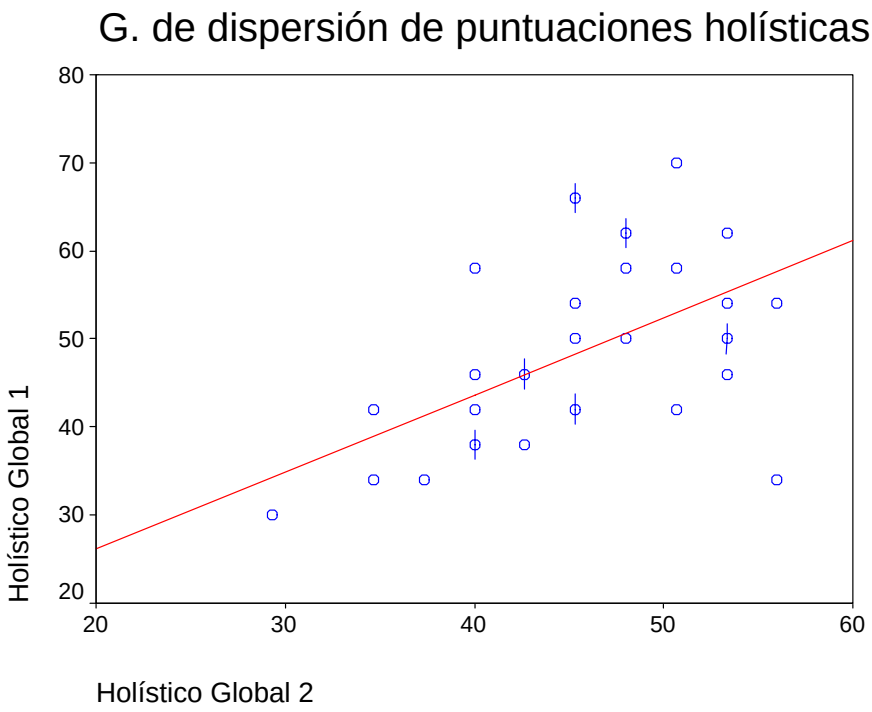
Así pues, nuestro siguiente paso en el estudio de la fiabilidad intra-corrector consistirá en examinar los resultados obtenidos a través de la implementación de la técnica de la asociación lineal. Concretamente, vamos a centrarnos en el análisis de los gráficos de dispersión.

10.3.1. Casos extremos

Los gráficos de dispersión nos muestran de forma plástica la relación que existe entre la actuación de los distintos correctores en las evaluaciones holísticas PRE y POST. Esta técnica también nos permite identificar los casos de correctores extremos o *outliers*.

La Fig. 10.8 nos presenta el primer gráfico.

Fig.10.8. Gráfico de dispersión de las puntuaciones holísticas PRE y POST



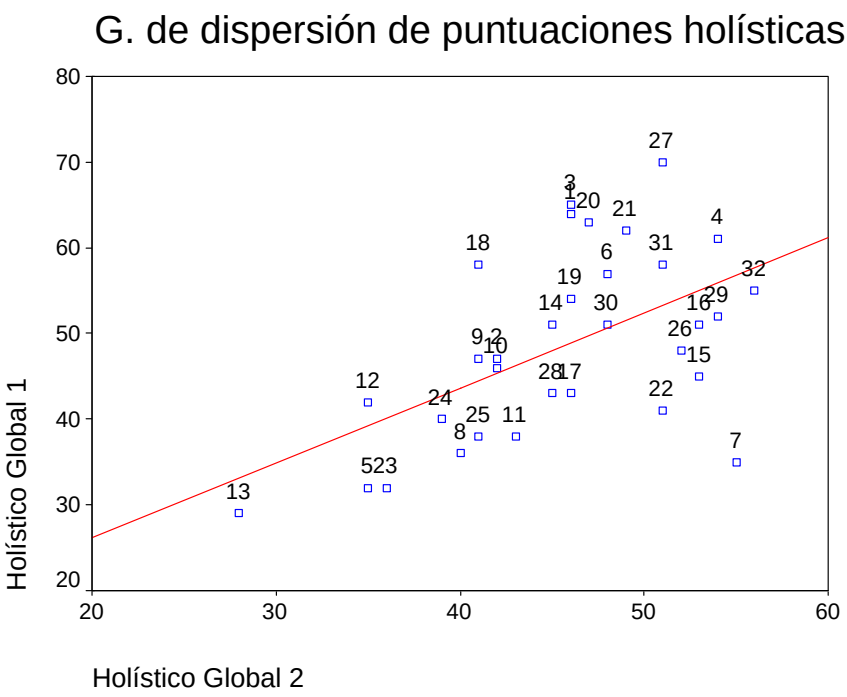
Como vemos, la dirección ascendente de la línea de regresión evidencia una relación positiva entre las puntuaciones holísticas en el PRE y en el POST. Esto es, a medida que las puntuaciones holísticas aumentan en el PRE también aumentan las puntuaciones holísticas en el POST.

No obstante, el gráfico nos muestra una cierta dispersión en las puntuaciones alrededor de la línea de regresión. Esto significa que hay ciertas discrepancias en las valoraciones realizadas por los correctores en ambas ocasiones (PRE y POST).

El siguiente gráfico (Fig. 10.9) permite identificar a los correctores por su número. Este sistema facilita el seguimiento de las actuaciones individuales y la localización de los casos extremos.

El gráfico revela que los correctores R27 y R10 son casos extremos. El R27 demuestra ser un corrector excesivamente benevolente que puntúa con valores $>$ 70 (7,0) y el corrector R10 demuestra ser un corrector excesivamente exigente dado que puntúa con valores $<$ 29 (2,9). La ventaja que nos ofrece la identificación de los casos extremos es que nos permite prescindir de la participación de dichos correctores en las futuras pruebas evaluadoras.

Fig.10.9. Gráfico de dispersión de las puntuaciones holísticas PRE y POST



CONCLUSIONES

Introducción

En este trabajo hemos analizado la fiabilidad de los las puntuaciones holísticas que asignan los correctores al ejercicio del ensayo que se incluye en la prueba de Inglés de Selectividad. Las puntuaciones holísticas se han contrastado con las analíticas y se ha estudiado tanto la fiabilidad inter-corrector como la fiabilidad intra-corrector en la implementación de ambos tipos de evaluación.

Dentro del estudio de la fiabilidad inter-corrector se han examinado, en primer lugar, el grado de consistencia y fiabilidad de las puntuaciones holísticas y analíticas. En segundo lugar, se han analizado los perfiles docentes, actitudinales y evaluadores de los correctores y se ha estudiado su incidencia en las puntuaciones holísticas de los ensayos. Asimismo, se ha investigado en profundidad la influencia de las variables género y situación laboral en las valoraciones del rendimiento de los candidatos que establecen los correctores. Finalmente, se ha abordado el estudio de los elementos sobresalientes o elementos que los correctores destacan en la evaluación holística de los ensayos. Entre dichos elementos, se ha prestado especial

atención a la relevancia del factor error en las puntuaciones finales que se asignan a los ensayos.

Las conclusiones se han organizado en seis apartados diferentes de acuerdo con las líneas de investigación que hemos abarcado. Empezaremos por exponer el estudio de la fiabilidad inter-corrector en las evaluaciones holísticas y analíticas de los ensayos (epígrafe A). En este epígrafe, abordaremos también la influencia que ejerce el nivel de dominio lingüístico de los ensayos en ambos tipos de evaluación. A continuación, bajo el epígrafe B, analizaremos la incidencia del género y situación laboral de los correctores en las puntuaciones holísticas y analíticas. Seguidamente, (epígrafe C) examinaremos los elementos de los ensayos que tienen un mayor peso en las evaluaciones holísticas y que determinan su puntuación final. Posteriormente (epígrafe D), se contemplará la evaluación de los errores contextualizados (i.e. en los ensayos) y descontextualizados (i.e. en frases individuales desprovistas de contexto). En el epígrafe E, abordaremos la fiabilidad inter-corrector. Todos los aspectos citados intentarán contrastarse con la visión teórica de los correctores sobre su perfil docente y evaluador (ver capítulo 12). Finalmente, en el epígrafe F, exploraremos las implicaciones pedagógicas de nuestro estudio.

A. Fiabilidad inter-corrector en las evaluaciones holísticas y analíticas

En este apartado se expondrán los resultados de las evaluaciones holísticas y analíticas a nivel general y las conclusiones sobre la influencia que ejerce el dominio lingüístico del ensayo en las puntuaciones holísticas.

A.1. Evaluaciones holísticas y analíticas a nivel general

Los resultados obtenidos del análisis de las correlaciones entre las evaluaciones holísticas y entre las analíticas (PRE y POST) a nivel general demuestran que las correlaciones son muy significativas (i.e. con un $p< 0,01$) en ambos casos. No obstante, los valores de las correlaciones son muy bajos: 0,508 en las evaluaciones holísticas y 0,475 en las evaluaciones analíticas. Estos datos indican que aproximadamente un 50% o un 47% de la variabilidad de las puntuaciones holísticas y analíticas respectivamente es consistente mientras que el resto de la variabilidad que se observa (i.e. 50% y 52%) se atribuye a factores fortuitos.

El cruce de evaluaciones holísticas y analíticas muestra de nuevo correlaciones muy significativas salvo en las correlaciones cruzadas entre las puntuaciones holísticas de la segunda ocasión (HG2) y las puntuaciones analíticas en la primera ocasión (AG1). No obstante, en general los valores de las correlaciones cruzadas se incrementan en relación a las correlaciones anteriores, especialmente en HG2 y AG2[1] (0,706).

---

[1]Las abreviaturas HG1 y HG2 se refieren a las evaluaciones holísticas globales de la primera y la segunda ocasión (PRE y POST respectivamente). De igual modo, las abreviaturas AG1 y AG2 se utilizan para referirse a las evaluaciones analíticas de la primera y la segunda ocasión (PRE y POST).

De este modo ratificamos las hipótesis de los estudios que señalan que las evaluaciones holísticas y las analíticas correlacionan positivamente (Cooper 1977; Evans 1981; Bacha 2001). Los datos indican que no existen grandes discrepancias entre las puntuaciones que se obtienen con la implementación de ambos métodos evaluativos. Así pues, no se demuestra la superioridad del método holístico vs. el analítico o viceversa (ver Cast 1939; Brooks 1980). La elección entre ambos métodos responderá a consideraciones pedagógicas (i. provisión de feedback a los estudiantes, intervención educativa etc.) o de carácter práctico (i.e. cualidad de la factibilidad).

Por otra parte, el nivel de consistencia de las puntuaciones holísticas y analíticas, según nos indica el coeficiente de correlación intra-clase (CCI) y el coeficiente de fiabilidad α es de 0,6556 y 0,6967 respectivamente. Estos datos señalan un grado de concordancia que se considera *regular-buena* (Fleiss, 1986).

Según Hatch y Lazaraton (1991), la obtención de una fiabilidad inter-corrector en torno al 0,75 no es inusual entre aquellos correctores que evalúan siguiendo su propio criterio o que no han recibido ningún tipo de formación previa en los procesos de evaluación. Estos autores explican que un coeficiente de correlación mucho más alto, en circunstancias como las anteriores, resultaría sorprendente. Por su parte, un coeficiente de correlación mucho más bajo significaría que los correctores habrían aplicado criterios evaluativos totalmente diferentes. Así, los datos de nuestro estudio

confirman que, a pesar de carecer de criterios de evaluación establecidos, los correctores evalúan los ensayos aplicando criterios evaluativos *similares*.

Los resultados anteriores corroboran, además, las conclusiones de los estudios que afirman que tanto los métodos holísticos de evaluación (Cooper 1977; Hughes 1989) como los analíticos producen resultados fiables (Jacobs *et al.* 1981; Bachman 1991; Hamp-Lyons 1991; Gamaroff 2000).

No obstante, es importante hacer notar que los coeficientes de correlación intra-clase y los coeficientes de fiabilidad α que se alcanzan tanto en las puntuaciones holísticas como en las analíticas son tan sólo moderamente altos. Este dato no puede obviarse en el contexto de las PAAU donde las décimas y centésimas de punto condicionan la posibilidad de elegir la futura carrera universitaria de los candidatos.

Es más, las pruebas de significación entre las evaluaciones holísticas y analíticas (PRE y POST) a nivel general, señalan diferencias significativas en todos los pares de muestras relacionadas (i.e. AG1-AG2; HG1-AG1; HG2-AG2) salvo en el par HG1-HG2. Estos datos sugieren que la actuación de los correctores en las evaluaciones analíticas y en el cruce de evaluaciones holísiticas y analíticas en las dos ocasiones es muy dispar.

Sorprendentemente, las evaluaciones holísticas en la primera y segunda ocasión, con la desviación típica más alta (DT =9,1068) no muestran diferencias significativas. Se confirma, por tanto nuestra Hipótesis 1. La hipótesis nula no puede ser rechazada dado que la diferencia entre las puntuaciones holísticas en las dos ocasiones (PRE y POST) no es significativa.

Como vimos, el grado de acuerdo absoluto entre los correctores era superior en las evaluaciones holísticas (0,6390) que en las analíticas (0,5993). Este resultado tiene sus antecedentes en los estudios de O'Loughlin (1994) y Bacha (2001) que admiten que las evaluaciones holísticas evidencian un mayor consenso entre los correctores. No obstante, según O'Loughlin (1994) este dato no significa que las valoraciones holísticas sean más válidas que las analíticas.

De hecho, conviene recordar que las puntuaciones medias sobre las que se basan los resultados de nuestro estudio tienden a neutralizar la dispersión de las puntuaciones. Así, cuando tomamos como punto de referencia el rango de las puntuaciones holísticas en los ensayos 1 (h1), 5 (h5) y 10 (h10), representativos de los niveles de dominio lingüístico alto, medio y alto respectivamente, la amplitud es considerable. En estos tres casos, la diferencia de puntos entre las puntuaciones holísticas máximas y mínimas es de 5 puntos y 6 puntos sobre una escala de 10. Este dato es claramente preocupante dadas las consecuencias sociales y personales que se derivan de las puntuaciones que se obtienen en las PAAU (ver Hamp-Lyons 1991; Messiclk 1993).

Asimismo, los datos anteriores nos conducen a rechazar nuestra Hipótesis 3. En ella se planteaba que no había diferencias significativas entre las puntuaciones holísticas y las analíticas. Nuestro estudio, sin embargo, muestra diferencias significativas entre ambos métodos evaluativos. De ahí se deduce que, a pesar de que las puntuaciones holísticas y analíticas correlacionan positivamente, la elección del método

evaluativo afectará las evaluaciones realizadas por los correctores (Hamp-Lyons 1990; Bachman y Palmer 1996). Los resultados revelan que las puntuaciones analíticas de los ensayos son más altas que las holísticas en las dos ocasiones (PRE y POST).

A.2. El nivel de dominio lingüístico de los ensayos

El estudio de la fiabilidad en los ensayos h1, h5 y h10, representativos del dominio que se tiene de la lengua: bajo, medio y alto respectivamente indica que la correlación entre las evaluaciones holísticas (PRE y POST) es mayor en aquellos ensayos con un dominio de la lengua superior.

Las correlaciones más bajas se dan en los ensayos con un dominio lingüístico medio-bajo. Así, estos ensayos son los únicos que no correlacionan consigo mismo. Estos datos sugieren que las puntuaciones holísticas que se asignan a estos ensayos son más arbitrarias y, por tanto, registran mayores discrepancias. Como se recordará, la amplitud del rango del ensayo con un dominio lingüístico medio era la mayor de los ensayos ya que alcanzaba los 6 puntos sobre una escala de 10. De este modo, se confirman las hipótesis establecidas por Connor-Linton (1994) y Vaughan (1991) que señalan que los ensayos con un dominio de la lengua medio son los que presentan una evaluación más problemática.

Asimismo, el coeficiente de correlación intra-clase del análisis de la fiabilidad, que evalúa consistencia y acuerdo absoluto, corrobora los datos anteriores que indican que el ensayo con un dominio de la lengua alto (h10)

es el que muestra un grado de consistencia (0,7220) y de acuerdo absoluto (0,7274) mayor. Como vemos, ambos valores se hallan en el umbral de los valores de concordancia clasificados como *muy buenos* (Fleiss, 1986). Por su parte, los ensayos con un dominio lingüístico medio y bajo registran un grado de concordancia que se clasifica como de *regular-bueno* (Fleiss, 1986).

Cabe señalar que el ensayo con un dominio de la lengua medio (h5), si bien muestra un grado de consistencia ligeramente mayor que el ensayo con un nivel lingüístico bajo (h1), alcanza un grado de acuerdo absoluto menor. Este resultado coincide con el mayor rango de variabilidad de las puntuaciones holísticas que se asignan al ensayo de tipo medio.

A la luz de estos resultados podemos afirmar que el nivel de dominio lingüístico de los ensayos afecta parcialmente las valoraciones de los correctores (Brown 1991; Sweedler-Brown 1993; Connor-Linton 1995; Milanovic *et al* 1996). Los ensayos con un dominio medio-bajo son los que presentan mayor problema de correlación entre los correctores. Por su parte, los ensayos con un dominio de la lengua alto tienden a incrementar las correlaciones significativas tanto entre las puntuaciones holísticas individuales como entre las puntuaciones holísticas y analíticas en su conjunto. Estos ensayos son, además, los que registran puntuaciones más homogéneas y consistentes.

B. El género y la situación laboral de los correctores

Los resultados del análisis de varianza (ANOVA) parecen indicar que las variables género y situación laboral de los correctores no influyen, en general, de forma significativa en la evaluación holística y analítica de los ensayos. Los datos indican que las diferencias que se observan en la actuación de los correctores se pueden atribuir a fluctuaciones ocasionales más que a diferencias sistemáticas producidas por estas variables (ver Baird, 1988). Se requiere una muestra más amplia de sujetos para poder confirmar estas hipótesis.

No obstante, en las evaluaciones holísticas y analíticas del POST (HG2 y AG2) se observan diferencias significativas en cuanto al género y situación laboral. Estos resultados impiden que se rechace la hipótesis nula número 4 que afirma que no hay diferencias significativas en las valoraciones de los correctores de acuerdo con las variables género y situación laboral.

En estas últimas evaluaciones (HG2 y AG2), los resultados de la variable género indican que, en general, el grupo de hombres asigna puntuaciones más altas a los ensayos que el grupo de mujeres tanto en las evaluaciones holísticas como en las analíticas. De ahí, se deduce que el grupo de hombres se muestra más benévolo e indulgente que el de mujeres corrigiendo los ensayos.

Cabe señalar, sin embargo, que si tenemos en cuenta ambas variables, es decir género y situación laboral, se observa que el grupo de mujeres de secundaria es menos estricto que el grupo de hombres de

secundaria en la evaluación holística de los ensayos. Este dato confirma la hipótesis de Vann Lorenz y Meyer (1991) que señalan que las mujeres tienden de forma general a mostrar una actitud más tolerante que los hombres en la evaluación de los ensayos.

Así, la tendencia general, si bien estadísticamente no significativa, que se observa en el estudio del resto de las evaluaciones indica que el grupo de mujeres de secundaria otorga puntuaciones más altas en todas las evaluaciones (i.e. holísticas PRE y POST y analíticas PRE) salvo en las analíticas del POST. En este último caso, el grupo de hombres de secundaria se muestra más indulgente que el de mujeres de secundaria en la asignación de puntuaciones.

Por el contrario, en el marco de la enseñanza universitaria, el grupo de mujeres de universidad otorga puntuaciones más bajas a los ensayos que el grupo de hombres. Estos datos ratifican, por tanto, la hipótesis de Herrera (2000/2001) que afirma: "As for gender rating behaviour among University raters, leniency is greater among men than among women..." (p. 176).

Cabe resaltar que el grupo de hombres de universidad es el grupo de correctores más indulgente en todos los casos estudiados.

Si comparamos la actuación del grupo de mujeres de secundaria y universidad, observamos que este último grupo es más estricto corrigiendo los ensayos en todas las evaluaciones salvo en la evaluación analítica del POST. En esta evaluación, la puntuación media del grupo de mujeres de universidad es superior a la del grupo de mujeres de secundaria. Estos resultados contradicen la hipótesis de Herrera (2000/2001) que señala que el

grupo de mujeres de secundaria es el grupo de correctores más exigente a nivel general.

C. Elementos que los correctores destacan en los ensayos.

El análisis de los estadísticos descriptivos de las evaluaciones analíticas (PRE y POST) indica que la *gramática* es la categoría que registra las medias más bajas tanto en el PRE como en el POST. De ahí se deduce que la *gramática* es la categoría que los correctores penalizan con mayor rigor en la evaluación de los ensayos. Como vemos, este dato coincide con el perfil del corrector sistemático y dominante de los correctores de nuestro estudio (capítulo 12, subepígrafe 12.2.1)

Es interesante destacar que las pruebas de significación que investigan las correlaciones entre las distintas muestras relacionadas indica que la relación entre las categorías analíticas es significativa en todos los casos salvo en el caso de la *gramática*. Estos datos coinciden con los resultados obtenidos en las correlaciones de Spearman entre las distintas categorías analíticas que indican que la *gramática* no correlaciona de forma significativa ni consigo misma ni con el resto de los componentes analíticos. Por tanto, se puede afirmar que los correctores juzgan la *gramática* de forma más arbitraria y más estricta que el resto de los componentes analíticos.

Por otra parte, las pruebas de significación señalan que la diferencia entre las medias de las muestras relacionadas es significativa en todos los casos salvo en el caso de la *presentación*. Sin embargo, la *gramática* es la

que presenta un nivel de significación mayor. A la luz de estos resultados, se puede inferir que existen grandes discrepancias en la evaluación de las categorías analíticas en la primera y en la segunda ocasión, especialmente en la evaluación de la gramática que, como vemos, constituye un punto clave de referencia para los correctores. Estos resultados contradicen las hipótesis que tienen los correctores sobre su perfil evaluador ya que dichos correctores admiten ser mayoritariamente consecuentes en sus valoraciones (ver capítulo 12, epígrafe 6.4). No obstante, los resultados revelan actuaciones muy dispares en las distintas evaluaciones (i.e. holísticas y analíticas) y en las distintas ocasiones (PRE y POST).

Por último, el análisis de frecuencias de los comentarios positivos y negativos que realizan los correctores sobre los ensayos demuestra, una vez más, que la *gramática* desempeña un papel esencial en la evaluación tanto positiva como negativa de los ensayos. La *organización*, por su parte, tiende a juzgarse como un aspecto fundamentalmente positivo. Estas dos categorías son las que mayoritariamente centran la atención de los correctores.

El resto de las categorías analíticas recibe una importancia mucho menor. El *vocabulario* se juzga tanto de forma positiva como negativa. Asimismo, los datos de nuestro estudio confirman los resultados de Brown (1991) que señalan que el *contenido* se juzga, en general, de forma positiva. No obstante, esta categoría recibe una importancia relativa.

Es interesante observar que la *mecánica* se evalúa exclusivamente de forma negativa ya que según el criterio de la comprensibilidad (ver Politzer

1978; Albrechtsen *et al.* 1980; Chastain 1980; Dordick 1996), este tipo de error es insignificante desde el punto de vista comunicativo.

Estos datos contradicen los resultados del cuestionario que señalan que el perfil docente del profesor comunicativo ocupa un segundo lugar en orden de importancia por encima del perfil del profesor dominante que ocupa el cuarto y último lugar (ver capítulo 12, subepígrafe 12.2.1.). No obstante, la actuación práctica de los correctores demuestra que los elementos formales más que los comunicativos son los que despiertan el mayor interés de los correctores.

C.1. Elementos que se destacan según el dominio lingüístico de los ensayos.

Las correlaciones entre las categorías analíticas en los ensayos h1, h5 y h10 representativos de los niveles de dominio de la lengua bajo, medio y alto respectivamente, demuestran que los correctores adecuan el *feedback* al nivel de dominio lingüístico de los ensayos.

Los datos indican que se enfatiza de forma especial el *vocabulario* en los ensayos con un dominio de la lengua medio y la *gramática* en los ensayos con un dominio lingüístico alto. Los correctores se muestran, sin embargo, dubitativos en cuanto a los aspectos que se han de primar en los ensayos con un dominio de la lengua bajo. Así, en este tipo de ensayos, el *contenido* es el elemento que correlaciona de forma más significativa con la evaluación holística del PRE. No obstante, en el POST se subrayan los

aspectos relacionados con el léxico (i.e. el *vocabulario*) y, especialmente, con la forma del ensayo (i.e. la *mecánica* y la *gramática*).

Estos resultados evidencian la controversia que se da entre *forma* y *contenido* y la tensión que experimentan los correctores de segundas lenguas a la hora de decidir qué elementos se han de evaluar y cómo se han de evaluar (Mohan y Low, 1995).

No obstante, los resultados de la técnica de regresión múltiple son determinantes. Los datos indican que los tres elementos que explican las puntuaciones holísticas de los ensayos con un dominio de la lengua bajo son: la *organización*, el *vocabulario* y la *gramática*, este último elemento se halla en el umbral de la significación. Se confirman, por tanto, los resultados de la III Parte del cuestionario que señala que estos tres últimos tipos de error son los que se juzgan de forma más seria (ver capítulo 12, epígrafe 6.5).

De todo lo que antecede se deduce que los correctores prestan especial atención a los aspectos discursivos (ver Freedman 1979; Freedman y Pringle; 1980 y Santos 1988), léxicos (Grobe 1981; Santos 1988; Engber 1995) y gramaticales (Sweedler-Brown 1993; Amengual y Herrera 2000) del ensayo. De esta forma, se contradicen los resultados de los estudios que defienden la hipótesis de que los aspectos temáticos son decisivos en las puntuaciones holísticas de los ensayos (Freedman 1979; Santos 1988; Song y Caruso 1996).

Asimismo, estos resultados nos obligan a rechazar nuestra hipótesis 5. En ella planteábamos que todos los componentes del ensayo ejercen el

mismo peso en las evaluaciones holísticas del ensayo. Nuestro estudio, sin embargo, señala que los elementos que determinan las puntuaciones holísticas son la *organización*, el *vocabulario* y la *gramática*. Asimismo, se rechaza la visión teórica de los correctores que enfatiza la importancia de la carencia de conocimientos generales previos en la evaluación de los ensayos (ver capítulo 12, epígrafe 6.5). Los aspectos de *contenido* apenas se consideran.

Los datos anteriores se confirman en el análisis de frecuencias de los comentarios positivos y negativos que realizan los correctores según el dominio lingüístico de los ensayos. Los resultados demuestran que la *gramática* polariza la mayoría de dichos comentarios independientemente del nivel de dominio de la lengua que demuestren los ensayos. La *organización* es el segundo elemento que se destaca en los ensayos de forma positiva.

Por el contrario, la *mecánica* se juzga de forma negativa tan sólo en los ensayos con un dominio de la lengua alto. Este dato contradice los resultados de Brown (1991) que concluye que la *mecánica* se evalúa de forma negativa en los ensayos con un dominio de la lengua medio y bajo.

La hipótesis que planteamos para explicar este caso es que los errores de *mecánica* son elementos que destacan o sobresalen en un *buen* ensayo (i.e. los ensayos con un dominio de la lengua alto) y, por tanto, llaman la atención de los correctores. En estos ensayos, este tipo de error puede producir cierta *irritabilidad*, lo que provoca la reacción negativa de los correctores (ver Vann *et al.* 1984; Santos 1988).

Probablemente, en los ensayos con un dominio de la lengua inferior los correctores centran su interés en los aspectos gramaticales o discursivos más relevantes desde el punto de vista comunicativo. La importancia de los errores de *mecánica* se tiende a relativizar en estos contextos dado que el criterio que se aplica es el de la *comprensibilidad* (Chastain 1980; Dordick 1996).

Por lo que respecta al *contenido*, este elemento se evalúa de forma positiva en todos los ensayos si bien la importancia que se le concede es muy inferior a la que se le atribuye a los elementos formales y discursivos (i.e. *gramática* y *organización*). Así pues, a pesar del gran interés suscitado por el enfoque basado en el proceso en el desarrollo de la producción escrita (ver Miller y Judy en Kroll 1990), el *contenido* y la expresión de ideas apenas se consideran en la evaluación de los ensayos (ver Applebee 1981; Perkins 1981; Homburg 1984; Ziv N. 1984; Zamel 1985)

Por último, los resultados del análisis de los comentarios positivos y negativos de los ensayos indican que previsiblemente los correctores incrementan el número de comentarios negativos a medida que desciende el dominio lingüístico del ensayo. No obstante, el número de comentarios positivos es similar en los diferentes ensayos independientemente del dominio lingüístico que demuestren. En otras palabras, se tienden a subrayar las cualidades positivas de un ensayo *malo*. Este tipo de *feedback* positivo puede resultar muy motivador para los estudiantes con un dominio de la segunda lengua inferior. Sin embargo, los datos indican que cuando se

evalúa un *buen* ensayo, los correctores se muestran más reacios a resaltar sus cualidades o aspectos positivos, lo que debiera hacernos replantear nuestra actuación en la evaluación de este último tipo de ensayos.

D. Evaluación de los errores

Los estadísticos descriptivos demuestran que las medias de las evaluaciones de las categorías analíticas contenidas en los ensayos son superiores a las medias de los errores contenidos en las frases descontextualizadas. Esto es así para todos los elementos salvo para las categorías de *registro* y *mecánica* cuyas puntuaciones medias son más altas en la evaluación de las frases descontextualizadas.

Asimismo, las pruebas de significación de las correlaciones entre las muestras relacionadas indican que la relación entre las puntuaciones de las categorías analíticas contenidas en el discurso y los errores contenidos en las frases descontextualizadas no es significativa salvo en el caso del *registro* y de la *mecánica*. Estos datos nos permiten inferir que dichas puntuaciones son independientes dado que no guardan relación alguna.

Por último, las pruebas de significación indican que la diferencia entre las medias de muestras relacionadas son significativas en la evaluación de los elementos de *mecánica*, *presentación* y, especialmente, *organización* y *gramática*. El elemento de *vocabulario* se halla en el umbral de la significación.

Estos datos indican que nuestra hipótesis 6 no se cumple. Debemos rechazar la hipótesis nula dado que la actuación de los correctores en la

evaluación de las categorías analíticas contenidas en el discurso y en los errores contenidos en las frases descontextualizadas muestra diferencias significativas. Los datos de nuestro estudio sugieren que los errores se evalúan con mayor dureza y rigor cuando se hallan descontextualizados. Estos datos tienen sus antecedentes en algunos estudios (Palmer 1973; Davies 1982; Dordick 1996) que señalan que los errores que se encuentran fuera del ámbito del discurso tienden a crecer fuera de proporción sin la ayuda del contexto y son juzgados, por consiguiente, de forma mucho más severa que aquellos contenidos en el discurso (i.e. los ensayos).

Se cuestiona, por tanto, la validez de algunos métodos de investigación que obtienen sus resultados basándose en al análisis de errores contenidos en frases descontextualizadas (Ludwig 1982; Davies 1983; Santos 1988; Dordick 1996).

Cabe subrayar que los errores de *gramática* son los que registran un mayor grado de significación y los que obtienen las medias más bajas en su evaluación tanto dentro del discurso como cuando se hallan en frases descontextualizadas. Estos datos revelan que dicha categoría se juzga con especial dureza (ver Burt y Kiparsky 1974). Por el contrario, los resultados sugieren que el *contenido* se evalúa de forma positiva, especialmente dentro del discurso (i.e. los ensayos).

E. Fiabilidad Intra-corrector

El estudio de la fiabilidad inter-corrector señala que los valores de la correlación entre un mismo corrector en las evaluaciones holísticas PRE y

POST son muy altos. En general, el nivel de correlación es ≥ 0,90. Asimismo, el coeficiente de correlación intra-clase (CCI), que mide el nivel de consistencia y de acuerdo absoluto, señalan un grado de concordancia *muy bueno* (≥ 0,90) de acuerdo con el criterio de Fleiss (1986). Estos datos sugieren que los correctores muestran un grado de consistencia y fiabilidad mucho más alto en sus actuaciones individuales que a nivel de grupo. En este sentido, el perfil del corrector consecuente que define a los correctores de nuestro estudio queda demostrado. (Ver capítulo 12, epígrafe 12.4).

De este modo, se confirma nuestra hipótesis 2 en la que postulábamos que las puntuaciones de un mismo corrector en diferentes ocasiones son consistentes (i.e. fiabilidad inter-corrector). Así, los resultados del análisis de la fiabilidad sugieren que los correctores se muestran estrictos o indulgentes de forma sistemática en las dos ocasiones (PRE y POST).

Como se recordará, la diferencia en el rango de las puntuaciones holísticas PRE y POST oscila mayoritariamente entre los 0 y 1 punto de diferencia, lo que indica un nivel de consistencia interna muy razonable. De hecho, la diferencia global entre las medias de las evaluaciones holísticas PRE y POST de los correctores es inferior a 1 punto (0,7).

A la luz de estos resultados podemos afirmar que los correctores parecen tener percepciones muy claras sobre las cualidades que definen un *buen* o un *mal* ensayo. Los criterios personales que establecen son los que guían sus actuaciones individuales. Se confirma, por tanto, la hipótesis de

Vaughan (1991) que asegura que: "holistic assessmenty is a lonely act" (p.120).

No obstante, en la evaluación de los ensayos también se observan algunos casos de correctores extremos que registran un rango de diferencia entre las puntuaciones holísticas del PRE y del POST de hasta 3 y 4 puntos de diferencia sobre una escala de 10 (ver R8, R11, R29 y R23). Este dato es realmente preocupante dada la trascendencia de los resultados de las prueba de Selectividad.

La técnica estadística de la asociación lineal se nos presenta como una herramienta muy útil para identificar el comportamiento extremo de los correctores, tanto por su excesiva indulgencia como por su excesiva exigencia con respecto a la media (ver, por ejemplo, R27 y R13). Como dijimos, la identificación de los casos extremos nos permitirá prescindir de la participación de estos correctores en futura pruebas evaluadoras.

En resumen, las principales conclusiones de nuestro estudio que conviene mencionar son las siguientes:

1. Las puntuaciones holísticas de los correctores en las dos ocasiones (PRE y POST) no muestran diferencias significativas. No obstante, la diferencia que se observa en el rango de dichas puntuaciones (de 5 y hasta 6 puntos sobre una escala de 10) así como el grado de concordancia que señala el coeficiente de

correlación intra-clase (CCI), indican un grado de fiabilidad únicamente moderado.

2. Las evaluaciones holísticas y analíticas correlacionan de forma significativa. El grado de consistencia entre ambas evaluaciones es de 0,684 (en HG1 y AG1) y 0,706 (en HG2 y AG2). Estos datos indican un grado de concordancia que se puede considerar *regular-bueno* (Fleiss, 1986).
3. La evaluación holística se ve afectada por el dominio lingüístico de los ensayos. Las correlaciones más altas entre las evaluaciones holísticas y las analíticas se observan en aquellos ensayos que presentan un dominio de la lengua alto. Estos últimos ensayos son los que, además, muestran un grado de concordancia mayor. Los ensayos con un dominio de la lengua medio presentan una evaluación más problemática. Estos ensayos poseen un rango de puntuaciones más amplio (i.e. de hasta 6 puntos sobre una escala de 10) y registran un grado de acuerdo absoluto entre los correctores inferior al de los ensayos con un dominio de la lengua alto y bajo.
4. Las evaluaciones analíticas (PRE y POST) y el cruce de evaluaciones analíticas y holísticas PRE y POST muestra diferencias significativas. No obstante, estas diferencias no pueden atribuirse a diferencias sistemáticas en cuanto a género o situación laboral de los correctores. El único caso que muestra una interacción de los factores género y situación laboral se da en el

cruce de las evaluaciones holísticas y analíticas en la segunda ocasión (HG2 y AG2). Los resultados parecen indicar que el grupo de hombres de universidad es el grupo de correctores más indulgente corrigiendo los ensayos.

5. Se observan diferencias significativas entre las categorías analíticas. La *gramática* es la categoría que presenta el mayor grado de significación. Este elemento es el que se juzga de forma más arbitraria y más estricta.
6. El análisis de frecuencias de los comentarios positivos y negativos de los ensayos indica que la *gramática* ejerce un papel decisivo en la evaluación positiva y negativa de los ensayos independientemente del nivel de dominio de la lengua que éstos presenten. La *organización* destaca como un elemento positivo en las valoraciones de los correctores mientras que la *mecánica* tiende a evaluarse como un elemento negativo, especialmente en los ensayos con un dominio de la lengua alto
7. La técnica estadística de regresión múltiple señala que los elementos analíticos que determinan la puntuación holística de los ensayos con un dominio de la lengua bajo son: la *organización*, el *vocabulario* y la *gramática* (esta última se halla en el umbral de la significación). Los datos sugieren que los aspectos discursivos, léxicos y gramaticales son los que centran el interés de los correctores. Por el contrario, el *contenido* y la expresión de ideas

apenas se contemplan, a pesar de la introducción del enfoque basado en el proceso en la enseñanza de la producción escrita.

8. Se observan diferencias significativas en la evaluación de los errores contenidos en el discurso (i.e. los ensayos) y los errores contenidos en frases descontextualizadas. Estos últimos errores se penalizan de forma más severa ya que el contexto ayuda a mitigar la reacción negativa de los correctores ante el error (Davies 1982; Dordick 1996). Los errores de *gramática* son los que se juzgan de forma más estricta tanto dentro como fuera del discurso.
9. Las puntuaciones de un mismo corrector en las distintas ocasiones (PRE y POST) son, generalmente, consistentes. El coeficiente de correlación intra-clase (CCI) señala un grado de concordancia *muy bueno* (Fleiss, 1986). Asimismo, el rango de diferencia global entre las puntuaciones holísticas se sitúa en los 0,7 puntos, es decir, una diferencia de puntos inferior a 1. Estos datos sugieren que los correctores mantienen el mismo grado de indulgencia o exigencia en sus distintas actuaciones.
10. Por último, los datos de la parte empírica revelan que el perfil docente de nuestros correctores es el de un profesor sistemático y dominante. Su perfil evaluador es consecuente en sus actuaciones individuales más que en sus actuaciones a nivel de grupo. La discrepancia que se observa en las puntuaciones holísticas y analíticas sugiere que no son correctores expertos. En general, se

demuestra que la visión teórica de la actuación de los correctores no se corresponde con su actuación práctica.

Estos resultados, a pesar de ser significativos desde el punto de vista estadístico, se han de interpretar con cautela dado el tamaño de la muestra.

G. IMPLICACIONES PEDAGÓGICAS

Los resultados de este estudio son relevantes desde el punto de vista pedagógico. Como hemos podido comprobar, a pesar de la importancia del enfoque basado en el proceso, la composición y el desarrollo de ideas en la enseñanza de la producción escrita, los datos demuestran que los correctores parecen centrar su atención en los aspectos formales de la lengua (Perkins 1981; Homburg 1984; Ziv N. 1984; Zamel 1985).

En este sentido, la *gramática* desempeña un papel decisivo. Las pruebas de significación indican que la *gramática* muestra el mayor grado de significación. El análisis de los comentarios positivos y negativos de los ensayos revela que la *gramática* es un elemento clave en la evaluación tanto positiva como negativa del ensayo. Asimismo, el análisis de errores corrobora que los errores de *gramática* son los que se juzgan con mayor dureza. A la luz de estos resultados se deduce que los profesores consideran la *gramática* como un subcomponente importante de la producción escrita.

De ser así, cabe replantearse el enfoque instruccional de la enseñanza de la producción escrita. Dada la importancia que le conceden los correctores a los aspectos formales de la lengua, sería conveniente acomodar las expectativas de *proceso* y *producto* estrechamente vinculadas a los aspectos de *forma* y *contenido*. Es evidente que en la evaluación de la producción escrita se ha de atender el aspecto comunicativo del texto. No obstante, la producción escrita también debe adecuarse a las demandas y exigencias de la comunidad académica. Sweedler-Brown (1993) afirma que el mejor servicio que podemos prestar a nuestros estudiantes es el de convertirnos en profesores de lengua y profesores de composición al mismo tiempo.

Para conseguir la evaluación conjunta de los aspectos de *forma* y *contenido* (ver Mohan y Low 1995; Fathman y Whalley 1990) se proponen diversas alternativas. Probablemente, la solución más factible dentro del contexto de la PAAU sea la construcción de una escala holística que combine ambos aspectos y les otorgue el mismo valor (Kroll, 1990). De este modo, nos aseguraremos de que tanto los aspectos retóricos como los formales se contemplan (Sweedler-Brown, 1993). Esta idea viene respaldada por la hipótesis del estudio de Norton y Starfield (1997) que señala que los estudiantes desean que se les facilite información sobre ambos aspectos.

BIBLIOGRAFÍA

Albrechtsen *et al*. (1980). Native Speaker Reactions to Learners' Spoken Inter-language. *Language Learning*, 30 (2), 365-396.

Alcaraz, E. (2000). *El inglés profesional y académico*. Madrid: Alianza Editorial, S.A.

Alcaraz, E. y Ramón, J. (1980). *La evaluación del inglés: Teoría y práctica*. Madrid: Sociedad General Española de librería, S.A

Alderson, J.C. (1981). Report of the Discussion on Communicative Language Testing. En J.C. Alderson y A. Hughes (Ed.), *Issues in Language Testing* (ELT Documents, Vol. 111). London: The British Council.

Alderson, J.C. (1983). Who Needs Jam? En Hughes y Porter (eds.). *Current Developments in Language Testing*, London: Academic Press.

Alderson, J.C. (1988). New procedures for Validating Proficiency Tests of ESP? Theory and Practice. *Language Testing*, 5 (2), 220-32.

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 S. Anivan (Ed.). *Current Developments in Language Testing,* Vol. 25,1-26.

Alderson, J.C. (1993). The Relationship between Grammar and Reading in an English for Academic Purposes Test Battery. En D.Douglas y C. Chapelle (Eds.), *A New Decade of Language Testing Research: Selected papers from the 1990 Language Testing Research Colloquium* Alexandria, Va.: TESOL: 203-19.

Alderson, J.C. (1995). Assessing Student Performance in the ESL Classroom. *TESOL Quartely*, V 29 (1), 184-187.

Alderson, J.C. y Banerjee, J. (2001). Language Testing and Assessment (Part I). *Language Teaching: The International Abstracting Journal*. Cambridge University Press, 213-236.

Alderson, J.C. y Hamp-Lyons, L. (1996). TOEFL Preparation Courses: A Study of Washback. *Language Testing*, 13 (3), 280-97

Alderson, J. C. y North, B. (1991). (ed.). *Language Testing in the 1990s.* UK: Modern English Publications and the British Council.

Alderson, J.C. y Urquhart, A.H. (1985). The Effect of Students' Academic Discipline on their Performance on ESP Reading Tests. *Language Testing*, 2 (2), 192-204.

Alderson, J.C. y Wall, D. (1993). Does Washback Exist?. *Applied Linguistics* 14, 115-129.

Alderson, J.C. y Windeatt, S. (1991). Learner-adaptive Computer-based Language Tests. *Language Testing Update*, 9, 18-21.

ALTE (1998). *ALTE handbook of European Examinations and Examination systems*. Cambridge: UCLES.

Amengual, M y Herrera, H (2000). Rater's Assumptions about Form and Content. Congreso de AESLA XVIII, Barcelona, mayo 2000

Anastasi, A. (1982/88). *Psychological Testing* (5th ed. y 6th ed.). London: Collier Macmillan.

Angoff, W. (1988). Validity; An Evolving Concept. En H. Wainer y H. Braun (eds.). *Test Validity*. Hillsdale, N.J.; Erlbaum: 19-32

Applebee, A. N. (1981). *Writing in the Secondary School*. NCTE Research Rep. No. 21. Urbana, ill: National Council of Teachers of English.

Applebee, A. (1983). Writing and Learning in School Settings. En M. Nystrand (Ed.), *What Writers Know: The Language, Process and Structure of Written Discourse*. New York: Academic Press: 365-382

Arndt, V. (1987). Six Writers in Search of Texts: A Protocol based Study of L1 and L2 Writing. *ELT Journal*, 41, 257-267.

Árva, V. y Medgyes, P. (2000). Native and Non-native Teachers in the Classroom. *System* 28, 355-372.

Bacha, N. (2000). Writing Evaluation: What can Analytic versus Holistic Essay Scoring Tell Us?. *System* (2001) 371-383

Bachman, L.F. (1988). Problems in Examining the Validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition* 10: 149-164.

Bachman, L.F. (1989). The Development and Use of Criterion-referenced tests of Language Proficiency in Language Program Evaluation. En Johnson R.K. (ed.) *The Second Language Curriculum*, Cambridge: Cambridge University Press.

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. *et al* (1995). Investigating Variability in Tasks and Rater Judgements in a Performance Test of Foreign Language Speaking. *Language Testing* 12, 238-57

Bachman, L.F y Palmer, A.S. (1981). The Construct Validation of the FSI oral Interview. *Language Learning* 31 (1), 67-86.

Bachman, L.F. y Palmer, A.S. (1982). The Construct Validation of some Components of Communicative Proficiency. *TESOL Quartely* 16, 49-65

Bachman, L. F. y Palmer, A.S. (1996). *Language Testing Practice*. Oxford: Oxford University Press.

Banerjee,J. y Luoma, S. (1997). Qualitative Approaches to Test Validation. En C. Clapham y D. Corson (edit), *Language Testing and Assessment*. Dordrecht, The Netherlands: Kluwer Academic Publishers: Vol 7, 275-87

Bailey, K. (1996). Working for Washback: A Review of the Washback Concept in Language Testing. *Language Testing*, 13 (3), 257-79.

Baird, J. A. (1988). What's in the Name?. Experiments with Blind Marking in A-Level Examinations. *Educational-Research* 40, 2: 191-202.

Barrit, L. Stock, P. y Clarke, F. (1986). Researching Practice: Evaluating Assessment Essays. *College Composition and Communication*, 37, 315-327.

Bennesch, S. y Rorschach, B. (1989). Academic Writing Workshop II. Belmont, C.A.: Wadsworth.

Bennet, R.E. *et al*. (1990). The Relationship of Expert System Scored Constrained Free Response Items to Multiple-choice and Open-ended Items. *Applied Psychological Measurement*. 14. 151-162

Berlin, J.A. (1982). Contemporary Composition: The Major Pedagogical Theories. *College English*, 44, 765-777

Bernárdez, E. (1995). *Teoría y epistemología del texto.* Madrid: Cátedra.

Berninger *et al.* (1996). Assessment of Planning, Translating, and Revising in Junior High Writers. *Journal of Scholl Psychology* 34, 1: 23-52.

Berry, V. (1993). Personality Characteristics as a Potential Source of Language Test Bias. En Huhta, A. Sajavaara, K. Y Tácala, S., editores,

Language Testing: new openings. Jyväskyla: University of Jyväskyla, 114-124

Boldt, R. (1992). Crossvalidation of Item Response Curve Models Using TOEFL Data. *Language Testing*. 9, 79-95

Breland *et al.* (1987). Assessing Writing Skill: Research Monograph. No. 11. New York: *The College Entrance Examination Board*.

Brière, E. (1966). Quantity before Quality in Second Language Composition. *Language Learning*, 16, 141-151.

Bridgeman, B. y Carlson, S. (1983). Survey of Academic Writing Tasks. *Written Communication,* 1, 247-280.

Brindley, G. (1989). The Role of Needs Analysis in Adult ESL Programme Design. En R.K. Johnson (ed.). *The Second Language Curriculum*. Cambridge: Cambridge University Press: 63-78.

Brindley, G. (2001). Outcomes-based Assessment in Practice: some Examples and Emerging Insights. *Language Testing*, 18 (4), 393-407.

Brooks, B. S. (1980). News Reporting and Writing. Missouri Group, St. Martin's Press.

Brown, A. (1993). The role of Test-taker Feedback in the Development Process: Test-takers' Reactions to a Tape-mediated Test of Proficiency in Spoken Japanese. *Language Testing.* 10, 277-304

Brown, A. (1995). The Effect of Rater Variables in the Development of an Occupation-specific Language Performance Test. *Language Testing* 12, 1-15

Brown, J.D. (1987). Principles of Language Learning and Teaching., 2nd ed. Englewood Cliffs, N.J.: Prentice Hall.

Brown, J.D. (1988). Improving ESL Placement Tests Using two Perspectives. *TESOLl Quartely*. 23, 65-83

Brown, J.D. (1991). Do English and ESL Faculties Rate Writing Samples Differently? *TESOL Quartely*. Vol 25, No. 4: 587-603.

Brown, J.D. (1992). Statistics as a Foreign Language- Part II: More Things to Consider in Reading Statistic Language Studies. *TESOL Quartely*, 26 (4), 629-664.

Brown, J.D. (1999). The Relative Importance of Persons, Items Subtests and Languages to TOEFL Test Variance. *Language Testing* 16(2), 217-38

Brown, J.D. y Hudson, T. (1998). The Alternatives in Language Assessment. *TESOL Quartely* 32(4), 653-75

Brown, P. y Levinson, S. (1978). *Politeness: Some Universals in Language in Language Usage.* Cambridge: Cambridge University Press.

Brownig, G. L. (1982). A Listener-based Hierarchy of Acceptability for Non-native Features of Oral English. Doctoral dissertation, University of California, Los Angeles, California.

Buck (1988), G. Testing Listening Comprehension in Japanese University Entrance Examinations. *JALT Journal* 10, 15-42.

Buck, G. (1991). The Testing of Listening Comprehension: An Introspective Study. *Language Testing*, 8 (1), 67-91

Buck, G. (1994). The Appropriacy of Psychometric Measurement Models for Testing Second Language Listening Comprehension. *Language Testing*. 11, 275-87145-170.

Burstein, J., *et al* (1996). Technologies for Language Assessment. *Annual Review of Applied Linguistics*,16, 240-60

Burstein, J. y Leacock, C. (2001). Applications in Automated Essay Scoring and Feedback. Comunicación presentada en el congreso “the Association of Language Testers in Europe” (ALTE), Barcelona.

Burt, M.K. y Kiparsky, C. (1974). Global and Local Mistakes. En J.H. Schumann y N. Stenson (Eds.), *New Frontiers in Second Language Learning*. Rowel, M.A: Newbury House: 71-80.

Burtoff. M. (1983). The Logical Organization of Written Expository Discourse in English: A Comparative Study of Japanese, Arabic, and Native Speaker Strategies. Doctoral Dissertation, Georgetown University.

Campbell, C. y Fiske, D. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. En: *Psychological Bulletin*, 56, 2.

Canale, M. (1983). On some Dimensions of Language Proficiency. En J.W. Oller (ed.). *Issues in Language Testing Research*. Rowley, M.A: Newbury House, 333-42

Canale, M. (1984). A Communicative Approach to Language Proficiency Assessment in a Minority Setting. En Rivera, C., edit, *Communicative Competence Approaches to Language Proficiency Assessment: Research and Application*. Clevedon: Multilingual Matters, 107-22.

Canale, M. y Swain, M. (1980/1). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics* 1 (1), 1-47

Carlson *et al*. (1985). *Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English* (TOEFL Research Rep. No (9). Princeton, NJ.: Educational Testing Service.

Carlson, S y Bridgeman, B. (1986). Testing ESL Student Writers. En K. Greenberg, H. Wiener, y R. Donovan (Eds.), *Writing Assessment: Issues and Strategies*. New York: Longman: 126-152.

Carroll, J.B. (1961). Fundamental Considerations in Testing for English Proficiency of Foreign Students. *Testing the English Proficiency of Foreign*

Students, Washington, D.C: Center for Applied Linguistics: 31-40. Reprinted in Allen, H.B. y Campbell, R.N. (eds.) (1972): 313-320.

Carroll, J.B. (1968). The Psychology of Language Testing. En Davies, A., edit, *Language Testing Symposium: a Psycholinguistic Approach*. Oxford: Oxford University Press, 46-69

Cason, G.J. y Cason, C.L. (1984). A Deterministic Theory of Clinical Performance Rating. *Evaluation and the Health Professions* 7, 221-47.

Celce-Murcia, M. (1997). Direct Approaches in L2 Instruction: A Turning Point in Communicative Language Teaching. *TESOL Quartely* 31: 141-52.

Chalhoub-Deville, M. (1997). Theoretical Models, Assessment Frameworks and Test Construction. *Language Testing*, 14 (1), 3-22.

Chapelle, C. (1994). Are C-tests Valid Measures for L2 Vocabulary Research? *Second Language Research*. 10, 157-187

Chapelle, C. y Abraham, R.G. (1990). Cloze Method: GAT Difference does it Make? *Language Testing.* 7, 121-146.

Chaplen, F. (1970). *Paragraph Writing*. London: Oxford University Press.

Charney, D. (1984). The Validity of Using Holistic Scoring to Evaluate Writing. *Research in the Teaching of English* 18, 65-81.

Chastain, K. (1980). Native Speaker Reaction to Instructor-identified Student Second-language Errors. *Modern Language Journal,* 64, 210-15.

Chastain, K. (1988). The ACTFL Proficiency Guidelines: A Selected Sample of Opinions. *ADFL Bulletin* 20, 47-51.

Cherry, R. (1989). Book Review: Assessing Writing Skills. *Composition Chronicle*: 11-12.

Clapham, C. (1993). Can ESP Testing be Justified? En D. Douglas y C. Chapelle (edit.) *A New Decade of Language Testing Research*. Alexandria, VA: TESOL Publications: 257-271

Clapham, C. (2000). Assessment and Testing. *Annual Review of Applied Linguistics*, 20, 147-61.

Coates, J. (1993). *Women, Men and Language*, 2nd edition (First Edition 1986). Longman, Harlow.

Coffman, W. (1971). On the Reliability of Ratings of Essay Examinations in English. *Research in the Teaching of English* 7, 356-71.

Coffman, W.E. y Kurfman, D. (1968). A Comparison of two Methods of Reading Essay Examinations. *American Educational Research Journal* 5, 101-20

Cohen, A. (1984). On Taking Language Tests: What the Students Report. *Language Testing* 1, 1: 70-81

Cohen, A. D. y Cavalcanti, M. C. (1990). Feedback on Compositions: Teacher and Student Verbal Reports. En B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom*. Cambridge: Cambridge University Press: 155-177.

Connor, U. (1984). A Study of Cohesion and Coherence in English as a Second Language Students' Writing. *Papers in Linguistics: International Journal of Human Communication* 17, 301-316.

Connor, U. (1987). Research Frontiers in Writing Analysis. *TESOL Quartely,* 21, 677-696.

Connor-Linton, J. (1995). Looking behind the Curtain: What do L2 Composition Ratings Really Mean? *TESOL Quartely* 29 (4), 762-765.

Constable, E. y Andrich, A. (1984). Inter-Judge Reliability: Is Complete Agreement among Judges the Ideal?. Annual National Council of Measurement in Education, New Orleans, L.A.

Cooper, C.R. (1977). Holistic Evaluation of Writing. En Cooper, C.R. y Odell, L. (Eds.), *Evaluating Writing: Describing, Measuring, Judging.* Urbana, IL: NCTE, 3-31

Coulthard, M y Ashby, M.C. (1975). Talking with the Doctor. *Journal of Communication*, 25, 240-247. En L. Hamp-Lyons (ed.). *Assessing Language Writing in Academic Contexts.* Norwood, NJ: Ablex: 127-153.

Crammer, N.A. (1985). *The Writing Process: 20 Projects for Group Work*. Rowley, MA: Newbury House.

Cripper, C. y Davies, A. (1988). ELTS: Research Reports 1 (i): Final Report of the ELTS Validation Project. London: British Council.

Cronbach, L. (1988). Five Perspectives on Validity Argument. En H. Wainer y H. Braun (eds.) *Test Validity*. Hillsdale, N.J.: Erlbaum: 3-17.

Cumming, A. (1987). Decision Making and Text Representation in ESL Writing Performance. 21st Annual TESOL Convention, Miami, abril.

Cumming, A. (1990). Expertise in Evaluating Second-Language Compositions. *Language Testing*, 7 (1), 31-51

Cumming, A. (1995). Introduction: The Concept of Validation in Language Testing. En A. Cumming y R. Berwick. *Validation in Language Testing*. Multilingual Matters LTD. Modern Languages in Practice 2.

Cummins, J. (1979). Cognitive Academic Language Proficiency, Linguistics Interdependence, the Optimum Age Question, and some other Matters. *Working Papers on Bilingualism* 19, 197-205.

Cummins, J y Swain, M. (1986). *Bilingualism in Education*. New York: Longman.

Davies, A. (1968). Introduction. En A. Davies (ed.) *Language Testing Symposium: A Psycholinguistic Approach*. London: Oxford University Press: 1-18.

Davies, A. (1977). The Construction of Language Tests. En Allen, J.P y Davies, A. (eds.). *Testing and Experimental Methods. Edinburgh Course in Applied Linguistics*, Vol, 4, London: Oxford University Press.

Davies, A. (1978). Language Testing: Survey Articles 1 and 2. *Language Teaching and Linguistics Abstracts*, 11, 145-59 y 215-31

Davies, A. (1982). *Survey 1. Eight State of-the-art Articles on Key Areas in Language Teaching*. Kinsella (ed.). Cambridge Language Teaching Surveys, Cambridge: 127-141 y 141-159

Davies, A. (1983). The Validity of Concurrent Validation. En A. Hughes y D. Porter (eds.) 1983. *Current Developments in Language Testing*. Academic Press: 141-145.

Davies, A. (1990). *Principles of Language Testing*. Oxford: Blackwell.

Davies, A. (1995). Proficiency of the Native Speaker: What are we Trying to Achieve in ELT. En: Cook, G. Seidlhofer, G. (Eds.). *Principle and Practice in Applied Linguistics*. Oxford University Press, Oxford: 145-157.

Davies, A. (1997). Demands of Being Professional in Language Testing. *Language Testing*, 14 (3), 328-39.

Davies, A (2001). The Logic of Testing Languages for Specific Purposes. *Language Testing*, 18 (2), 133-47.

Davies, A., (edit) (1997). Special Issue: Ethics in language testing. *Language Testing* 14(3).

Diaz, D. (1986). The Adult ESL Writer: The Process and the Context. 76th Annual NCTE Convention, San Antonio, Texas, Nov.

Diederich, P.B. (1974). *Measuring Growth in English*. Champaign, IL: National Council of Teachers of English.

Diederich *et al.* (1961). *Factors in Judgments of Writing Ability. Research Bulletin* 61-15. Princeton, NJ: Educational Testing Service

Dordick, M. (1996). Testing for a Hierarchy of the Communicative Interference Value of ESL Errors. *System*, Vol 24, No. 3, pp. 299-308.

Douglas, D. (1995). Developments in Language Testing: *Annual Review of Applied Linguistics*, 15, 167-87.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Dreyer, C. (1998). Teacher-Student Style Wars in South Africa: the Silent Battle. *System* 26, 115-126.

Edelsky, C. (1982). Writing in a Bilingual Program: The Relation of L1 and L2 Texts. *TESOL Quartely*, 16, 211-228.

Edgeworth, F.Y. (1980). The Element of Chance in Competitive Examinations. *Journal of the Royal Statistical Society* 53, 460-75 y 644-63

Eggington, W.G. (1987). Written Academic Discourse in Korean: Implications for Effective Communication. En U. Connor y R.B. Kaplan (Eds.) *Writing across Languages: Analysis of L2 Text*. Reading, MA: Addison-Wesley: 153-168.

Elbow, P. (1993). Ranking, Evaluating, and Liking: Sorting out Three Forms of Judgment. *College English*, 55, 187-206.

Elder, C. (1993). Are the Raters' Judgements of Language Teacher Effectiveness wholly Language Based?. 15th Language Testing Research Colloquium, Cambridge, agosto.

Elder, C. (1997). What does Test Bias have to Do with Fairness? *Language Testing*, 14 (3), 261-77.

Erazmus, E. (1960). Second Language Composition Teaching at the Intermediate Level. *Language Learning*, 10, 25-31.

Evans, P. (1979). Evaluation of Writing in Ontario: Grades 8, 11, and 13. *Review and Evaluation Bulletins*, 1 (2). Toronto: Ministry of Education.

Faigley, L. (1986). Competing Theories of Process: A Critique and a Proposal. *College English*, 48 (6), 527-542.

Farhady, H (1983). On the Plausibility of the Unitary Language Proficiency Factor. En Oller, J.W. (edit), *Issues in Language Testing Research,* Rowley, MA: Newbury House: 11-28.

Fathman, A. y Whalley, E. (1990). Teacher Response to Student Writing: Focus on Corm versus content. En Kroll, B. (Ed.), *Second Language Writing: Research Insights for the Classroom*. Cambridge University Press, Cambridge: 178-190.

Fayer, J.M.y Krasinki, E. (1987). Native and Nonnative Judgements of Intelligibility and Irritation. *Language Learning* 47, 313-26.

Fishman, P.M. (1978). Interaction: the Work Women Do. *Social Problems* 24, 397-406.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments.* New York: WILEY.

Flower, L. y Hayes, J.R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, 32, 365-387.

Follman, J. y Alderson, J. (1967). An Investigation of the Reliability of Five Procedures for Grading English Themes. *Research in the Teaching of English* 1, 190-200.

Freedman, S. W. (1979). How Characteristics of Student Essays Influence Teachers' Evaluations. *Journal of Educational Psychology*, 71, 328-338.

Freedman, S. (1991). Evaluating Writing: Linking Large-scale Testing and Classroom Assessment. No. 27. Berkeley, CA.: Center for the Study of Writing.

Freedman, S.W. y Calfe, R.C. (1983). Holistic Assessment of Writing: Experimental Design and Cognitive Theory. En P. Mosenthal, L. Tamor, y S. Walmsley (Eds.), *Research in Writing: Principles and Methods*. New York: Longman: 75-98.

Freedman, A. y Pringle, I. (1980). Writing in the College Years: Some Indices of Growth. *College Composition and Communication*, 31, 311-324.

Friendlander, A. (1990). Composing in English: Effects of a First Language on Writing in English as a Second Language. En B. Kroll (Ed.). *Second Language Writing: Research insights for the Classroom*. New York: Cambridge University Press: 109-125.

Fries, C. (1945). *Teaching and Learning English as a Second Language*. Ann Arbor: University of Michigan Press.

Fulcher, G. (1998). The Testing of Speaking in a Second Language. En Clapham, C., Corson, D. (edit), *Language Testing and Assessment. Encyclopaedia of Language and Education*. Vol. 7. Kluwer Academic Publishers, Dordrecht, 75-86.

Fulcher, G. (2000). Computers in Language Testing. En Brett, P., Motteram, G. (edit.), *Computers in Language Teaching*. IATEFL Publications, Manchester, 97-111

Fulcher, G. (2000). The 'Communicative' Legacy in Language Testing. *System* 28, 483-497.

Galloway, V. B. (1980). Perceptions of the Communicative Efforts of American Students of Spanish. *Modern Language Journal*, 64, 428-433.

Gamaroff, R. (2000). Rater Reliability in Language Assessment: the Bug of all Bears. *System* 28 (1), 31-53.

Gaskill, W. (1986). Raising in Spanish and English as a Second Language: A Process oriented Study of Composition. Doctoral Dissertation, University of California, Los Angeles.

Gass, S. y Varonis, E.M. (1986). Sex Differences in NSS / NNSS Interaction. Talking to Learn: Conversation in Second Language Acquisition. Ed. R. Day. Rowel, M.A.: Newbury House.

Gipps, C.V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London: The Palmer Press.

Goddart-Spear, M. (1983). Sex Bias in Science Teachers' Rating of Work. Second GASAT Conference, Oslo, Noruega.

Goldstein, H. (1993). Assessing Group Differences. *Oxford Review of Education* 19, 2, 141-50.

Goldstein, H. (1994). Recontextualising mental Measurement. *Educational Measurement: Issues and Practice*, Vol 13, 1.

Grabe, W. y Kaplan, R. B. (1989). Writing in a Second Language: Contrastive Rhetoric. En D. M. Johnson y D. H. Roen (Eds.), *Richness in Writing: Empowering ESL Students*. New York: Longman: 263-283.

Green, A. (1998). *Verbal Protocol Analysis in Language Testing Research: A Handbook*. Studies in Language Testing Series, Vol. 5. Cambridge: University of Cambridge Local Examinations Syndicate / Cambridge University Press.

Greenberg, K.L. *et al.* (ed.). (1986). *Writing Assessment: Issues and Strategies*. New York / London: Longman.

Greenberg, K.L. (1988). A Review of Assessing Writing Skills. *College Composition and Communication*, 39, 478-479.

Grobe, C. (1981). Syntactic Maturity, Mechanics and Vocabulary as Predictors of Quality Ratings. *Research in the Teaching of English*, 15, 75-86.

Gruba, P. y Corbel, C. (1997). Computer-based Testing. En Clapham, C y Corson, D., editores, *Language testing and assessment*. Volume 7. Language testing and assessment. Dordrecht: Kluwer Academic, 141-49.

Gyagenda, I.S. y Engelhard, G. (1998). Pattern, Domain and Gender Influences on the Assessed Quality of Students Writing Using Weighted and Unweighted Scoring. Annual Meeting of the American Educational Research Association (San Diego, CA, abril 13-17, 1998).

Hales, L.W. y Tokar, E. (1975). The Effect of the Quality of Preceding Responses on the Grades Assigned to Subsequent Responses to an Essay Question. *Journal of Educational Measurement*, 12, 27-45.

Hale-Benson, J. (1986). *Black Children, their Roots, Culture and Learning Styles,* rev. ed. Provo, Utah: Brigham Young University Press.

Hamayan, E. (1995). Approaches to Alternative Assessment. *Annual Review of Applied Linguistics*, 15, 212-26.

Hamp-Lyons, L. (1989). Preparing for the TOEFL Test of Written English. New York: Newbury House, Harper and Row.

Hamp-Lyons, L. (1990). Second Language Writing: Assessment Issues. En Kroll, B. (Ed.), *Second Language Writing: Research Insights for the Classroom*. Cambridge University Press, Cambridge: 69-87.

Hamp-Lyons, L. (ed.). (1991a). *Assessing Language Writing in Academic Contexts.* Norwood, NJ: Ablex.

Hamp-Lyons, L. (1991b). Basic Concepts. En L. Hamp-Lyons (ed.). *Assessing Language Writing in Academic Contexts.* Norwood, NJ: Ablex: 5-9.

Hamp-Lyons, L. (1991c). The Writer's Knowledge and Our Knowledge of the Writer. En L. Hamp-Lyons (ed.). *Assessing Language Writing in Academic Contexts.* Norwood, NJ: Ablex: 51-68.

Hamp-Lyons, L (1991d). Pre-Text: Task-Relate Influences on the Writer. En L. Hamp-Lyons (ed.). *Assessing Language Writing in Academic Contexts.* Norwood, NJ: Ablex: 87-107.

Hamp-Lyons, L (1991e). Reconstructing “Academic Writing Proficiency”. En L. Hamp-Lyons (ed.). *Assessing Language Writing in Academic Contexts.* Norwood, NJ: Ablex: 127-153.

Hamp-Lyons, L (1991f). Scoring Procedures for ESL Contexts. En L. Hamp-Lyons (ed.). *Assessing Language Writing in Academic Contexts.* Norwood, NJ: Ablex: 241-276.

Hamp-Lyons, L. (1996). Applying Ethical Standards to Portfolio Assessment of Writing in English as a Second Language. En M. Milanovic y N. Saville (editores), *Performance testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research* Colloquium. Studies in Language Testing Series. Cambridge: Cambridge University Press: Vol.3, 151-64.

Hamp-Lyons, L. (1998). Ethics in Language Testing. En C.M. Clapham y D. Corson (editores), *Language Testing and Assessment*. (Vol. 7). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Hamp-Lyons, L, Henning, G y DeMauro, G. (1988). Construct Validation of Communicative Writing Profiles. Tenth Annual Colloquium on Language Testing Research, University of Illinois en Urbana Champaign.

Harlen, W. (ed.) (1994). *Enhancing Quality in Assessment,* BERA policy Task Group on Assessment, Paul Chapman Publishers.

Harris, D.P. (1969). *Testing English as a Second Language*. New York: Mc Graw-Hill.

Harrison, A. (1991). Language Assessment as Theatre: Ten Years of Communicative Testing. En Alderson, C.H. y North, B. (ed.). *Language Testing in the 1990s.* London: Modern English Publications and the British Council.: 95-106.

Hartog, P (1936). English Composition at the School Certificate Examination and the 'Write anything about Something for Anybody' Theory. En M. Sadler *et al.* (Eds.), *Essays on Examinations*. London: Macmillan.

Hartog, P. *et al* (1941). *The Marking of English Essays.* London: Macmillan.

Hatch, E y Lazaraton, A. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. New York: Newbury House.

Hawthorne, L. (1997). The Political Dimension of Language Testing in Australia. *Language Testing*, 14 (3), 248-60.

Hayes, J.R. y Flower. L. (1983). Uncovering cognitive Processes in Writing: an Introduction to Protocol Analysis. En P. Monsenthal, L. Tamar y S. A. Walmsley, (Eds.), *Research in Writing*. New York: Longman: 206-220.

Heaton, J.B. (1975). *Writing English Language Tests*. London: Longman.

Heaton, J.B. (1982) (ed.). *Language Testing*. Oxford: Modern English Publications.

Heaton, J.B. (1990). *Classroom Testing*. UK: Longman.

Henning, G. (1987). *A Guide to Language Testing Development, Evaluation and Research*. Rowel, Mass: Newbury House.

Henning, G. (1991). Issues in Evaluating and Maintaining and ESL Writing Assessment Program. En Hamp-Lyons (ed.) *Assessing Second Language Writing in Academic Contexts.* New Jersey: Ablex Publishing Corporation Norwood: 279-291.

Henning, G. (1996). Accounting for Nonsystematic Error in Performance Ratings. *Language Testing*, Vol 13, 1, 53-61

Henning, G., *et al*. (1993). Computer-assisted Testing of Reading Comprehension: Comparisons among Multiple-choice and Open-ended

Scoring Methods. En D. Douglas y C. Chapelle (edit.). *A New Decade of Language Testing Research*. Alexandria, VA: TESOL Publications: 123-131.

Henning, G. Y Davidson, F. (1987). Scalar Analysis of Composition Ratings. En K. Bailey *et al. Language Testing Research*. Montery, Cal: Defense Language Institute.

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*. Madrid, 7: 89-107.

Herrera Soler, H. (2000a). The Effect of Gender and Working Place of Raters on University Entrance Examination Scores. *RESLA* 14, 161-179.

Herrera Soler, H. (2000b). The Influence of Scoring Procedure on Reliability and Validity of an Achievement Test. Proceedings of the International Conference on Measurement and Multivariate Analysis. Banff, Alberta, Canadá: Vol II: 161-163.

Herrera Soler, H. (2002). A New Insight into Examinee Behaviour in a Multiple-Choice Test: a quantitative Approach. *Estudios Ingleses de la Universidad Complutense*. Madrid, 10: 113-137.

Hill, K. (1991). Test Item Banker: An Item Banker for a very Small Micro. En Alderson, CH. y North, B. *Language Testing in the 1990s.* UK: Modern English Publications and the British Council: 247-255

Hill, K. (1996). Who Should be the Judge? The Use of Non-native Speakers as Raters on a Test of English as an International Language. En H. A. Viljokonen *et al.* (Ed.), *Current Developments and Alternatives in Language Assessment Proceedings of LTRC*: 275-290.

Hinds, J.(1987). Reader vs. Writer Responsibility: A New Typology. En U. Connor y R.B. Kaplan (Eds.), *Writing across Languages: Analysis of L2 Text* Reading, Mass: Addison-Wesley: 141-152.

Hirsch, E.D. y Harrington, D.P. (1980). Measuring the Communicative Effectiveness of Prose. En C. Fredericksen y J. Dominic (eds.). *Writing: the*

Nature, Development and Teaching of Written Communication. Hillsdale, N.J*.. Lawrence Erlbaum Associates.*

Hoetker, J. y Brossell, G. (1986). A procedure for Writing Content-fair Essay Examination Topics for Large-scale Writing Assessments. *College Composition and Communication*, 37 (3), 328-335.

Homburg, T. (1984). Holistic Evaluation of ESL Compositions: Can it be Validated Objectively? *TESOL Quartely* 18 (1), 87-107

Horowitz, D.M. (1986a). Process, not Product: Less than Meets the Eye. *TESOL Quartely*, 20 (1), 141-144.

Horowitz, D.M. (1986b). What Professors Actually Require: Academic Tasks for the ESL Classroom. *TESOL Quartely* 20 (3), 445-462.

Horowitz, D.M. (1991). ESL Writing Assessments: Contradictions and Resolutions. En Hamp-Lyons (ed.). *Second Language Research in Academic Contexts.* Cambridge University Press, Cambridge

Hudson, T y Lynch, B.K. (1984). A Criterion-referenced Approach to ESL Achievement Testing. *Language Testing* 1, 171-201

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press, GB.

Hughes, D.C. Keeling, B. y Tuck, B.F. (1980). The Use of Model Essays to Reduce Context Effects in Essay Scoring. *Journal of Educational Measurement*, 21, 277-281.

Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research* 60, 2: 237-263.

Hymes, D. (1972). On Communicative Competence. En J.B .Pride y J.Holmes (ed.), *Sociolinguistics: Selected Readings*. Harmondsworth. Middlesex: Penguin: 267-93

Ingram, E. (1985). Assessing Proficiency: an Overview on some Aspects of Testing. En Hyltenstam, K., Pienemann, M. (Eds.), Modelling and Assessing Second Language Acquisition. Multilingual Matters Ltd, Clevedon, Avon.

Jacobs, S. (1982). *Composing and Coherence: The writing of Eleven Pre-medical Students*. Linguistics and Literacy Series 3. Washington, D.C.: Center for Applied Linguistics.

Jacobs *et al.* (1981). *Testing ESL Composition: a Practical Approach*. Rowel, MA: Newbury House.

James, C. (1977). Judgments of Error Gravity. *English Language Teaching Journal*, 31, 116-124.

Johns, A.M. (1984). Textual Cohesion and the Chinese Speaker of English. *Language Learning and Communication*, 3, 69-74.

Johns, A.M. (1990). L1 Composition Theories: Implications for Developing Theories of L2 Composition. En B. Kroll (Ed.). *Second Language Writing: Research Insights for the Classroom*. New York: Cambridge University Press: 24-36

Johns, A.M. (1991). Faculty Assessment of ESL Student Literacy Skills: Implications for Writing Assessment. En L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic* Contexts. Norwood, NJ: Ablex: 167-179.

Johnson, C. (1985). The Composing Process of six ESL Students. Doctoral Dissertation, Illinois State University. En Krapels (1990). Second Language Writing Process Research.

Johnson, K. (1981). *Communicate in Writing.* London: Longman.

Janopoulus, M. (1992). University Faculty Tolerance of NS and NNS Writing Errors: A Comparison. *Journal of Second Language Writing* 1 (2), 109-121

Jones, S. (1982). Attention to Rhetorical Form while Composing in a Second Language. En C. Campbell, V. Flashner, T. Hudson, y J. Lubin (Eds.), *Proceedings of the Los Angeles Second Language Research Forum*, Vol. 2 Los Angeles: University of California, Los Angeles: 130-143. En Krapels

Jones, S. Y Tetroe, J. (1987). Composing in a Second Language. En A. matsushashi (Ed.), *Writing in Real Time: Modelling Production Processes*. Norwood, N.J.: Ablex: 34-57.

Kachru, B.B. (1988). Teaching World Englishes. ERIC/CLL *News Bulletin*, 12 (2-4),8.

Kaplan, R. (1967). Contrastive Rhetoric and the Teaching of Composition. *TESOL Quartely*, 1, 10-16.

Kaplan, R. B. (1987). Cultural Thought Patterns Revisited. En U. Connor y R. B. Kaplan (eds.) *Writing across Languages: Analysis of L2 Text*. Reading, M.A: Addison-Wesley: 9-20.

Kaplan, R. B. y Shaw, P. A. (1983). *Exploring Academic Discourse.* Rowley, M A: Newbury House.

Kelly, G.A. (1955). *The Psychology of Personal Constructs*. Vols. I y II. New York, NY: Norton.

Kelly, R. (1978). On the Construct Validation of Comprehension Tests: an Exercise in Applied Linguistics (PhD). University of Queensland.

Kinsella, K. (1995). Understanding and Empowering Diverse Learners in the ESL Classroom. En Reid, J. (Ed.). *Learning Styles in the ESL /EFL* Classroom. Heinle and Heinle, Boston.

Krapels, A. (1990). An Overview of Second Language Writing Process Research. En B. Kroll (1990). *Second Language Writing: Researh Insights for the Classroom*. Cambridge: Cambridge University Press: 37-56

Krashen, S.D. (1984). *Writing Research, Theory and Applications*. Oxford: Pergamon Institute of English.

Kroll, B. (ed.) (1990). *Second Language Writing: Research Insights for the Classroom*. Cambridge: Cambridge University.

Kroll, B. (1990b). Introduction. En B. Kroll (1990). *Second Language Writing: Research Insights for the Classroom*. Cambridge: Cambridge University Press: 1-5

Kroll, B. (1990c). What does Time Buy?. ESL Student Performance on Home versus Class Compositions. *Second Language Writing: Research Insights for the Classroom*. Cambridge: Cambridge University Press: 140-154.

Kroll, B. (1990d). The Rhetoric / syntax Split: Designing a Curriculum for ESL Students. *Journal of Basic Writing*, 9, 40-55.

Kunnan, A. (1992). An Investigation of a Criterion-referenced Test using G-theory, and Factor and Cluster analyses. *Language Testing*. 9, 30-50.

Kunz, L. (1972). *26 Steps: A Course in Controlled Composition for Intermediate and Advanced ESL Students*. New York: Language Innovations.

Lado, R. (1961, 1964). *Language Testing*. New York: McGraw-Hill

Laurier, M. (1991). What we Can do with Computerized Adaptive Testing_ and What we Cannot Do! En S. Anivan (ed.) *Current Developments in Language Testing*. Singapore: SEAMEO Regional Language Centre: 244-255.

Lay, N. (1982). Composing Processes of Adult ESL Learners. *TESOL Quartely*, 16, 406.

Lázaro, A (1996). Teaching and Assessing Writing Skills. En *Acquisition and Assessment of Communicative Skills*. Alcalá de Henares: Servicio de publicaciones de la universidad de Alcalá: 89-111.

Lehmann, R.H. (1983). Rating the quality of student writing: findings from the IEA study of achievement in written composition. En Huhta, A., Sajavaara, K. Y Tácala, S., editores, *Language Testing: new openings*. Jyväskyla: University of Jyväskyla, 186-204.

Leki, I. (1990). Potential Problems with Peer Responding in ESL Writing Classes. *CATESOL Journal* 3 (1), 5-19.

Linacre, J.M. (1989). *Many-faceted Rasch Measurement*. Chicago, IL: MESA Press.

Linacre, J.M. y Wright, B. (1992). *Facets: Rasch Measurement Computer Program, version 2.6.* Chicago, IL: MESA Press.

Locke, C. (1984). The Influence of the Interviewer on Student Performance Interests Of Foreign Language Oral Aural Skills. MA Report, University of Reading.

Low, G. (1982). The Direct Testing of Academic Writing in a Second Language. *System*, Vol, 10, No 3, 247-257.

Ludwig, J. (1982). Native-speaker judgements of Second-language Learners' Efforts at Communication: A review. *Modern Language Journal*, 66, 274-283.

Lumley, T. y McNamara, T.F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12(1), 54-71

Lunz, M.E. y Stahl, J.A. (1990). Judge Consistency and Severity across Grading Periods. *Evaluation and the Health Professionals* 13, 425-44.

Lukmani, Y. (1996). Linguistic Accuracy versus Coherence in Academic Degree Programs: a Report on Stage 1 of the Project. En Milanovic, M. y Saville, N., (edit.), *Performance Testing, Cognition and Assessment*. Cambridge: University of Cambridge Local Examinations Syndicate y Cambridge University Press, 130-50.

Lynch, B. (1997). In Search of the Ethical Test. *Language Testing*, 14 (3), 315-27.

Lynch, B. y Davidson, F. (1993). Criterion-referenced Language Test Development for Teaching and Research. Annual TESOL Convention. Atlanta, GA, abril.

Madaus, G. (1988). The Influence of Testing on the Curriculum. En Tanner (ed.) *Critical Issues in Curriculum*, 87th Yearbook of NSSE Part I, Chicago, IL: University of Chicago Press.

Madaus, G. (1990). Testing as a Social Technology. BOISI Lecture in Education and Public Policy, Boston College, 6 diciembre.

Martin-Betancourt, M. (1986). The Composing Processes of Puerto Rican College Students of English as a Second Language. Doctoral Dissertation, Fordham University.

Martínez-Arias, C (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Síntesis: Madrid.

McCarthy, C.P. (1987). A Stranger in Strange Lands: A College Student Writing across the Curriculum. *Research in the Teaching of English*, 21, 233-265.

McColly, W. (1970). What does Educational Research Say about the Judging of Writing Ability? *Journal of Educational Research* 64, 148-56.

McDaniel, B.A. (1985). *Ratings vs. Equity in the Evaluation of Writing*. 36th Annual Conference on College Composition and Communication. Minneapolis, MN.

McEldowney, P.L. (1976). *Test in English (overseas): the Position after Ten Years*, Manchester: Joint Matriculation Board.

McIntyre, P.N. (1993). The Importance and Effectiveness of Moderation Training on the Reliability of Teacher Assessments of ESL Writing Samples. MA. Thesis, University of Melbourne.

McNamara, T.F. (1990) Item Response Theory and the Validation of an ESP Test for Health Professionals. *Language Testing.* 7, 52-76.

McNamara, T. F. (1995). Modelling Performance: Opening Pandora's Box. *Applied Linguistics*, 16 (2), 159-75.

McNamara, T.F. (1998). Policy and Social Considerations in Language Assessment. *Annual Review of Applied Linguistics*, 18, 304-19.

McNamara, T.F. y Adams, R.J. (1991/1994). Exploring Rater Characteristics with Rasch Techniques. En *Selected Papers of the 13th Language Testing Research Colloquium (LTRC).* Princeton, N.J.: Educational Testing Service, International Testing and Training Program Office.

McNamara, T.F. y Lumley, T. (1993). The Effects of Interlocutor and Assessment Mode Variables in Offshore Assessment of Speaking Skills in Occupational Settings. 15th Language Testing Research Colloquium, Cambridge, agosto (ERIC ED 364 066)

Medgyes, P. (1994). *The Non-Native Teacher*. Macmillan Publishers, London.

Messick, S. (1989a). Validity. En R.L. Linn (Ed.), *Educational Measurement. Third Edition*. New York: Macmillan: 13-103

Messick, S. (1989b). Meaning and Values in Test Validation: the Science and Ethics of Assessment. Educational Researcher 18 (2), 5-11.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, 23 (2), 13-23.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice* 14(4), 5-8

Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing*,13 (3), 241-56.

Meuffels, B. (1989). The Abused Reader. An Article on the Skills of Dutch Language Experts in Evaluating Essays. *Journal of Language Performance*, 11, 161-76.

Michael *et al.* (1980). A Comparison of the Reliability and Validity of Ratings of Student Performance on Essay Examinations by Professors of English and by Professors in other disciplines. *Educational and Psychological Measurement* 19, 37-47.

Miller, J y Judy, S. (1978). *Writing and Reality*. New York: Harper and Row.

Mlynarcyk, R. (1992). Personal and Academic Writing: A False Dichotomy? *TESOL Journal*.

Mohan, B.A. y Low, M. (1995). Collaborative Teacher Assessment of ESL Writers: Conceptual and Practical Issues. *TESOL Journal*, vol. 5, no1, otoño.

Moller, A. (1982). A Study of the Validation of Proficiency Tests of English as a Foreign Language. Doctoral Dissertation, University of Edinburgh.

Morrow, K. (1977). *Techniques of Evaluation for a Notional Syllabus.* London: Royal Society of Arts.

Morrow, K. (1979). Communicative Language Testing: Revolution or Evolution? En Brumfit, C.J. y Johnson, K., (edit.), *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press, 9-25

Morrow, K. (1991). Evaluating Communicative Tests. En: Anivan, S. (Ed.), *Current Developments in Language Testing*, RELC, Singapore, 111-118.

Moss, P. (1992). Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research* 62, 229-58.

Moss, P. (1994). Can There be Validity without Reliability? *Educational Research* 23, 5-12.

Mullen, K. (1980). Evaluating Writing Proficiency in ESL. En J.R. Oller, J y K. Perkins (Eds.), *Research in Language Testing*. Rowley, MA.: Newbury House: 160-170

Munby, J. (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Myers, M. (1980). *A Procedure for Writing Assessment and Holistic Scoring*. Urbana, IL: National Council of Teachers of English and Educational Resources Information Center.

Nagy, P. *et al.* (1988). Exploratory Analysis of Disagreement among Holistic Essays Scores. *The Alberta Journal of Educational Research*, Vol. XXXIV, No. 4 diciembre, 355-374.

Neel, J. (1988). *Plato, Derrida, and Writing*. Berkeley: University of California Press.

Newcomb, J.S. (1977). The Influence of Readers on the Holistic Grading of Essays. Doctoral Dissertation, University of Michigan.

Norton, B. y Starfield, S. (1997). Covert Language Assessment in Academic Writing. *Language Testing*, 14 (3), 278-94

North, B. (1993). The Development of Descriptors on Scales of Language Proficiency. Washington, D.C.: National Foreign Language Center.

Nuttall, D. (1987). The Validity of Assessments, *European Journal of Psychology of Education*, 11, 2, 109-18.

O'Hare, F. (1973). *Sentence Combining: Improving Student Writing without Formal Grammar Instruction*. Urbana, IL: National Council of Teachers of English.

Oller, J.W. (1976). Evidence of a General Language Proficiency Factor: An Expectancy Grammar. *Die neuren Sprachen* 76, 165-74

Oller, J.W. (1979). *Language Tests at School*. London: Longman

Oller, J.W. y Richards, J.C. (1973). *Focus on the Learner: Pragmatic Perspectives for the Language Teacher.* Rowley Mass: Newbury House.

O'Loughlin, K. (1994). The Assessment of Writing by English and ESL Teachers. *Australian Review of Applied Linguistics* 17, (1) 23-44.

O'Sullivan, B. (2000). Exploring Gender and Oral Proficiency Interview Performance. *System* 28, 373-386.

Oxford *et al.* (1991) Style Wars: Teacher-Student Conflicts in the Language Classroom. En Magnan, S.S. (ed.) *Challenges in the 1990s for College Foreign Language Programs*. Heinle and Heinle, Boston, MA.

Oxford *et al.* (1992). Language Learning Styles: Research and Practical Considerations for Teaching in the Multicultural Tertiary ESL/EFL Classroom. *System* 20 (4), 439-456.

Palmer, L. (1973). A Preliminary Report on a Study of the Linguistic Correlates or Raters' Subjective Judgements of Non-native English Speech. En R.G. Shuy y R.W. Fasold (eds.), *Language Attitudes: Current Trends and Prospects*. Washington D.C.: Georgetown University Press.

Palmer, L. y Spolsky, B. (ed.) (1975). *Papers on Language Testing*, *1967-1974*, Washington D.C.: TESOL.

Papajohn, D. (1999). The Effect of Topic Variation in Performance Testing: The Case of the Chemistry TEACH Test for International Teaching Assistants. *Language Testing* 16 (1) 52-81.

Paulston, C. B. y Dykstra, G. (1973). *Controlled Composition in English as a Second Language*. New York: Regents.

Perkins, K. (1980). Using Objective Methods of Attained Writing Proficiency to Discriminate among Holistic Evaluations. *TESOL Quartely*, 14, 61-69.

Perkins, K.(1983). On the Use of Composition Scoring Techniques, Objective Measures and Objective Tests to Judge ESL Writing Ability. *TESOL Quartely* 17 (4), 651-671.

Piazza, L. (1980). French Tolerance for Grammatical Errors made by Americans. *Modern Language Journal*, 64, 422-427.

Pica *et al.* (1989). Comprehensible Output as an Outcome of Linguistic Demands on the Learner. *Studies in Second Language Acquisition* 11 / 1: 63-90.

Pienemann *et al* (1988). Constructing an Acquisition-based Procedure for Language Assessment. *Studies in Second Language Acquisition* 10, 217-43

Pollit *et al.* (1985). What Makes Exam Questions Difficult? An Analysis of 'O' Grade Questions and Answers. Research Reports for Teachers No. 2 Edinburgh: Scottish Academic Press.

Pollit, A y Hutchinson (1987). Calibrating Graded Assessments: Rasch Partial Credit Analysis of Performance in Writing. *Language Testing* 4: 72-92.

Politzer, R. (1978). Errors of English Speakers of German as Perceived and Evaluated by German Natives. *Modern Language Journal*, 62, 253-61.

Pollit, A. y Murray, N. (1996). What Raters *Really* Pay Attention To. En M. Milanovich y N. Saville (Eds.). *Performance, Testing, Cognition and Assessment*: *Selected papers from the 15th Language Testing Research Colloquium*, Cambridge y Arnhem.

Porter, D. (1983). The Effect of Quantity of Context on the Ability to Make Linguistic Productions: A Flaw in a Measure of General Proficiency.

Porter, D. (1991a). Affective Factors in Language Testing. En Alderson, C.J. North, B. (Eds.). *Language Testing in the 1990s*. London: Modern English Publications: 32-40.

Porter, D. (1991b). Affective Factors in the Assessment of Oral Interaction: Gender and Status. En Arnivan, S. (Ed.). *Current Developments in Language Testing*. Anthology Series 25. Singapore: SEMEO Regional Language Centre: 92-102.

Purpura, J.E. (1992). *Rater Consistency between and among ESL Teachers and Writing Program Teachers*. Department of TESL/Applied Linguistics. University of Los Angeles, California.

Raimes, A. (1983). Anguish as a Second Language? Remedies for Composition Teachers. En A. Freedman, I. Pringle y J. Yalden (Eds.), *Learning to Write: First Language / Second Language*. Londres: Longman, 258-272

Raimes, A. (1985a). An Investigation of the Composing Processes of ESL Remedial and Nonremedial Students. 36th Annual Convention, Minneapolis, Minn., marzo.

Raimes, A. (1985b). What Unskilled Writers Do as they Write: A Classroom Study of Composing. *TESOL Quartely*, 19, 229-258.

Raimes, A. (1987). Language Proficiency, Writing Ability, and Composing Strategies: A Study of ESL College Student Writers. *Language Learning*, 37, 439-467.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for Concern. *TESOL Quartely*, 24 (3), 427-42.

Raimes, A. (1991). Out of the Woods: Emerging Traditions in the Teaching of Writing. *TESOL Quartely*, Vol. 25, No. 3, otoño: 407-430

Raymond, J.C. (1982). What We don't Know about the Evaluation of Writing. *College Composition and Communication* 33 (4), 399-403.

Rea, P.M. (1978). Assessing Language as Communication. *MALS Journal*, verano, 45-68.

Rea, P.M. (1991). Response to Andrew Harrison Paper: Language Assessment as Theatre. En Alderson, CH. Y North, B. (1991). *Language Testing in the 1990s. Language Testing in the 1990s.* UK: Modern English Publications and the British Council: 106-111

Rea-Dickins, P. (1997). So Why do We Need Relationships with Stakeholders in Language Testing? A View from the UK. *Language Testing*, 14 (3), 304-14.

Read, J. (1990). Providing Relevant Content in an EAP Writing Test. *English for Specific Purposes*.9, 109-122.

Read, J. (2000). *Assessing Vocabulary Knowledge and Use*. Cambridge: Cambridge University Press.

Reid, J. (1990). Responding to Different Topic Types: A Quantitative Analysis from a Contrastive Rhetoric Perspective. En B. Kroll (Ed.). *Second Language Writing: Research Insights for the Classroom*. New York: Cambridge University Press: 191-210

Rosen, H. (1969). Towards a Language Policy across the Curriculum. En *Language, the Learner, and the School*. London: Penguin.

Ruth, L. y Murphy, S. (1984). Designing Writing Tasks for the Assessment of Writing. Ablex, Norwood, NJ.

Santos, T. (1988). Professors' Reactions to the Writing of Non-native-speaking Students. *TESOL Quartely* 22(1), 69-90.

Sasaki, M. (1996). *Second Language Proficiency, Foreign Language Aptitude, and Intelligence: Quantitative and Qualitative Analyses*. New York: Peter Lang.

Sasaki, M. (1999). Development of an Analytic Rating Scale for Japanese L1 Writing. *Language Testing* 16 (4) 457-478.

Savignon, S. (1982). *Communicative Competence: Theory and Classroom Practice*. New York: Addison-Wesley.

Selinker, L, Todd-Trimble, M y Trimble, L. (1978). Rhetorical Function Shifts in EST Discourse. *TESOL Quartely*, 12 (3), 311-320.

Schils *et al.* (1991). The Reliability Ritual. *Language Testing* 8, 2, 125-138.

Schoonen *et al.* (1997). The Assessment of Writing Ability: Expert Readers versus Lay Readers. *Language Testing* 14 (2) 157-184.

Shavelson, R.J. *et al.* (1992). Performance Assessments: Political Rhetoric and Measurement Reality. *Educational Researcher* 21, 4: 22-27.

Shehadeh, A. (1999). Gender Differences and Equal Opportunities in the ESL Classroom. *ELT Journal.* 53 (4), 56-261.

Sheorey, R. (1985). Goof Gravity in ESL: Native vs. Nonnative Perceptions. 19th Annual TESOL Convention, New York.

Shohamy, E. (1994). The Validity of Direct versus Semi-direct Oral Tests. *Language Testing*, 11 (2) 99-124.

Shohamy, E. (1997). Testing Methods, Testing Consequences: Are They Ethical? *Language Testing*, 14 (3), 340-9

Shohamy, E. (1999). Critical Language Testing: Uses and Consequences of Tests, Responsibilities of Testers and Rights of Test-takers. 21st Annual Language Testing Research Colloquium, Tsukuba, Japón.

Shohamy, E. (2001a). *The Power of Tests*. London: Longman.

Shohamy, E. (2001b). Democratic assessment as an Alternative. *Language Testing*, 18 (4). 373-92.

Shohamy, E y Reeves (1985). Authentic Language Tests: Where from and Where to. En *Language Testing*, 2, London: Edward Arnold.

Silva, T. (1990). Second Language Composition Instruction: Developments, Issues, and Directions in ESL. En B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom*. New York: Cambridge University Press: 11-23.

Skehan, P. (1987). Variability and Language Testing. En R. Ellis (Ed.), *Second Language Acquisition in Context*, Englewood Cliffs, N.J.: Prentice Hall.

Skehan, P. (1988). Language Testing: Survey Article, Part 1. *Language Teaching abstracts.* 21(4), 211-21

Skehan, P. (1989). Language Testing: Survey Article, Part 2. *Language Teaching abstracts.* 22 (1), 1-13

Skehan, P. (1991). Progress in Language Testing. The 1990s. En Alderson, J.C. y North, B., editores, *Language testing in the 1990s*. London: Macmillan, 3-21

Song, B. y Caruso, I. (1996). Do English and ESL Faculty Differ in Evaluating the Essays Of Native English-Speaking and ESL Students?. *Journal of Second Language Writing*, 5 (2), 163-182

Spaan, M. (1993). The Effect of Prompt in Essay Examinations. En D. Douglas y Chapelle (ed.). *A New Decade of Language Testing Research*. *Selected Papers from the 1990*. Language Testing Research Colloquium: 93-122

Spack, R. (1988). Initiating ESL Students into Academic Discourse Community: How Far should We Go? *TESOL Quartely*, 22, 29-51

Spolsky, B. (1975). Language Testing or Science?. Fourth AILA International Congress, Stuttgart.

Spolsky, B. (1981). Some Ethical Questions about Language Testing. En Klein-Braley, C. Y Stevenson, D.K. (edit.), *Practice and Problems in language testing*. Frankfurt: Peter Lang: 5-21.

Spolsky, B. (1985). The Limits of Authenticity in Language Testing. *Language Testing* 2 (1), 31-40.

Spolsky, B. (1997). The Ethics of Gatekeeping Tests: What have We Learned in a Hundred Years? *Language Testing*, 211-21.

Stahl, J y Lunz, M. (1992). *Judge Performance Reports.* 19th Annual TESOL Convention, New York.

Stansfield, C.W. (1993). Ethics, Standards and Professionalism in Language Testing. *Issues in Applied Linguistics* 4 (2), 15-30

Stern, H.H. (1983). *Fundamental Concepts of Language Teaching*. New York: Oxford University Press.

Stevens, V. (1991). Strategies in Solving Computer-based Cloze: Is it Reading? Annual TESOL Convention. New York, marzo.

Stewart, M. y Grobe, C. (1979). Syntactic Maturity, Mechanics of Writing and Teachers' Quality Ratings. *Research in the Teaching of English*, 13, 207-215.

Storey, P. (1997). Examining the Test-taking Process: a Cognitive Perspective on the Discourse Cloze test. *Language Testing*, 14 (2), 214-31.

Swain, M. (1984). A Review of Immersion Education in Canada: Research Evaluation Studies. *ELT Documents 119, Language Issues and Educational Policies.* London: British Council.

Sweedler-Brown, C.O. (1993). The Effects of ESL Errors on Holistic Scores Assigned by English Composition Faculty. *College ESL*, 3, 53-69

Swinford, F. (1964). Test Analysis, Advanced Placement Examination in American History Form MBP (ETS SR No.53) Princeton, NJ: Educational Testing Service.

Taylor, B. (1981). Content and Written Form: A Two-way Street. *TESOL Quartely*, 15, 5-13.

Taylor *et al.* (1988). (eds.) *Literacy by Degrees.* Milten Keynes: Open University Press.

Teddick, D.J. (1990). ESL Writing Assessment: Subject-matter Knowledge and its Impact on Performance. *English for Specific Purposes*, 9 (2), 123-43.

Tittle, C.K. (1989). Validity: Whose Construction is it in the Teaching and Learning Context?, *Educational Measurement: Issues and Practice,* 8, 1, primavera.

Tomiyama, M. (1980). Grammatical Errors Communication Breakdown. *TESOL Quartely*, 19 (1), 71-79.

Underhill, N. (1982). The Great Reliability / Validity Trade-Off: Problems in Assessing the Productive Skills. En J. B. Heaton (ed.) *Language Testing*. London: Modern English Publications.

Upshur, J y Turner, C.E. (1999). Systematic Effects in the Rating of Second-Language Speaking Ability: Test Method and Learner Discourse. *Language Testing*, 16 (1), 82-111.

Valette, R.M. (1967). *Modern Language Testing: a Handbook*. New York: Harcourt Brace Jovanovich

Vann, R. Lorenz, F. y Meyer, D. (1991). Error Gravity: Faculty Response to Errors in the Written Discourse of Nonnative Speakers of English. En L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex: 181-194

van Essen, A. (1994). Language Imperialism. NELLE, Innsbruck.

Van *et al*. (1984). Error Gravity: a Study of Faculty Opinion of ESL Errors. *TESOL Quartely*, 18 (3), 427-440

Van, R.J. y Meyer, F.Q (1984). Error Gravity: A Study of Faculty Opinion of ESL Errors. *TESOL Quartely* 18, 427-440.

Vaughan, C. (1991). Holistic Assessment: What Goes on in the Rater's Mind?. En L. Hamp-Lyons (ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex: 111.125.

Vollmer, H.J. (1981). Why are We Interested in General Language Proficiency? En Alderson , CH. Y Hughes, A (ed.). Issues in Language Testing: ELT Documents, 111

Wall, D. y Alderson, J.C. (1993). Examining Washback: The Sri Lankan Impact Study. *Language Testing*, 10 (1), 41-69.

Watanabe, Y. (1996). Introducing New Tests into Traditional Systems: Insights from General Education and from Innovation Theory. *Language Testing*, 13 (3), 334-54.

Weigle, S.C. (1994). Effects of Training on Raters of ESL Compositions. *Language Testing*, 11(2), 197-223.

Weigle, S.C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, 15(2), 263-87

Weir, C.J. (1983). *Identifying the Language Problems of Overseas Students in Tertiary Education in the UK.* PhD dissertation, University of London, London

Weir, C. (1988). Academic Writing _ Can we Please all the People all the Time? *ESL Documents* 129, 24-34.

Weir, C. (1990). Communicative Language Testing. Prentice-Hall, London.

Weir, C.J. (1993). *Understanding and Developing Language Tests*. London: Prentice Hall.

White, E. (1984). Holisticism. *College Composition and Communication* 35, 400-409.

White, E.M. (1985). *Teaching and Assessing Writing*. San Francisco: Jossey-Bass.

Widowson, H.G. (1978). *Teaching Language as Communication*. Oxford: Oxford University Press.

Wigglesworth, G. (1993). Exploring Bias Analysis as a Tool for Improving Rater Consistency in Assessing Oral Interaction. *Language Testing*, Vol. 10, No 3.

Wilkins, D.A. (1973, 1976). *Notional Syllabuses*. Oxford University Press, Oxford.

Wilkinson, A. (1983). Assessing Language Development: The Credition Project. En A. Freedman *et al.* (ed.), *Learning to Write: First Language / Second language*. New York: Longman: 67-86.

Wiseman, S. (1949). The Marking of English Composition in English Grammar School Selection. *British Journal of Education Psychology* 19, 200-209.

Wu, Y. (1998). What do Tests of Listening Comprehension Test? A Retrospection Study of EFL Test-takers Performing a Multiple-choice Task. *Language Testing* 15(1), 21-44.

Zamel, V. (1976). Teaching Composition in the ESL Classroom: What We can Learn from Research in the Teaching of English. *TESOL Quartely*, 10, 67-76.

Zamel, V. (1982). Writing: The Process of Discovering Meaning. *TESOL Quartely*, 16, 195-209.

Zamel, V. (1983). The Composing Processes of Advanced ESL Students: Six Cases Studies. *TESOL Quartely*, 17, 165-187.

Zamel, V. (1985). Responding to Student Writing. *TESOL Quartely* 19 (1), 79-101

Zemelman, S. (1979). Writing in other Disciplines: A Questionnaire for Teachers. *Conference on Language Attitudes and Composition Newsletter.* Portland State University, 5, 12-16.

Ziv, N. (1984). The Effects of Teacher Comments of the Writing of Four College Freshmen. En R. Beach y L. S. Bridwell (Eds..), *New Directions in Composition Research*. New York: Guildford Press: 362-380.

Página WEB:

http: //www.surrey.ac.uk/ELI/ltr.html

Técnica estadística:

SPSS 11.0.1. *Statistical Package for the Social Sciences*

Apéndice 1.

Instrucciones

Recuerde:

- Rellene primero el cuestionario. Tan sólo tendrá que completarlo una vez
- No olvide contestar todas las preguntas. (Las respuestas son totalmente confidenciales).
- Corrija los ensayos siguiendo las indicaciones que se señalan en las mismas. El nivel de dominio de la lengua de los alumnos es el de SELECTIVIDAD
- POR FAVOR, CORRIJA LOS ENSAYOS SIGUIENDO EL ORDEN NUMÉRICO DE LOS MISMOS
- Entregue los ensayos corregidos dentro del sobre facilitado junto con el cuestionario

¡MUCHAS GRACIAS POR SU COLABORACIÓN!

Apéndice 2.

CUESTIONARIO SOBRE LA CORRECCIÓN DE LOS ENSAYOS EN LENGUA INGLESA

I Parte

Por favor rellene la información que se indica en este apartado anotando una X en el lugar correspondiente. No es necesario que escriba su nombre.

Sexo: □ Hombre □ Mujer

Edad: □ 20-30 □ 31-40 □ 41-50 □ 51-60

Lengua nativa/ L1: _____

Estudios:

Doctorado: □ Sí □ No □ En proceso

Otras titulaciones: _____

Profesión actual: □ Profesor/a E. Secundaria □ Profesor/a Universidad

Años que lleva trabajando en el puesto actual:

□ De 0 a 5 □ De 6 a 10 □ De 11 a 15 □ Más de16

II Parte. Actitud ante la corrección de ensayos en lengua inglesa
Le agradeceríamos que señalara su opinión sobre los siguientes enunciados. Para ello, lea cada enunciado y responda anotando el grado de acuerdo con cada uno de ellos respecto a la escala indicada.

A) Estilo de enseñanza
Por favor rodee con un círculo el número de la escala elegido de 1 (totalmente en desacuerdo) a 5 (totalmente de acuerdo).

1. Prefiero actividades como la conversación a aquellas que conlleven análisis y aplicación de reglas 1 2 3 4 5

2. Considero fundamental la corrección en el uso de la lengua y me molesta mucho cometer errores 1 2 3 4 5

3. Normalmente escribo los elementos clave en la pizarra y ofrezco a los estudiante ejemplos gráficos que les ayuden a entender
los nuevos conceptos 1 2 3 4 5

4. Hago poco uso de la pizarra y de los medios audiovisuales; prefiero la conferencia, el debate y el discurso oral para
impartir los conocimientos 1 2 3 4 5

5. Me gusta desplazarme mientras explico ofreciendo actividades variadas en la clase 1 2 3 4 5

6. Favorezco la planificación, el orden y el trabajo bien hecho; las sorpresas no me agradan 1 2 3 4 5

7. Prefiero trabajar en compañía más que solo y a menudo me resulta difícil concentrarme durante un largo período de tiempo
en una misma tarea 1 2 3 4 5

8. Normalmente intento buscar relaciones entre hechos y datos para llegar a tener una visión global de las cosas 1 2 3 4 5

9. Trabajo de forma sistemática el planteamiento de mis lecciones y busco que todo salga de acuerdo al plan establecido 1 2 3 4 5

10. Me agrada la sistematicidad en la preparación y presentación de las lecciones y raramente me suelo desviar de los objetivos
específicos de cada lección 1 2 3 4 5

11. Prefiero la flexibilidad en el trabajo y me molesta trabajar con fechas que tengan un fin de plazo determinado 1 2 3 4 5

B) Autoevaluación del profesorado

Por favor rodee con un círculo el número de la escala que mejor refleje su opinión de 1 a 5 (5= totalmente de acuerdo).

1. Soy una persona optimista	1 2 3 4 5
2. Me gusta mi trabajo y me parece estimulante	1 2 3 4 5
3. Me siento confiado y seguro en el trabajo	1 2 3 4 5
4. Tengo una gran autoestima	1 2 3 4 5
5. Me siento realizado en mi trabajo y este me resulta enriquecedor	1 2 3 4 5

C) Autoevaluación del profesorado como corrector

Por favor rodee con un círculo UNO de los números que mejor refleje su opinión de 1 a 5 en cada una de las escalas siguientes:
(Recuerde: no debe dejar ninguna pregunta sin contestar; asegúrese de que rodea un número para cada par de palabras).

1. Mi dominio de la lengua inglesa es:						
Escaso	1	2	3	4	5	muy bueno
2. Me considero un corrector:						
condescendiente	1	2	3	4	5	duro
3. Me considero un corrector:						
inflexible	1	2	3	4	5	flexible
4. Me considero un corrector:						
Consecuente	1	2	3	4	5	inconsecuente
5. Me considero un corrector:						
No cualificado	1	2	3	4	5	experto

D) Identificación de aspectos problemáticos

Por último, señale los TRES principales problemas o dificultades que presentan los alumnos a la hora de escribir un ensayo en lengua inglesa de acuerdo con los siguientes enunciados:

1. Carencia de conocimientos generales previos
2. Problemas con la identificación del objetivo/propósito general del ensayo
3. Falta de planificación
4. Dificultad a la hora de hacer inferencias y establecer conexiones entre la realidad, conocimientos previos y los distintos aspectos del texto
5. Carencia de vocabulario básico o esencial
6. Incapacidad de ser objetivos y desligarse emocionalmente de sus propias creencias y valores

Por favor, anote el número de los enunciados elegidos por orden de preferencia:

1º □
2º □
3º □

Asimismo, indique cualquier otro criterio que considere relevante y desee añadir

_____

Apéndice 2

EVALUACIÓN DE LOS ENSAYOS

PUNTUACIÓN GLOBAL
Por favor, juzgue el ensayo que acaba de leer basándose en la impresión general o global que este le produce de acuerdo a una escala de **1** (pésimo) a **10** (muy bueno).
Por favor recuerde: NO haga uso de números decimales

PUNTUACIÓN GLOBAL DEL ENSAYO: ☐

PUNTUACIÓN DE ACUERDO A LA SIGUIENTE ESCALA ANALÍTICA
Por favor, rodee con un círculo un número de **1** a **5** de acuerdo con la opinión que le merece la actuación del estudiante en cada uno de los aspectos que se mencionan. Por favor, asegúrese de que rodea un número para cada uno de los aspecto que se señalan.

1. Malo 2. Flojo 3. Mediano 4. Bueno 5. Muy bueno

1. CONTENIDO (relevante, bien argumentado, interesante)	1	2	3	4	5
2. ORGANIZACIÓN (claridad, coherencia, buen desarrollo, párrafos)	1	2	3	4	5
3. GRAMÁTICA (corrección, complejidad de estructura oracional)	1	2	3	4	5
4. VOCABULARIO (corrección, riqueza, variedad)	1	2	3	4	5
4. REGISTRO (formal/coloquial, apropiado)	1	2	3	4	5
5. MECÁNICA (puntuación, ortografía)	1	2	3	4	5
6. PRESENTACIÓN (descuidada, ordenada/limpia, confusa)	1	2	3	4	5

Por último, ¿qué término, expresión o estructura del ensayo destacaría positivamente?

_____

_____

A su entender ¿cuál es el término, expresión o aspecto más negativo que destacaría en este ensayo?

_____

_____

III Parte

Finalmente, en estos diez enunciados le presentamos en letra cursiva errores desde un punto de vista morfo-sintáctico, léxico, discursivo o de estilo que anotamos en el examen de Selectividad. Le agradeceríamos que puntuase la gravedad de los mismos de acuerdo a una escala de **1** (sin importancia) a **5** (muy importante).

1. I want to go to London to see *the Statue of Liberty.* 1 2 3 4 5

2. Thank you, I will accept your *invitacion.* 1 2 3 4 5

3. If I went to London I would *bought* a lot of things. 1 2 3 4 5

4. I like speaking *english* very much. 1 2 3 4 5

5. I find London very interesting but the food is *wrong.* 1 2 3 4 5

6. I would visit all the museums, *furthermore* Big Ben. 1 2 3 4 5

7. *I have been to London and I have visited a few shops. Perhaps you think I'm stupid but don't worry about that*. (style). 1 2 3 4 5

8. One of my main *hopenness* is to visit London 1 2 3 4 5

9. I like London. *However,* my parents and I will visit it next year 1 2 3 4 5

10. *I don't like London and I think it's a horrible island* (content) 1 2 3 4 5

11. A: Are you sure it would not be a problem for your cousin to put me up?
B: *No, silly*. (style) 1 2 3 4 5

12. I would visit the most *importants* places in London. 1 2 3 4 5

Apéndice 4.

Criterios evaluativos para la corrección de los ensayos (junio, 2000)

Se valorará primordialmente: el manejo del léxico, la organización de ideas, la coherencia y la capacidad de transmitir un mensaje así como la creatividad y el grado de madurez. A modo indicativo, señalamos la siguiente distribución de puntuaciones:

0,5 por la adecuación al número de palabras estipulado (100/150)
0,5 por la corrección ortográfica
1,5 por la corrección sintáctica y la organización de ideas
1,5 por la utilización adecuada del léxico, su riqueza y creatividad.