Análisis de la prueba de inglés de selectividad de la Universitat de les Illes Balears

Marian Amengual Pizarro
Universitat de les Illes Balears
marian.amengual@uib.es

Resumen

Este estudio se propone investigar la validez y la fiabilidad de la prueba de Inglés (PI) que se incluye en las Pruebas de Acceso a la Universidad (PAAU). 16 correctores participaron en las dos fases de este trabajo. En la primera fase, se analizaron las puntuaciones de 10 correctores que evaluaron 50 PI en las PAAU de la Universitat de les Illes Balears (UIB), junio de 2002. En la segunda fase, se examinaron las puntuaciones de 6 correctores que participaron en la corrección de las 50 PI de las PAAU en septiembre de 2002. Los resultados demuestran que la PI discrimina a los candidatos de acuerdo con su nivel de dominio lingüístico de la lengua. Los resultados también indican que la discriminación de los candidatos se establece principalmente a partir de las preguntas de carácter subjetivo que se incluyen en la Prueba. No obstante, la fiabilidad de dichas preguntas es cuestionable. Este último dato nos obliga a estudiar distintas medidas destinadas a garantizar la fiabilidad que esta prueba de dominio de le lengua inglesa debería tener.

Palabras clave: Pruebas de Acceso a la Universidad, evaluación, corrector, validez, fiabilidad

Abstract

Analysis of the English Test in the Spanish University Entrance Examination at the University of the Balearic Islands

This study aims to investigate the validity and the reliability of the English Test (ET) in the Spanish University Entrance Examination (SUEE). 16 raters participated in two different data collection sessions. In the first session, the ratings of 10 raters assessing 50 different ET of the University of the Balearic Islands Entrance Examination (UBIEE, June 2002) were analysed. In the second session the ratings of 6 raters judging 50 different ET of the UBIEE (September 2002) were examined. The results show that the ET discriminates among students and, therefore, ranks students according to their linguistic proficiency. Results also reveal that the discrimination of students' performance appears to rest on the subjective items of the ET. However, the reliability of those items

is called into question. This last finding obliges us to study different alternatives in order to guarantee the reliability that this English proficiency Test should have.

Key words: University Entrance Examination, testing, rater, validity, reliability

1. Introducción

La Prueba de Inglés (PI) forma parte junto con otras pruebas pertenecientes a asignaturas diversas (Lengua Castellana y su literatura, Historia, Filosofía, etc.) de las Pruebas de Acceso a la Universidad (PAAU) o Selectividad. Estas pruebas, de carácter común y obligatorio en todo el estado español, se plantean un doble objetivo: en primer lugar, homogeneizar las calificaciones obtenidas por los estudiantes al término de la enseñanza secundaria o formación profesional de segundo grado o grado superior[1] y, en segundo lugar, discriminar a los candidatos en función de los resultados obtenidos. En este sentido, los resultados que se obtengan en las PAAU, van a condicionar la carrera universitaria de no pocos estudiantes españoles.

La PI se puede definir como una prueba de dominio de la lengua (*proficiency test)*. A diferencia de las pruebas de rendimiento o progreso (*achievement tests*), que evalúan el dominio de la materia estudiada, las pruebas de dominio lingüístico evalúan el dominio real de la habilidad que se pretende medir. Las pruebas de dominio lingüístico suelen ser de carácter normativo ya que, al contrario de lo que sucede con las pruebas criteriales,[2] se intenta relacionar la actuación de un candidato con la del resto de los candidatos de modo que ambos tipos de actuación puedan compararse. Su objetivo principal es conseguir altos niveles de discriminación entre los candidatos y repartir las puntuaciones a lo largo de una curva de distribución normal (Gipps, 1994; Herrera, 1999).[3]

De acuerdo con este planteamiento, la PI pretende discriminar entre los candidatos de la forma más fiable que sea posible. Se busca con ello obtener una puntuación numérica o calificación fiable que permita a las autoridades académicas clasificar a los candidatos de acuerdo con su nivel de dominio de la lengua inglesa. Cabe señalar que los candidatos van a poder elegir la Facultad en la que deseen cursar sus estudios universitarios de acuerdo con la puntuación obtenida en las PAAU. De ahí, la importancia social que se le atribuye a las mismas. De hecho, el problema que se plantea en las PAAU no es el de aprobar o suspender, dado que el porcentaje de aprobados en estos últimos años se sitúa, por lo general, en torno al 85-90%. La importancia de las PAAU reside en el resultado final que los estudiantes obtienen en dichas Pruebas, ya que este último condicionará su futuro académico, laboral e incluso personal. Conviene recordar en este punto que, en ocasiones, son décimas o

incluso centésimas de punto las que impiden a un candidato acceder a la carrera que realmente desea cursar.

Así pues, y dado que vamos a sacar conclusiones y tomar decisiones a partir de los resultados de las PAAU, en general, y de la PI, en particular, una de nuestras responsabilidades profesionales y éticas fundamentales será la de asegurarnos de que las puntuaciones finales que se obtienen son válidas y, por lo tanto, razonablemente fiables (ILTA, 2000; Alderson & Banerjee, 2001).

La Prueba de Inglés (PI)

Si examinamos el diseño y la estructura de la PI que se desarrolla en las distintas universidades españolas, podremos observar que, en general, la PI evalúa la capacidad de los candidatos para comprender y expresarse por escrito en lengua inglesa. Ello permite a los estudiantes universitarios de nuevo ingreso acceder a la comprensión de textos o publicaciones de interés para su formación universitaria.

No obstante, se ha de admitir que el diseño de la PI de Selectividad no se corresponde exactamente con los objetivos y contenidos que se desarrollan en la asignatura de Lengua Inglesa que se imparte en 2º de Bachillerato. Una clara prueba de ello lo constituye el hecho de que, por ejemplo, el aspecto oral de la lengua no se evalúe en la PI de Selectividad. De hecho, se puede decir que más que basarse en los modelos de competencia comunicativa (Canale & Swain, 1980; Canale, 1983), la PI se fundamenta en un único componente del modelo de competencia comunicativa (Bachman, 1990), que se identifica como la competencia lingüística[4] (Herrera, 1999). El elemento principal de esta última es el de la competencia organizativa, que incluye a su vez la competencia gramatical y la competencia textual, las cuales se subdividen, finalmente, en gramática, léxico, comprensión lectora y redacción, para dar una descripción más detallada de las habilidades que se quieren medir o del constructo.[5] Este es el marco operativo en el que se basan la mayoría de las PAAU aunque la naturaleza de las preguntas de la PI y los criterios evaluativos pueden variar entre las distintas universidades españolas hasta cierto punto.

Las condiciones que se establecen para la realización de las PAAU y, por ende, para la PI, son las habituales en este tipo de Pruebas, es decir, control e identificación de los candidatos, administración de la Prueba, vigilancia, adaptación para los estudiantes con discapacidad, etc. Los candidatos disponen de una hora y treinta minutos para la realización de la PI. Con el fin de evitar cualquier tipo de sesgo o discriminación, se omite todo tipo de información personal sobre los candidatos

(nombre, género, nacionalidad, etc.). De este modo, se asegura el anonimato de los mismos durante todo el proceso de evaluación.

El diseño de la PI debe reunir los requisitos básicos y esenciales de toda prueba de dominio: validez y fiabilidad. Esto le permitirá justificar el uso ético posterior que se haga de las calificaciones obtenidas en la Prueba (Messick, 1989). Dada la trascendencia de estas Pruebas y las consecuencias e implicaciones, tanto personales como sociales, que se derivan del uso que se hace de las puntuaciones finales obtenidas, resulta ciertamente desconcertante observar las pocas investigaciones llevadas a cabo dirigidas a garantizar la validez y fiabilidad de las mismas (Herrera, 1999; Sanz, 1999; Amengual, 2004). Conscientes de ello, la Universitat de les Illes Balears (UIB) decidió llevar a cabo este estudio con el fin de recoger datos empíricos destinados a comprobar el funcionamiento de la PI y mejorar, de este modo, la creación de las pruebas futuras.

Componentes de la Prueba de Inglés (PI)

La PI presenta una opción única de examen y se basa en un solo texto, de interés general y lenguaje común no especializado, a partir del cual se formulan cinco preguntas. El candidato tiene que contestar por escrito y en lengua inglesa sin la ayuda del diccionario u otro material didáctico. La estructura de la PI de la UIB sobre la que se basa este estudio es la siguiente:

- La 1ª pregunta pide a los candidatos realizar un resumen del texto original de una extensión inferior a las cincuenta palabras. No se permite la copia literal del texto.
- La 2ª pregunta es una pregunta de comprensión que se realiza mediante la técnica verdadero/falso. No obstante, para contrarrestar el factor azar, los candidatos deben justificar su repuesta a partir del texto.
- La 3ª pregunta es una pregunta de comprensión abierta. Los candidatos no pueden copiar literalmente las frases del texto.
- La 4ª pregunta consta de dos apartados:
  - Uno de ellos (4º a) se relaciona con el léxico y se valora el dominio del vocabulario que tienen los candidatos. Se pide a los candidatos que encuentren una palabra o expresión en el texto que sea sinónima a las palabras o expresiones que se les señala.
  - El segundo apartado (4º b) se relaciona con los aspectos gramaticales (morfológicos y sintácticos) del texto y se subdivide, a su vez, en dos

subapartados (4b1 y 4b2). En dichos apartados, se pide a los candidatos que transformen o completen frases, que hagan una remodelación o *rephrasing* o que hagan preguntas para respuestas determinadas.

- La 5ª pregunta consiste en el desarrollo de una única opción: una redacción o un diálogo en lengua inglesa, ambos basados en un tema relacionado con el texto de la Prueba. La extensión de este ejercicio es de 120-150 palabras aproximadamente.

Criterios de evaluación

Los criterios específicos de corrección para la PI se encuentran resumidos en la Tabla 1 que se presenta a continuación. En dicha Tabla se especifica, además, la naturaleza objetiva o subjetiva de cada una de las preguntas y la técnica que se utiliza para evaluar cada una de las competencias específicas.

Pregunta	Puntuación	Tipo de ítem	Técnica
1	0-2	Subjetivo	Resumen
2	0-1	Objetivo	Verdadero/Falso
3	0-1	Subjetivo	Pregunta abierta
4a	0-1	Objetivo	Sinónimos
4b	0-1	Objetivo	Gramática
5a	0-4	Subjetivo	Redacción
o	o		
5b	0-4	Subjetivo	Diálogo

Tabla 1. Componentes de la PI.

La puntuación máxima de la PI es de 10 puntos y los candidatos deben obtener una calificación definitiva de 5 puntos o superior para considerar dicha Prueba superada.[6] Los correctores pueden hacer uso de notas enteras y de fracciones de medio punto. Las puntuaciones máximas que corresponden a cada una de las preguntas de la PI se indican entre paréntesis al final del enunciado de las mismas. Los correctores deben corregir entre cien y ciento ochenta Pruebas en cada convocatoria y se han de comprometer a entregar las Pruebas corregidas en un plazo máximo de cinco días a partir del inicio de las PAAU.

Como se aprecia en la Tabla 1, arriba descrita, un 30 % de la calificación de la PI se compone de preguntas de naturaleza objetiva (verdadero/falso, léxico y gramática) y el 70% restante, de preguntas de carácter subjetivo (resumen, pregunta de comprensión abierta y redacción/diálogo). En principio, a las pruebas objetivas se les atribuye una mayor fiabilidad que a las subjetivas. Por el contrario, las pruebas subjetivas se

consideran más válidas que las objetivas ya que son las que evalúan la producción real de los candidatos, de ahí que tengan un mayor peso en la puntuación final de la Prueba.

En las PAAU, con la finalidad de homogeneizar las valoraciones de los diversos correctores se estipulan unos criterios de corrección que, además de la puntuación concreta de cada pregunta, incluyen una serie de descriptores sobre los aspectos o elementos concretos que deben evaluarse.

A pesar de que se dedican sesiones al comentario y discusión de los criterios evaluativos, el reglamento oficial de las PAAU no contempla la formación obligatoria del profesorado en técnicas de evaluación para poder participar en la corrección de la PI. Ello dificulta la obtención de puntuaciones homogéneas, especialmente en las preguntas de naturaleza subjetiva como, por ejemplo, la redacción/diálogo, dada la gran cantidad de factores fortuitos y aleatorios que se asocian con la figura del corrector (cansancio, prejuicios, rigor, benevolencia, etc.) (Moss, 1994; Lumley & McNamara, 1995; Milanovic et al., 1996; Gamaroff, 2000; Amengual, 2003, 2004).

Por el contrario, en la evaluación de las preguntas de naturaleza objetiva, se prevé, normalmente, obtener unos resultados más homogéneos entre los distintos correctores ya que el margen de actuación de estos últimos se halla más limitado en este tipo de preguntas.

Objetivos

Este trabajo se plantea como objetivo fundamental analizar la validez relacionada con el poder discriminatorio de cada pregunta que se incluye en la PI y la fiabilidad o grado de consistencia que existe entre los correctores (fiabilidad inter-corrector) que evalúan dicha Prueba.

Se pretende averiguar, en primer lugar, si las preguntas que se incluyen en la Prueba discriminan de forma adecuada y, por lo tanto, reparten las puntuaciones en una curva de distribución normal y, en segundo lugar, si las puntuaciones entre los diversos correctores de la Prueba son consistentes y fiables.

2. Método

Los datos de este estudio se obtienen a partir del análisis de las PI pertenecientes a las PAAU de la UIB durante las convocatorias de junio y septiembre de 2004 (ver Apéndices 1 y 2). Para el tratamiento de los datos se utilizó la técnica estadística

Statistical Package for Social Sciencies (SPSS 11.0.1). Se aplicaron las correlaciones de Pearson, el análisis de la fiabilidad interna de la Prueba y el análisis de la varianza (ANOVA). Asimismo, se analizaron los estadísticos descriptivos, histogramas, y diagramas de barras de los resultados obtenidos en este estudio.

Sujetos

Sobre una población total de 2.466 sujetos, 500 candidatos concursaron en las PAAU de junio; sobre una población total de 924 sujetos, 300 candidatos se presentaron en las PAAU de septiembre.

De los 15 correctores que formaban parte del tribunal corrector de las PAAU de junio de 2004 en la UIB, 10 participaron de forma voluntaria en este estudio. Todos los correctores procedían de centros públicos de Enseñanza Secundaria de les Illes Balears a excepción de uno de ellos que era profesor de la UIB.

En la convocatoria de las PAAU de septiembre de 2004 en dicha universidad, 6 de los 9 correctores pertenecientes al tribunal corrector aceptaron participar en la segunda parte de este estudio. De nuevo, todos los correctores procedían de centros de Enseñanza Secundaria salvo uno de ellos que era profesor de la UIB. A todos los correctores se les garantizó el anonimato de sus actuaciones y se les agradeció su intervención.

Procedimiento

La recopilación del material para este estudio se efectuó en dos partes: junio y septiembre de 2004.

A todos los participantes se les pidió que corrigieran las PI, las cuales les habían sido asignadas de forma aleatoria por el Comité Organizador de las Pruebas, y que anotaran las puntuaciones parciales así como la puntuación final de las 50 primeras PI corregidas. En la convocatoria de septiembre de 2004, se repitió el proceso y se procedió del mismo modo.

De este modo, se consiguió una muestra de 500 ejercicios en la convocatoria de junio y una muestra de 300 ejercicios en la convocatoria de septiembre, lo cual se considera una muestra muy razonable para cualquier trabajo empírico a efectos de generalización de resultados.

3. Resultados de la Prueba de Inglés (PI) de Selectividad de junio: justificación psicométrica

Los datos que presentamos a continuación corresponden a la PI de Selectividad realizada en junio de 2004 (ver Apéndice 1).

Puntuaciones totales de la Prueba de Inglés (PI)

En una primera aproximación a los datos, se detallan los principales estadísticos descriptivos (Tabla 2) basados en las puntuaciones totales pertenecientes a las distintas preguntas que se incluyeron en la PI de Selectividad (junio, 2004).

	N	Rango	Media	Desv.Típ.	Asimetría		Curtosis	
	Estad.	Estad.	Estad.	Estad.	Estad.	Err. T.	Estad	Err. T.
Resumen	500	2,00	1,1055	,5216	−,084	,109	−,959	,218
Verdadero/Falso	500	1,00	,7220	,2571	−,684	,109	−,159	,218
Pregunta abierta	500	1,00	,6340	,3130	−,466	,109	−,774	,218
Sinónimos	500	1,00	,5750	,3026	−,299	,109	−,877	,218
Gramática	500	1,00	,2005	,3144	1,331	,109	,616	,218
Redacción	426	4,00	2,1626	1,0346	−,020	,118	−,897	,236
Diálogo	74	4,00	1,8750	1,0536	,066	,279	−,989	,552
TOTAL	500	9,00	5,3795	2,0450	,051	,109	−,623	,218
N válido (según lista)	0							

Desv.Típ. = Desviación Típica; Estad. = Estadístico; Err. T. = Error Típico.

Tabla 2. Estadísticos descriptivos de la PI.

Como puede observarse, la media de las puntuaciones más baja hace referencia al dominio de la gramática ($\bar{x}$ = 0,20). Este resultado nos sugiere, a simple vista, que los correctores fueron más estrictos juzgando dicha categoría. El escaso valor de la media de las puntuaciones globales referida a los aspectos gramaticales, también pone de manifiesto la dificultad que plantea a los candidatos la resolución correcta de los problemas estructurales de la lengua. Este dato resulta interesante dada la importancia que se le suele conceder habitualmente a estos últimos aspectos en el desarrollo diario de las clases de lengua extranjera.

El resto de las preguntas parece no plantear, en general, graves problemas de resolución a los candidatos ya que la media de las puntuaciones totales de cada una de ellas se sitúa ligeramente por encima de la mitad de su valor total.

La única excepción se halla en la pregunta del diálogo, la cual se ofrece como alternativa al desarrollo de la redacción. Como puede observarse, la media correspondiente a la pregunta del diálogo ($\bar{x}$ = 1,87) es ligeramente inferior a la media obtenida en la pregunta de la redacción ($\bar{x}$ = 2,16). Esto es, los candidatos que optaron por realizar la redacción obtuvieron puntuaciones más altas que aquellos que

desarrollaron el diálogo. Dado que el contenido de los temas a tratar en ambas opciones era muy similar (ver pregunta 5, Apéndice 1), la diferencia de puntuaciones pudiera deberse a la aplicación de criterios de corrección más estrictos en la evaluación de este último tipo de texto. No obstante, estos resultados deben interpretarse con cautela ya que la muestra de los ejercicios de diálogo (T = 74) es muy inferior a la de los ejercicios de redacción (T = 426).

De hecho, este último dato resulta interesante ya que destaca la preferencia de la gran mayoría de los candidatos por desarrollar un tipo de texto frente a otro. Como se verá posteriormente, esta misma tendencia se repite en septiembre. Una posible explicación de estos resultados puede que se deba al nivel de conocimiento de la lengua inglesa de los candidatos. Probablemente, los candidatos con un mayor dominio de la lengua inglesa realizan el ejercicio de redacción porque desean demostrar sus conocimientos y habilidades en el uso de dicha lengua. Los candidatos con un dominio de la lengua inglesa inferior, por el contrario, optan por el ejercicio del diálogo que les permite expresar algunos conocimientos básicos en inglés sin la necesidad de utilizar construcciones discursivas y sintácticas complejas. Puede que esto último explique también el hecho de que la nota media del diálogo sea inferior a la de la redacción en ambas convocatorias.

Por último, cabe señalar que la pregunta verdadero/falso es la que obtiene la puntuación media más alta ($\bar{x}$ = 0,72). El valor de la desviación típica de esta pregunta es también muy bajo, lo que indica que hay poca dispersión entre las puntuaciones de los distintos correctores, es decir, que las puntuaciones tienden a ser homogéneas y no se observan grandes discrepancias. Estos resultados son lógicamente previsibles ya que la naturaleza objetiva de esta última pregunta deja a los correctores poco margen de actuación para interpretaciones subjetivas o juicios personales.

Por otra parte, la puntuación media de la pregunta verdadero/falso, al ser mucho más alta que la del resto de las preguntas ($\bar{x}$ = 0,72), pone de manifiesto su escaso poder discriminatorio. En otras palabras, se trata de una pregunta fácil cuya resolución no parece plantear grandes dificultades a los candidatos.

En la figura 1 se representa el histograma de las puntuaciones totales obtenidas en la PI. Ello nos permite obtener una visión completa del conjunto de los resultados de una forma más plástica, lo que nos facilita su comprensión.

Como puede observarse, las puntuaciones totales dibujan una curva de distribución normal casi perfecta. En una prueba de dominio como la de Selectividad, el objetivo que se persigue es que las puntuaciones se aproximen a una distribución normal, que en este caso es un buen indicador de ausencia de sesgos de facilidad o dificultad de

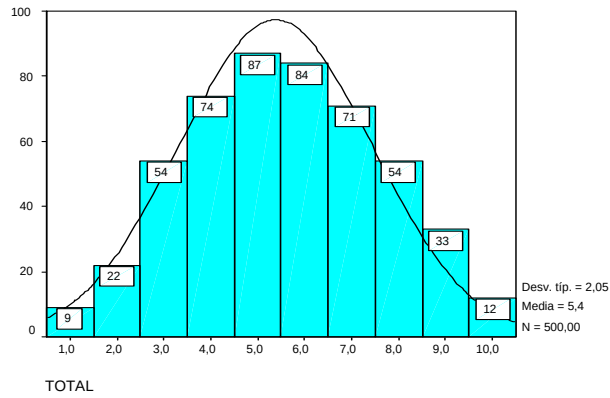

Figura 1. Puntuaciones totales de la Prueba de Inglés (PI) de junio.

la PI. A la vista de estos resultados (Figura 1), podemos afirmar que la PI de Selectividad de junio de 2004 cumplió su objetivo: distribuir a los estudiantes según su nivel de dominio de la lengua inglesa.

En la Figura 1, es prácticamente imperceptible una ligera asimetría[7] negativa (0,051) en la distribución de las puntuaciones ya que la curva se escora ligeramente hacia la derecha. Ello nos indica que el nivel de dificultad de la Prueba ha sido bastante asequible para el grupo en cuestión. Estos resultados son, hasta cierto punto, comprensibles ya que los candidatos que concursan en dicha Prueba son los que han superado con éxito la etapa de la Educación Secundaria, segundo grado o grado superior de Formación Profesional. En la convocatoria de Selectividad de septiembre, por el contrario, se prevén puntuaciones globales algo más bajas que las de junio, ya que los candidatos de septiembre no fueron considerados idóneos para presentarse en junio.

Correlaciones

La Tabla 3 nos muestra las correlaciones[8] entre las distintas preguntas de la PI. Como puede observarse, las correlaciones son muy significativas a p 0,01 en todos los casos. Sin embargo, los valores que se obtienen son relativamente bajos.

Lógicamente, la pregunta de redacción/diálogo[9] presenta los valores más altos de correlación con las preguntas de carácter subjetivo, especialmente, con el resumen (0,644) y con la pregunta abierta (0,474). Los valores de correlación de la redacción/diálogo son más bajos con el resto de preguntas de carácter objetivo. De

		Resumen	Verdadero / falso	Pregunta abierta	Sinónimos	Gramática	Redacción / diálogo
Resumen	Correlación de Pearson	1,000	,367**	,457**	,279**	,305**	,644**
	Sig. (bilateral)	,	,000	,000	,000	,000	,000
	N	500	500	500	500	500	500
Verdadero/ falso	Correlación de Pearson	,367**	1,000	,344**	,246**	,235**	,314**
	Sig. (bilateral)	,000	,	,000	,000	,000	,000
	N	500	500	500	500	500	500
Pregunta abierta	Correlación de Pearson	,457**	,344**	1,000	,276**	,277**	,474**
	Sig. (bilateral)	,000	,000	,	,000	,000	,000
	N	500	500	500	500	500	500
Sinónimos	Correlación de Pearson	,279**	,246**	,276**	1,000	,293**	,399**
	Sig. (bilateral)	,000	,000	,000	,	,000	,000
	N	500	500	500	500	500	500
Gramática	Correlación de Pearson	,305**	,235**	,277**	,293**	1,000	,438**
	Sig. (bilateral)	,000	,000	,000	,000	,	,000
	N	500	500	500	500	500	500
Redacción/ diálogo	Correlación de Pearson	,644**	,314**	,474**	,399**	,438**	1,000
	Sig. (bilateral)	,000	,000	,000	,000	,000	,
	N	500	500	500	500	500	500

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

Tabla 3. Correlaciones entre las distintas preguntas de la Prueba.

entre estas últimas, la pregunta verdadero/falso es la que presenta los valores de correlación más bajos (0,314).

Fiabilidad interna de la Prueba

A continuación se llevará a cabo un análisis de fiabilidad que nos permitirá averiguar, por un lado, el grado de discriminación de cada una de las preguntas de la PI y, por otro, la contribución de dichas preguntas a la fiabilidad total e interna de la Prueba.

Los datos que se analizarán se presentan en la Tabla 4. En concreto, nos interesan los valores que nos muestran la columna tres y cuatro de dicha Tabla. La columna tres "correlación elemento-total corregida" nos facilita la correlación de cada elemento o pregunta de la Prueba con la puntuación total menos ese elemento o pregunta. Las correlaciones se miden en una escala de 0 a 1. Este último valor indica una relación perfecta.

Como se aprecia en la Tabla 4, las preguntas o elementos que mejor correlacionan con la puntuación total de la PI son las preguntas de carácter subjetivo, es decir, la redacción/diálogo (0,7070), el resumen (0,6567) y la pregunta abierta (0,5348) por este orden (ver Apéndices 1 y dos, preguntas 5, 1 y 3 respectivamente). La correlación más baja se registra en la pregunta verdadero/falso (0,4029), lo cual pudiera relacionarse con su escaso poder discriminatorio.

La última columna de la Tabla 4 "Alfa de Cronbach si se elimina el elemento" explica hasta qué punto la fiabilidad interna de la PI, aquí denominada Alfa, aumentaría o disminuiría si ese elemento o pregunta en concreto se eliminara. Como se observa en

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento
1. Resumen	4,2460	2,7527	,6567	,5953
2. Verdadero/falso	4,6295	3,6970	,4029	,6929
3. Pregunta abierta	4,7175	3,4422	,5348	,6633
4a. Sinónimos	4,7765	3,5838	,4242	,6842
4b. Gramática	5,1510	3,5268	,4537	,6776
5. Redacción/diálogo	3,2370	1,3568	,7070	,6728

Tabla 4. Fiabilidad interna de la PI.

la Tabla, los valores más bajos corresponden, de nuevo, a las preguntas de carácter subjetivo, esto es, el resumen (0,5953), la pregunta abierta (0,6633) y la pregunta de la redacción/diálogo (0,6728). Estos datos nos indican que, si se eliminaran dichas preguntas, el valor Alfa se reduciría considerablemente y, por consiguiente, disminuiría la fiabilidad interna de la Prueba.

De acuerdo con ello, la redacción/diálogo es la pregunta que parece contribuir en mayor medida a la fiabilidad interna de la Prueba. Por el contrario, la pregunta verdadero/falso, con el valor Alfa más alto (0,6929), es la que menos contribuye a la fiabilidad interna de la Prueba y, por tanto, su eficacia y utilidad debiera reconsiderarse.

En resumen, y a la vista de estos resultados, podemos afirmar que las preguntas de carácter subjetivo son las que mejor correlacionan con el valor total de la PI y las que mejor contribuyen a su fiabilidad interna.

Análisis de la varianza (ANOVA)

Una vez analizada la fiabilidad interna de la Prueba nos interesa averiguar si las diferencias que se observan en la evaluación de cada una de las preguntas de la PI son significativas desde un punto de vista estadístico o si, por el contrario, estas últimas se deben a fluctuaciones ocasionales. Para ello aplicaremos la técnica estadística conocida como ANOVA o análisis de la varianza. En nuestro estudio, la hipótesis nula será que no existen diferencias significativas entre las puntuaciones medias de los distintos correctores. Los resultados de ANOVA se presentan en la Tabla 5.

Como se puede apreciar, los datos nos indican que hay diferencias significativas en la evaluación de las preguntas subjetivas de la PI. Así pues, la hipótesis nula debe ser rechazada. En otras palabras, las diferencias que se observan entre los correctores en la evaluación de la pregunta del resumen, la pregunta abierta y la pregunta de la redacción/diálogo son significativas y, por lo tanto, no pueden atribuirse a factores fortuitos.

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Resumen	Inter-grupos	7,259	9	,807	3,076	,001
	Intra-grupos	128,489	490	,262		
	Total	135,747	499			
Verdadero/ falso	Inter-grupos	,950	9	,106	1,616	,108
	Intra-grupos	32,033	490	6,537E-02		
	Total	32,983	499			
Pregunta abierta	Inter-grupos	5,919	9	,658	7,499	,000
	Intra-grupos	42,978	490	8,771E-02		
	Total	48,897	499			
Sinónimos	Inter-grupos	,497	9	5,528E-02	,599	,798
	Intra-grupos	45,190	490	9,222E-02		
	Total	45,688	499			
Gramática	Inter-grupos	1,291	9	,143	1,463	,159
	Intra-grupos	48,046	490	9,805E-02		
	Total	49,337	499			
Redacción/ diálogo	Inter-grupos	46,344	9	5,149	5,085	,000
	Intra-grupos	496,164	490	1,013		
	Total	542,507	499			

Tabla 5. Análisis de la varianza (ANOVA).

Pruebas "post hoc"

ANOVA nos permite realizar comparaciones múltiples entre los correctores mediante las pruebas *post hoc*. Dichas pruebas nos permiten localizar exactamente las diferencias significativas detectadas por ANOVA. Dado que estas últimas se hallan en el resumen, la pregunta abierta y la de redacción/diálogo, todas ellas preguntas de naturaleza subjetiva, vamos a analizarlas con más detenimiento. Los gráficos que se presentan a continuación nos ofrecen una representación gráfica de los resultados que nos facilita su comprensión.

Como se aprecia en la Figura 3, los resultados revelan que en la evaluación de la pregunta del resumen existen diferencias significativas en la aplicación de los criterios de evaluación de los dos correctores más indulgentes (correctores 4 y 9) y el del corrector más estricto (corrector 10).

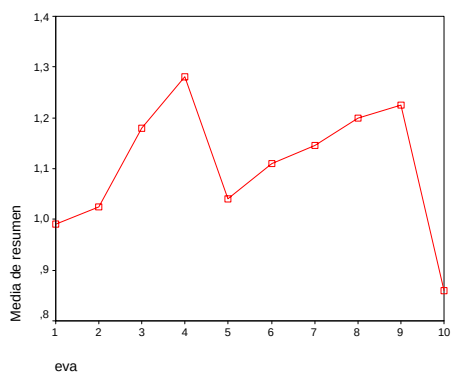

Figura 3. Gráfico de medias del resumen.

En cuanto a la evaluación de la pregunta abierta, la Figura 4 pone de manifiesto mayores discrepancias entre los distintos correctores. Como dato interesante cabe destacar que los correctores que presentan un mayor número de diferencias significativas con el resto son los correctores que aplican criterios de evaluación menos estrictos (correctores 1 y 6). Esto es, los criterios de evaluación que aplicaron dichos correctores fueron excesivamente indulgentes produciendo, por consiguiente, diferencias significativas con el resto de los correctores.

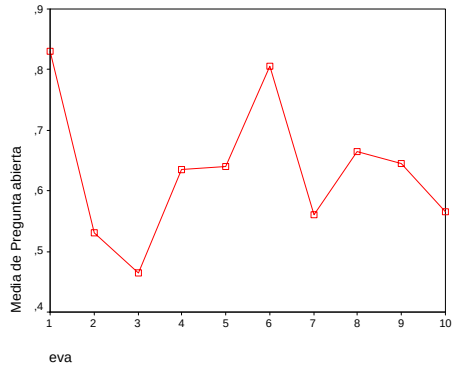

Figura 4. Gráfico de medias de la pregunta abierta.

La tendencia que se observa en la evaluación de la pregunta abierta se repite en la evaluación de la redacción/diálogo. Esta última pregunta es la que presenta un mayor número de diferencias significativas entre los correctores. De nuevo, tal como se observa en la Figura 5, los correctores que aplican los criterios de evaluación más indulgentes son los que mayores discrepancias muestran con el resto de los correctores.

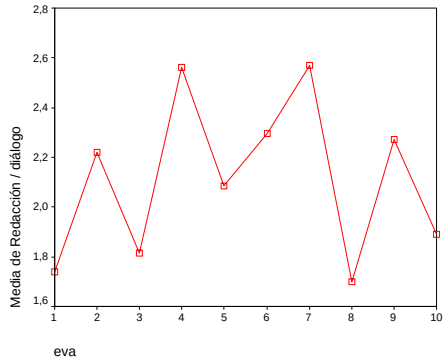

Figura 5. Gráfico de medias de la redacción/diálogo.

Así pues, y de modo sintético, estos últimos datos nos indican que, paradójicamente, el comportamiento excesivamente indulgente de algunos correctores incide negativamente en la evaluación de la PI de Selectividad de junio dado que disminuye de forma significativa la fiabilidad entre los diversos correctores (fiabilidad inter-corrector). Este hecho resulta interesante ya que, lejos de lo que pudiera pensarse, otorgar puntuaciones excesivamente altas provoca resultados tan injustos como los que pueda ocasionar otorgar puntuaciones excesivamente estrictas.

4. Resultados de la Prueba de Inglés (PI) de Selectividad de septiembre: justificación psicométrica

Análisis contrastivo de las puntuaciones totales de las Pruebas de Inglés (PI) de junio y septiembre

Los datos que se presentan a continuación (Tabla 6) corresponden a las PI de Selectividad de junio y septiembre de 2004. (Para más detalles véanse Apéndices 1 y 2). Dado que hemos aplicado los mismos procedimientos estadísticos a las PI en las convocatorias de junio y septiembre creemos más conveniente e ilustrativo realizar un análisis contrastivo de los resultados obtenidos en ambas convocatorias.

	Puntuación máxima de cada pregunta	Junio		Septiembre	
		Media	Desv. típ.	Media	Desv. típ.
Resumen	2	1,1055	0,5216	0,8133	0,5453
Verdadero/falso	1	0,7220	0,2571	0,5692	0,3087
Pregunta abierta	1	0,6340	0,3130	0,4850	0,3249
Sinónimos	1	0,5750	0,3026	0,4717	0,2666
Gramática	1	0,2005	0,3144	0,3800	0,3672
Redacción o	4 ó	2,1626	1,0346	1,6761	0,9285
Diálogo	4	1,8750	1,0536	1,4337	0,8366
TOTAL	10	5,3795	2,0450	4,4158	1,9563

Tabla 6. Estadísticos descriptivos: junio-septiembre.

A simple vista, el primer dato que destaca en la Tabla 6 es el hecho de que la media de las puntuaciones totales de cada una de las preguntas de la Prueba de septiembre es inferior a la media de las puntuaciones totales obtenidas en la Prueba de junio. Como ya avanzamos, este resultado resulta lógico dado que, excepto en ocasiones excepcionales, los estudiantes que realizan las pruebas de Selectividad de septiembre no se consideraron idóneos para concursar en junio por lo que el nivel de competencia de dichos candidatos es previsiblemente más bajo.

Resulta interesante apreciar que el único caso en el que la puntuación media de las preguntas de la Prueba de septiembre supera a la de junio es en la pregunta de gramática. La media de dicha pregunta es algo superior en la Prueba de septiembre ($\bar{x}$ = 0,3800) que en la de junio ($\bar{x}$ = 0,2005). Ello pudiera, quizás, atribuirse al grado de dificultad inherente en dicha pregunta. Probablemente, la pregunta de gramática en junio fue más difícil que la de septiembre. Sea como fuere, la gramática sigue siendo la categoría que se penaliza con mayor rigor o, quizás, la que presenta mayores problemas de resolución a los candidatos a juzgar por el escaso valor de las puntuaciones medias obtenidas en ambas convocatorias.

Por último, en la convocatoria de septiembre se observa que la puntuación media del diálogo ($\bar{x}$ =1,4337) es, de nuevo, ligeramente más baja que la de la redacción ($\bar{x}$ = 1,6761). Así pues, se repiten tendencias ya que, al igual que ocurriera en la Prueba de junio, la gran mayoría de candidatos opta por desarrollar la redacción (T = 247) frente al diálogo (T = 49). Esta notable diferencia en la elección de ambas opciones junto con la escasa respuesta positiva que obtiene el diálogo por parte de los candidatos nos obliga a replantearnos la necesidad de presentarlo como opción a la redacción en futuras convocatorias.

Porcentajes de los resultados de las preguntas subjetivas

Seguidamente, y antes de presentar el histograma de las puntuaciones totales obtenidas en la PI de septiembre, creemos ilustrativo mostrar algunos gráficos de barras comparando los porcentajes de algunas preguntas de la PI de Selectividad en las convocatorias de junio y septiembre. Por razones de espacio, nos centraremos exclusivamente en el análisis de las preguntas subjetivas dado que se les da un mayor peso en la puntuación final de la Prueba.

Como se puede apreciar en el gráfico de barras del resumen (Figura 6), los porcentajes más elevados correspondientes a las puntuaciones de la PI de septiembre se concentran en la parte izquierda del gráfico a partir de la mitad del valor total que se le asigna a la pregunta (1,00). Por el contrario, los porcentajes de las puntuaciones de la PI de junio se observan en la parte derecha del gráfico. A partir del valor 1,00, el porcentaje de las puntuaciones máximas es siempre superior en la PI de junio.

En la Figura 7, se repite la tendencia anterior si bien la distribución de los porcentajes de la PI de septiembre resulta algo más equilibrada. El incremento positivo de los porcentajes de la PI de junio es gradual. De hecho, el mayor porcentaje de puntuaciones en la convocatoria de junio se polariza en el valor 1,00 que representa el valor máximo que se le asigna a la pregunta abierta.

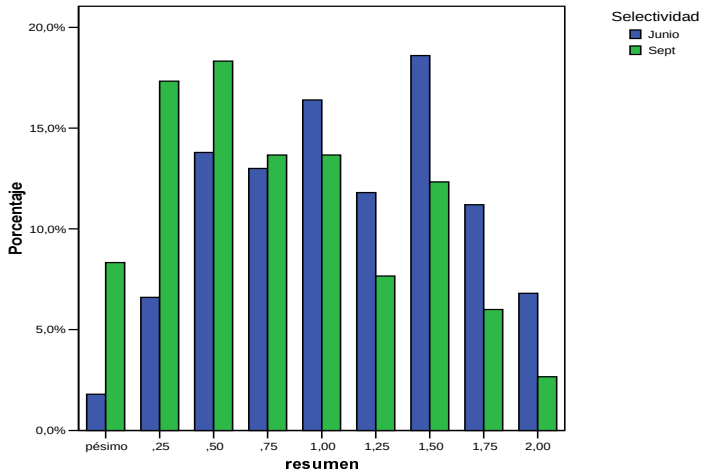

Figura 6. Gráfico de barras del resumen.

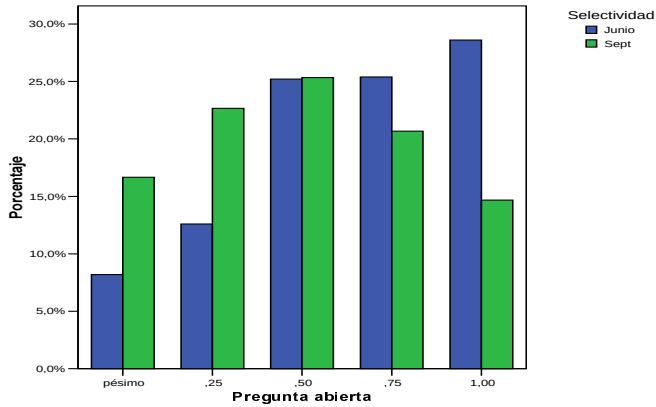

Figura 7. Gráfico de barras de la pregunta abierta.

Por último, la Figura 8 nos muestra los porcentajes relativos a la pregunta de la redacción/diálogo. Los resultados son similares a los observados hasta ahora. El incremento de los porcentajes de la PI de junio se detecta a partir del valor 2,5. A partir de este último valor, los porcentajes de la PI de junio son todos ellos superiores a los obtenidos en la PI de septiembre. Los porcentajes más elevados de la PI de esta última convocatoria se acumulan en el extremo izquierdo del gráfico a partir del valor 2,25.

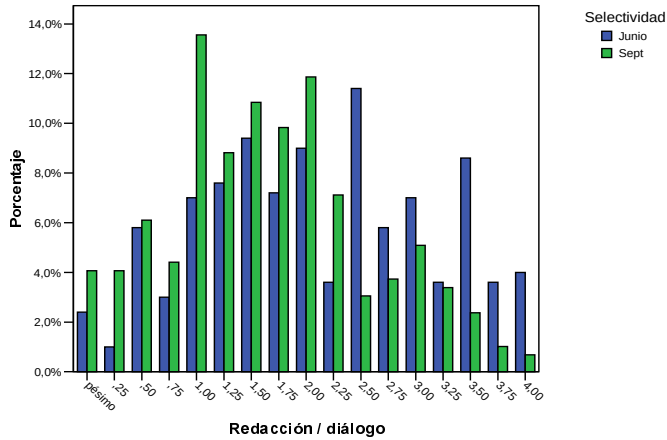

Figura 8. Gráfico de barras de la redacción/diálogo.

Los resultados que muestran estos gráficos sugieren que el nivel de competencia y conocimiento de la lengua inglesa de los candidatos es, en general, superior en la convocatoria de junio que en la de septiembre.

Puntuaciones totales de la Prueba de Inglés (PI) de septiembre

Una vez finalizado este análisis, presentamos el histograma de las puntuaciones totales obtenidas en la PI de septiembre.

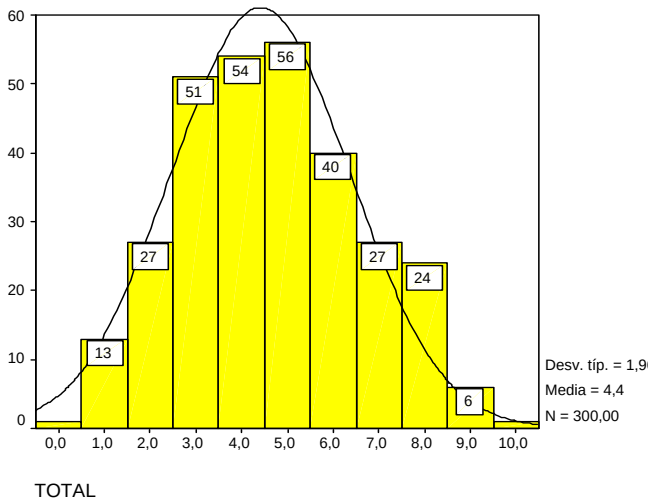

Figura 9. Puntuaciones totales de la PI de septiembre.

Al contrario de lo que ocurría en la Prueba de Selectividad de junio, en la Figura 9, puede observarse una ligera asimetría positiva de la distribución. La curva de distribución se escora ligeramente hacia la derecha, lo cual nos indica que hay un mayor número de puntuaciones bajas de lo que sería previsible en una distribución normal. Es decir, la mayoría de los candidatos encontró la Prueba algo difícil. Al tratarse de la convocatoria de Selectividad de septiembre, como ya se observó anteriormente, el nivel de competencia de los candidatos se prevé más bajo por lo que el grado de dificultad de la PI debería medirse teniendo en cuenta este último criterio.

Con el fin de agilizar la lectura de este análisis y dado que, tanto el estudio de las correlaciones como el de la fiabilidad interna de la PI de septiembre repite las tendencias observadas en la PI de junio, no creemos necesario incluir las tablas de estos últimos resultados.

Análisis de la varianza (ANOVA)

Así pues, y dada su mayor relevancia en este estudio, pasamos seguidamente a comentar los resultados obtenidos en el análisis de la varianza (ANOVA). Los resultados de ANOVA (Tabla 7) nos indican que hay diferencias significativas entre los correctores en la evaluación de las preguntas subjetivas, es decir, en el resumen, la pregunta abierta y la redacción/diálogo. Ello nos obliga a rechazar la hipótesis nula que establecía que no existen diferencias significativas entre las puntuaciones de los distintos correctores.

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Resumen	Inter-grupos	9,087	5	1,817	6,693	,000
	Intra-grupos	79,835	294	,272		
	Total	88,922	299			
Verdadero/ falso	Inter-grupos	2,814	5	,563	6,440	,000
	Intra-grupos	25,689	294	8,738E-02		
	Total	28,502	299			
Pregunta abierta	Inter-grupos	3,485	5	,697	7,300	,000
	Intra-grupos	28,072	294	9,548E-02		
	Total	31,557	299			
Sinónimos	Inter-grupos	,207	5	4,133E-02	,577	,717
	Intra-grupos	21,053	294	7,161E-02		
	Total	21,259	299			
Gramática	Inter-grupos	,373	5	7,450E-02	,549	,739
	Intra-grupos	39,933	294	,136		
	Total	40,305	299			
Redacción/ diálogo	Inter-grupos	31,913	5	6,383	8,640	,000
	Intra-grupos	213,491	289	,739		
	Total	245,404	294			

Tabla 7. Análisis de la varianza (ANOVA).

A diferencia de lo que ocurría en junio, y de forma aparentemente excepcional, los datos de este último análisis revelan que en la evaluación de la pregunta verdadero/falso

existen diferencias significativas entre los correctores. Por consiguiente, estas últimas no pueden atribuirse a factores fortuitos. Dado que la pregunta verdadero/falso es de carácter objetivo, estos últimos resultados parecen poco comprensibles ya que el criterio evaluativo que se aplica en la evaluación de este tipo de preguntas limita claramente las interpretaciones subjetivas y juicios personales de los correctores.

Pruebas "Post hoc"

Veamos, seguidamente, si la aplicación de las pruebas *post hoc* nos ayuda a entender estos últimos resultados. Tal y como hiciéramos en la PI de junio, aplicaremos este análisis a las preguntas en las que ANOVA ha detectado diferencias significativas. Los gráficos de estos resultados se muestran a continuación.

Como se observa en la Figura 10, los resultados indican que en la evaluación del resumen los correctores aplican, en general, criterios evaluativos heterogéneos. Es decir, no parece haber consenso entre los correctores a la hora de evaluar dicha pregunta por lo que las puntuaciones que se obtienen varían de forma significativa. Cabe destacar especialmente la actuación del corrector más estricto (corrector 1) ya que difiere de forma significativa con la del resto de los correctores.

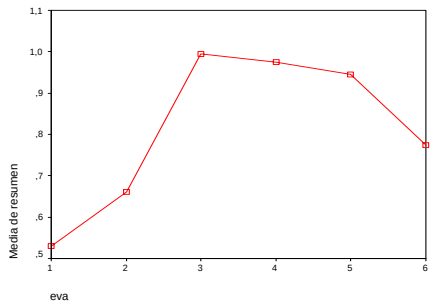

Figura 10. Gráfico de medias del resumen.

Por lo que respecta a la pregunta verdadero/falso, los datos de la Figura 11. demuestran que los correctores, en general, aplicaron criterios evaluativos homogéneos, tal y como se prevé en la evaluación de las preguntas objetivas. Las diferencias significativas detectadas por ANOVA en esta pregunta se localizan en un único corrector que, a juzgar por los resultados, aplica criterios de evaluación extremadamente estrictos que difieren de forma significativa con los del resto de los correctores sin excepción.

No se observan diferencias significativas en ningún otro caso. Así pues, el comportamiento extremo de uno de los correctores es el desencadenante de estos imprevisibles resultados. Sin embargo, este dato nos permite destacar la importancia que tiene la actuación individual de cada corrector en el cómputo total de unas Pruebas, que como ya dijimos, tienen una profunda trascendencia social y personal.

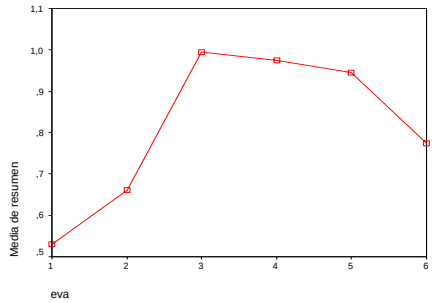

Figura 11. Gráfico de medias de la pregunta verdadero /falso.

La evaluación de la pregunta abierta revela también discrepancias significativas entre los correctores (Figura 12). Los datos nos indican que el corrector más indulgente (corrector 3) y el más estricto (corrector 2) respectivamente son los que mayores diferencias significativas presentan con el resto de los correctores. Como pudimos observar anteriormente (Figura 11), este último corrector resulta especialmente problemático ya que los criterios de evaluación que aplica son excesivamente estrictos si los comparamos con el resto de los correctores.

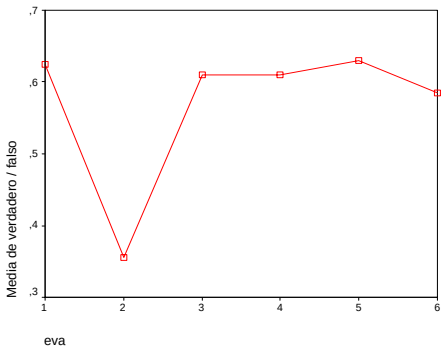

Figura 12. Gráfico de medias de la pregunta abierta.

Por ultimo, la evaluación de la pregunta de la redacción/diálogo (Figura 13), tal y como ocurría en la PI de junio, es la que registra un mayor número de diferencias

significativas. Ello nos indica que el grado de desacuerdo entre los correctores es mayor en la evaluación de esta última pregunta. Los correctores parecen guiarse por su experiencia personal o por interpretaciones subjetivas que varían de un corrector a otro de forma significativa. Ello produce resultados poco consistentes y fiables.

El gráfico también nos permite corroborar el comportamiento extremo del corrector más estricto que discrepa nuevamente de forma significativa con la actuación del resto de los correctores.

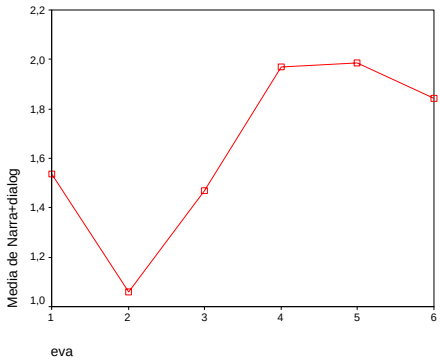

Figura 13. Gráfico de medias de la redacción /diálogo.

A la vista de estos resultados, cabe decir que, a diferencia de lo que ocurría en la PI de junio donde pudimos apreciar que la excesiva indulgencia de algunos correctores reducía sensiblemente la fiabilidad inter-corrector, en la PI de septiembre, se observa, por el contrario, que la aplicación de criterios evaluativos excesivamente rigurosos y estrictos por parte de algunos correctores disminuye la fiabilidad inter-corrector de forma significativa.

5. Discusión de resultados

La curva de distribución de las puntuaciones totales de la PI de Selectividad, tanto en la convocatoria de junio como en la de septiembre, se encuentra próxima a la de una distribución normal. Estos datos demuestran que se han evitado los posibles sesgos de facilidad/dificultad de las preguntas. Por consiguiente, se puede afirmar que la PI se halla bien estructurada y equilibrada ya que discrimina a los candidatos de acuerdo con su nivel de dominio de la lengua en ambas convocatorias.

No obstante, conviene señalar en este punto que el diseño actual de las PI de Selectividad de la mayoría de las universidades españoles mide básicamente un único componente del modelo de la competencia comunicativa (Bachman, 1990): la competencia lingüística. Es decir, las PI aquí analizadas discriminan a los candidatos de acuerdo con su nivel de competencia lingüística. Algunos autores consideran que este último criterio no es suficiente y critican la falta de consonancia que existe entre los objetivos y contenidos de la materia de lengua inglesa en 2º de Bachillerato y la PI de Selectividad (Sanz, 1999). En este sentido, pensamos que sería recomendable estudiar la viabilidad de evaluar el aspecto oral de la lengua inglesa que tanto peso tiene dentro del enfoque comunicativo que favorece el sistema educativo actual para la enseñanza de lenguas extranjeras.

Como hemos podido observar, en la convocatoria de la PI de junio (Figura 1), la curva de la distribución de las puntuaciones totales presenta una mínima asimetría negativa (0,051). Por el contrario, en la PI de septiembre (Figura 9), la curva de distribución se escora ligeramente hacia la derecha. Este último dato sugiere que los candidatos encontraron la Prueba de septiembre algo más difícil que la de junio o, muy probablemente, que el nivel de competencia del conjunto de los candidatos en septiembre fue algo inferior al de junio.

En cuanto a la importancia que reciben las distintas preguntas o componentes de la PI de Selectividad (Tabla 6), se observa que la pregunta de gramática obtiene la puntuación media más baja tanto en la convocatoria de Selectividad de junio como en la de septiembre. Estos resultados demuestran que, a pesar de la buena acogida del enfoque comunicativo en el campo de la enseñanza de segundas lenguas, los correctores parecen centrar su interés principal en los aspectos formales de la lengua que, por otra parte, son los que penalizan y evalúan con mayor dureza y rigor. En este sentido, se debiera trabajar con los correctores de la PI de Selectividad para conseguir acomodar las expectativas de forma y contenido de la lengua de modo que ambos aspectos se evalúen y obtengan el peso adecuado en la puntuación final de la Prueba.

Como dato interesante, cabe destacar que, a excepción de la gramática, el resto de las preguntas de la PI de junio registra una puntuación media superior al de las preguntas de la PI de septiembre. Esto nos lleva a pensar que el grado de dificultad de la pregunta de gramática fue mayor en la PI de junio que en la de septiembre. Asimismo, los resultados indican que la pregunta verdadero/falso es la pregunta que menos discrimina en las dos convocatorias de la PI de Selectividad. Los candidatos la resuelven sin problemas por lo que se considera una pregunta fácil tal y como demuestran los valores del parámetro Alfa (Tabla 4). Ello nos lleva a cuestionar su validez.

Por lo que respecta a las correlaciones (Tabla 3), este estudio revela que las preguntas subjetivas son las que mejor correlacionan con el resto de las preguntas de la PI. La pregunta de redacción/diálogo es la que mayores valores de correlación presenta, especialmente, con las dos preguntas de carácter subjetivo (resumen y pregunta abierta). Este dato es importante porque sugiere que los correctores son bastante consistentes en la aplicación de sus criterios evaluativos.

En cuanto a la fiabilidad interna de la Prueba (Tabla 4), los resultados indican que las preguntas subjetivas (resumen, pregunta abierta y redacción/diálogo) son las que mejor correlacionan con el valor total de la PI. Dichas preguntas son, también, las que mejor contribuyen a la fiabilidad interna de la Prueba ya que obtienen una mayor relevancia en el parámetro Alfa.

Así pues, se puede afirmar que la discriminación de la actuación de los candidatos en la PI de Selectividad reside en las preguntas de carácter subjetivo (resumen, pregunta abierta y redacción/diálogo). Estas últimas son las que mejor discriminan. Dichos resultados justifican el mayor peso (70% del valor total) que se concede a estas preguntas en la puntuación final de la PI de Selectividad de la UIB. De hecho, estos últimos resultados coinciden con los de Herrera (1999) quien cuestiona la validez de las preguntas objetivas que se incluyen en la en la PI de Selectividad de la Universidad Complutense de Madrid (UCM), dado que no discriminan a lo candidatos de forma adecuada: "Consequently, the discrimination of the students' performance merely rests on the sujective items" (Herrera, 1999: 18).

No obstante, cabe señalar que el análisis de varianza ANOVA (Tablas 5 y 7) detecta diferencias significativas en la evaluación de las preguntas subjetivas. Las preguntas objetivas (verdadero/falso, sinónimos y gramática), por el contrario, no presentan diferencias significativas entre los correctores. En resumen, se puede afirmar que si las preguntas subjetivas resultan más válidas que las objetivas son también menos fiables ya que, lógicamente, el nivel de subjetividad de los correctores se reduce de forma considerable en la evaluación de este último tipo de preguntas.

En el análisis de las preguntas objetivas, cabe destacar, de forma excepcional, la diferencia significativa que se detecta entre los correctores en la evaluación de la pregunta verdadero/falso que se incluye en la PI de septiembre (Tabla 7). No obstante, al examinar este caso con mayor detenimiento se comprueba que dicho resultado se debe a la actuación extrema de un único corrector que aplicó criterios de evaluación excesivamente estrictos en la valoración de las preguntas de la Prueba. En cualquier caso, este último dato destaca la importancia que tiene la actuación de un solo corrector en el cómputo total de la nota de Selectividad.

A nivel de grupo, podemos observar que en la PI de junio, el comportamiento excesivamente benévolo de algunos correctores discrepa de forma significativa con la actuación del resto de los correctores. Por el contrario, en la convocatoria de la PI de septiembre el comportamiento excesivamente estricto de algunos correctores disminuye la fiabilidad inter-corrector de forma notable. Esto demuestra que las actuaciones extremas de los correctores, tanto por exceso como por defecto, afectan y reducen seriamente la fiabilidad de las puntuaciones obtenidas en las PI de Selectividad.

6. Conclusiones

Los resultados más relevantes de este estudio se pueden resumir en los siguientes puntos:

a) El diseño de la PI de Selectividad de la UIB se halla bien estructurado y equilibrado ya que permite discriminar a los candidatos de acuerdo con su nivel de dominio de la lengua.

b) Las puntuaciones medias de las distintas preguntas evidencian que los correctores se muestran especialmente estrictos evaluando la gramática.

c) La pregunta verdadero/falso posee un escaso poder discriminatorio lo que nos lleva a cuestionar su validez.

d) La discriminación de la actuación de los candidatos en las PI de Selectividad reside en las preguntas de carácter subjetivo (resumen, pregunta abierta y redacción/diálogo).

e) Las preguntas subjetivas son más válidas que las preguntas objetivas pero son, también, menos fiables.

Respecto a este último punto, cabe decir que, las discrepancias que se evidencian entre los correctores de este estudio en relación al grado de exigencia y consistencia de sus valoraciones en la evaluación de las preguntas subjetivas de la PI de Selectividad señalan la necesidad de adoptar medidas para solventar esta situación. Dentro del contexto de las PAAU, dicha situación resulta especialmente delicada dadas las enormes consecuencias sociales y personales que se derivan de los resultados que se obtienen en Selectividad. Cabe recordar que, en muchas ocasiones, son décimas o milésimas de punto las que impiden a los candidatos acceder a los estudios que desean cursan.

Tal y como señala Weigle (1994), resulta difícil obtener resultados fiables entre los diversos correctores si éstos no han recibido ningún tipo de formación en técnicas de evaluación. Ello se debe al amplio número de factores fortuitos que se asocian con las puntuaciones de cada uno de ellos. La formación de los correctores en técnicas de

evaluación ha demostrado incrementar los niveles de fiabilidad inter-corrector. Otra posible solución dirigida a aumentar la fiabilidad entre los diversos correctores consiste en utilizar la técnica tradicional de la doble corrección.[10] De este modo, la puntuación final de cada ejercicio se obtiene a partir de la media de las puntuaciones que otorgan dos correctores distintos a un mismo ejercicio. En caso de desacuerdo extremo, se utiliza la puntuación de un tercer corrector. La puntuación final se obtiene a partir de la media de las dos calificaciones más próximas.

Dado el incremento en el coste personal y económico de las propuestas arriba citadas, se proponen estudiar otras alternativas quizás más viables en el contexto de las PAAU como, por ejemplo, redefinir los criterios evaluativos de las preguntas y los procedimientos que se llevan a cabo a la hora de evaluar los ejercicios de las Pruebas de Selectividad. Ello permitirá modificar las expectativas de los correctores y les ayudará a concienciarse de la necesidad de obtener niveles de fiabilidad satisfactorios en la evaluación de las Pruebas.

Evidentemente, y dada la gran trascendencia que tienen las pruebas de Selectividad, se requiere un esfuerzo compartido de todos los estamentos involucrados en la innovación educativa para:

a) Proponer cursos de formación para los correctores que participan en la corrección de las PI de Selectividad.
b) Seguir trabajando en el diseño de la PI y estudiar posibles alternativas.
c) Estudiar la viabilidad de incorporar la prueba oral en la PI de Selectividad.
d) Realizar estudios de carácter empírico que nos informen del funcionamiento de las Pruebas y nos permitan mejorar su diseño.

(Versión revisada recibida en diciembre 2005)

BIBLIOGRAFÍA

Alderson, J. C. & J. Banerjee (2001). "Language Testing and assessment (Part 1)". *Language Teaching* 34: 213-36.

Amengual, M. (2003). "A Study of Different Composition Elements that raters Respond to". *Estudios Ingleses de la Universidad Complutense* 11: 53-72.

Amengual, M. (2004). "Realibility Concerns on Evaluating ESL Compositions" en M. Carretero, H. Herrera, G. Kristiansen & J. Lavid (eds.), *Estudios de lingüística aplicada a la comunicación*, 15-27. Madrid: Universidad Complutense · CERSA.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. & A. S. Palmer (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Canale, M. (1983). "On Some Dimensions of Language Proficiency" en J. W. Oller (ed.), *Issues in Language Testing Research*, 333-42. Rowley, MA: Newbury House

Canale, M & M. Swain (1980). "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing". *Applied Linguistics* 1,1: 1-47.

Gamaroff, R. (2000). "Rater Reliability in Language Assessment: the Bug of all Bears". *System* 28,1: 1-53.

Gipps, C. (1994). *Beyond Testing: Towards a Theory of Educational*

Assessment. London: The Palmer Press.

Herrera, H. (1999). "Is the English Test in the Spanish University Entrance Examination as Discriminating as it should be?" *Estudios Ingleses de la Universidad Complutense* 7: 89-107.

ILTA (International Language Testing Association) (2000). Code of Ethics for ILTA. Disponible en www.Dundee.ac.uk/languagestudies/ltest/ilta/ilta.html (documento accedido en febrero 2003).

Lumley, T & T. F. McNamara (1995). "Rater characteristics and rater bias: implications for training". *Language Testing* 12,1: 54-71.

Messick, S. (1989). "Validity" en R. L. Linn (ed.). *Educational Measurement, 3ª ed.*, 13-103. New York: Macmillan.

Milanovic, M., N. Saville & S. Shuhong (1996). "A Study of the Decision-making Behaviour of Composition Markers" en M. Milanovic & N. Saville (eds.), *Performance, Testing, Cognition and Assessment*, 92-114. Cambridge: Cambridge University Press.

Moss, P. (1994). "Can there be Validity without Reliability?" *Educational Researcher* 23: 5-12.

Sanz, I. (1999). "El examen de Selectividad a examen". *Greta: Revista para profesores de inglés* 7,2: 16-29.

Weigle, S. C. (1994). "Effects of training on raters of ESL compositions". *Language Testing* 11,2: 197-223.

La Dra. **Marian Amengual Pizarro** es profesora titular del Departamento de Filología Española, Moderna y Latina (Filología Inglesa) de la Universitat de les Illes Balears (UIB) y actualmente es la coordinadora de lengua inglesa de las Pruebas de Acceso a la Universidad (PAAU) en dicha universidad. Su línea de investigación es la evaluación y ha publicado diversos artículos sobre estos temas.

NOTAS

1 Pueden participar en las PAAU, los estudiantes que hayan cursado de forma satisfactoria cualquiera de las siguientes opciones: bachillerato LOGSE (Ley Orgánica General de Secundaria), Formación Profesional de segundo grado o Formación Profesional de grado superior (módulo nivel 3 y ciclo formativo de grado superior) (Guia d'Accés a la Universitat, 2004).

2 Las pruebas criteriales miden el rendimiento del alumno a partir de unos criterios estándar de calidad que permiten definir el grado de competencia que manifiesta cada candidato de forma individual sin hacer referencia a la actuación del resto de los candidatos.

3 Es importante añadir que únicamente las pruebas de dominio lingüístico de carácter normativo permiten estimar la fiabilidad a través de los métodos de la teoría clásica de la evaluación. Ello explica, en parte, su uso mayoritario en pruebas de carácter nacional e internacional (por ejemplo, *Test of Written English* [TWE], etc.)

4 Bachman (1990) incluye tres componentes que interactúan en su modelo *Communicative Language Ability* (CLA): competencia lingüística, competencia estratégica y mecanismos psicofisiológicos.

5 Bachman y Palmer (1996: 21) definen 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".

6 Para que el candidato se considere apto para una vía de acceso se debe obtener una calificación global de al menos cuatro puntos por dicha vía. La calificación definitiva para el acceso a los estudios universitarios se calcula ponderando un 40% la calificación global de las PAAU y un 60% la nota media del expediente académico del candidato en el bachillerato.

7 El concepto de asimetría explica el desplazamiento de las puntuaciones sobre la escala de valores hacia la izquierda o la derecha con respecto al punto medio.

8 La correlación de Pearson es un procedimiento estadístico paramétrico que nos permite medir el grado de relación que existe entre dos variables.

9 En este análisis las puntuaciones de la *redacción* y del *diálogo* se han tomado conjuntamente dado el reducido tamaño de la muestra de esta última opción.

[10] Los candidatos que no consideren justa o adecuada la puntuación final de los ejercicios de las PAAU pueden optar por la vía denominada *doble corrección*. Esta vía contacta un nuevo corrector que 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.

Apéndice 1

Proves d'accés a la Universitat (2002)

Selectivitat (LOGSE)

Anglès

Model 1

Read the passage carefully and answer the questions in English. USE YOUR OWN WORDS AS FAR AS POSSIBLE.

Time allowed: 1 hour and 30 minutes. Total score: 10 points.

The Language

When I arrived in England I thought I knew English. After I'd been here an hour I realized that I did not understand one word. In the first week I picked up a tolerable working knowledge of the language and the next seven years convinced me gradually but thoroughly that I would never know it really well, let alone perfectly. This is sad. My only consolation being that nobody speaks English perfectly.

If you live here long enough you will find out to your greatest amazement that the adjective *nice* is not the only adjective the language possesses, in spite of the fact that in the first three years you do not need to learn or use any other adjectives. You can say that the weather is nice, a restaurant is nice, Mr Soandso is nice, Mrs Soandso's clothes are nice, you had a nice time, and all this will be very nice.

Then you have to decide on your accent. The easiest way to give the impression of having a good accent or no foreign accent at all is to hold an unlit pipe in your mouth, to mutter between your teeth and finish all your sentences with the question: 'isn't it?' People will not understand much, but they are accustomed to that and they will get a most excellent impression.

Anyway, this whole language business is not at all easy. After spending eight years in this country, the other day I was told by a very kind lady: 'But why do you complain? You really speak a most excellent accent without the slightest English.'

George Mikes, *How to Be an Alien* (1946) (edited)

1. Write a summary of the preceding passage in no more than 50 words. DO NOT copy from the text. **(2 points)**

2. Say whether the following statements are TRUE or FALSE. Explain WHY using your own words OR finding evidence in the text. NO marks are given for only TRUE or FALSE. **(1 point)**

a) It took the author seven years to know English well.

b) The adjective *nice* is surprisingly not the only adjective the language possesses.

c) English people tend to get an excellent impression of accents they do not usually understand.

d) The author ends up speaking English with a perfect accent.

3. In your own words and based on the ideas from the text, answer the following question. **(1 point)** Why does the author consider the adjective *nice* to be one of the most useful and productive in English? Explain.

4. Follow the instructions for each question.

a) Find in the text words or phrases which mean the same as the following:

1) reasonable 2) regardless of 3) to murmur 4) anyhow **(1 point)**

b) Finish each sentence so that it means the same as the sentence before it. **(1 point)**

b1) 'Please don't repeat every thing I say.'

Would you mind …

b2) 'If I were you, I'd try hard to acquire an Oxford accent.'

He advised me …

5. Choose ONLY ONE of the following options. **(4 points)**

a) Write a composition of 120-150 words on: If an English person wanted to come to your country to learn your language, what advice would you give him or her? Explain.

b) Write a dialogue of 120-150 words on: A conversation between you and a friend of yours who is going to spend six months in England and cannot speak a word of English.

Apéndice 2

Proves d'accés a la Universitat (2002)

Selectivitat (LOGSE)

Anglès

Model 1

Read the passage carefully and answer the questions in English. USE YOUR OWN WORDS AS FAR AS POSSIBLE.

Time allowed: 1 hour and 30 minutes. Total score: 10 points.

The Rights of Women

In 1918, some women over the age of thirty gained the right to vote after a long, hard struggle. John Stuart Mill, a radical thinker, had tried unsuccessfully to include votes for women in the 1867 Reform Bill. The industrial revolution had increased the power of men, and their feelings about property.

A man thought of his wife and daughters as his property, and so did the law. It was almost impossible for women to get a divorce, even for those rich enough to pay the legal costs. Until 1882, a woman had to give up all her property to her husband when she married him. And until 1891, husbands were still allowed by law to beat their wives with a stick "no thicker than a man's thumb", and to lock them up in a room if they wished. By 1850, wife beating had become a serious social problem in Britain. Women were probably treated worse in Britain than in any other industrialising European country at this time.

The war in 1914 changed everything. Britain would have been unable to continue the war without the women who took men's places in the factories. By 1918 29 per cent of the total workforce of Britain was female. Women had to be given the vote. But it was not until ten years later than the voting age of women came down to twenty-one, equal with men.

Once women could vote, many people felt that they had gained full and equal rights. But there was still a long battle ahead for equal treatment and respect both at work and at home.

D. McDowall, *An Illustrated History of Britain* (2003) (edited)

1. Write a summary of the preceding passage in no more than 50 words. DO NOT copy from the text. **(2 points)**

2. Say whether the following statements are TRUE or FALSE. Explain WHY using your own words OR finding evidence in the text. NO marks are given for only TRUE or FALSE. **(1 point)**

a) Women were granted full and equal rights during the industrial revolution.
b) Until 1891, wife beating was considered to be legal.
c) In 1914 women over the age of thirty gained the right to vote.
d) Once women could vote, they gained full and equal rights both at work and at home.

3. In your own words and based on the ideas from the text, answer the following question. **(1 point)** Why did the war in 1914 change the situation of women? Explain.

4. Follow the instructions for each question.

a) Find in the text words or phrases which mean the same as the following: **(1 point)**

1) belongings, possessions 2) to hand over
3) a long thin piece of wood 4) people in a country who are available for work

b) Answer the questions below. **(1 point)**

b1) Turn the following sentence into the passive voice:
Many celebrities attended the lecture on women's rights.

b2) Finish this sentence so that it means the same as the sentence before it:
'Why did many politicians refuse to support the 'suffragettes'?'
He asked…

5. Choose ONLY ONE of the following options. (4 points)

a) Write a composition of 120-150 words on: To what extent have things changed? Do you think women have nowadays gained full and equal rights? Explain.

b) Write a dialogue of 120-150 words on: A conversation between a man and a woman discussing the importance of equal treatment and respect.