CONTROL DE CALIDAD EN LA CALIFICACIÓN DE LA PRUEBA DE INGLÉS DE SELECTIVIDAD

FRANCES WATTS y AMPARO GARCÍA CARBONELL [*]

El objetivo de este estudio es examinar uno de los aspectos menos controlables de la Selectividad, la precisión de los criterios de corrección y la mediación de los jueces entre los criterios y la muestra a corregir. Expone dos experimentos realizados que basan su desarrollo en el examen de inglés de la convocatoria de junio de 1995 de las Pruebas de Aptitud para el Acceso a la Universidad en la provincia de Valencia. El estudio examina el problema de la concordancia entre jueces y propone la utilización de unos criterios de corrección detallados y uniformes, que bautiza como método global-definido. Asimismo propone recomendaciones de procedimiento para las pruebas de idioma extranjero de la Selectividad con el fin de aumentar la fiabilidad.

The objective of this study is to examine one of the less controllable aspects of the Selectividad, the precision of the rating criteria and the mediation of the raters between the criteria and the sample to be corrected. This paper discusses two experiments carried out using the English language test from the June, 1995 sitting of the Spanish university access examination battery administered in the province of Valencia. The study examines the problem of rater agreement and proposes the use of detailed, uniform scoring criteria of the type termed focused holistic. It also makes recommendations for scoring procedure for the foreign language tests of the Selectividad with the purpose of increasing reliability.

Introducción

En España, el acceso a la universidad y la elección de estudios superiores dependen de la culminación con éxito de la enseñanza secundaria y de la calificación obtenida en las Pruebas de Aptitud para el Acceso a la Universidad, la comúnmente llamada Selectividad. La Selectividad consta de una serie de pruebas, de las cuales la del idioma extranjero es obligatoria, siendo el inglés hoy día el idioma más elegido, seguido a una distancia considerable del francés, y aún más del alemán. La prueba de Inglés como

[*] FRANCES WATTS y AMPARO GARCÍA CARBONELL son Profesoras Titulares de Escuela Universitaria. Departamento de Idiomas. Universidad Politécnica de Valencia.

Lengua Extranjera juega un papel ponderativo importante en la batería española de exámenes que selecciona a los estudiantes aptos para los estudios universitarios.

Las pruebas de Selectividad casi siempre se definen como una prueba abierta. El alto grado de subjetividad e imprecisión que eso comporta centra las críticas más frecuentes. La fiabilidad de las calificaciones en la evaluación de lenguas, como en muchas otras materias, viene condicionada por distintas fuentes de variación. El resultado de un examen debería reflejar lo que sabe el examinando; sin embargo, las mediciones lingüísticas, según Bachman (1990), son una interacción compleja entre los aspectos controlables y menos controlables del contexto evaluador, y las características del examinando. Los aspectos controlables son el marco y los atributos de la prueba, tales como el tipo de examen, la respuesta esperada, y los parámetros y criterios de evaluación. Aspectos menos controlables son la suficiencia de la muestra de la lengua, la precisión de la escala con la que hay que juzgarla y la mediación de los jueces entre la escala y la muestra. Las características del examinando son su experiencia en el idioma, su cultura y madurez, entre otros.

Diferentes estudios llevados a cabo han demostrado que las calificaciones de jueces competentes e independientes pueden variar (Cuxart y otros, 1997; Wood, 1993; Huot, 1990; Marín Ibañez, 1980). Lo que es importante para un juez lo es menos para otro. Por ejemplo, al evaluar un ejercicio de redacción, un juez puede hacer hincapié en la corrección gramatical, mientras que otro puede considerar más el contenido o la organización del mismo. Incluso si todos los jueces considerasen el mismo componente, podrían interpretarlo de manera distinta.

Muchos errores en la calificación son atribuibles a la variable juez. Según Wolf (1990) y R. Neira y otros (1995) se han identificado tendencias en los errores tales como la gravitación hacia el punto medio de una escala, evitando los extremos; a otorgar puntuaciones similares a características que el evaluador considera similares, pero que no lo son; y al efecto "halo" o tendencia a permitir que la impresión general del ejercicio influya en la evaluación de lo específico. Un juez puede dejarse influir por sus prioridades en la evaluación en perjuicio de los criterios prefijados. Asimismo, la idiosincrasia y las creencias ideológicas pueden afectar a la neutralidad de la calificación.

Otros factores intrajuez son la fatiga, la prisa y el estado de ánimo, que también producen desviaciones en las calificaciones. Incluso la sensibilidad a los errores puede menguar durante el período evaluador. Un error que se estima importante en el primer examen, puede parecerlo menos

después de encontrarlo repetidas veces en exámenes siguientes. La inconstancia del juez es muy difícil de identificar y de compensar.

Respecto a los aspectos del entorno que influyen en la calificación, podemos mencionar la hora del día, las distracciones en la sala donde se califica y el orden de presentación de los ejercicios. Los aspectos influyentes que surgen de los propios ejercicios son la organización, el desarrollo, la calidad argumental, la longitud, la ortografía y la caligrafía. La velocidad con la que los jueces se ven obligados a evaluar y su actitud personal ante la presión a ceñirse a unos criterios también influyen. De exámenes de los que derivan consecuencias significativas para los examinandos, igualmente puede provenir una presión adicional sobre los jueces.

Estos problemas de tipo operativo, cuya magnitud e importancia pueden soslayarse en modo alguno, se complican cuando se unen a las preocupaciones teóricas por el "constructo" o concepto objeto de evaluación. En la evaluación de los idiomas se evalúa la competencia en la lengua junto con la habilidad en la comunicación escrita y la adecuación del contenido.

En las tres últimas décadas la discusión sobre lo que significa competencia en una lengua ha sido especialmente rica, desde que Hymes formulara en 1972 su modelo de competencia lingüística, favoreciendo la extensión del movimiento comunicativo en la enseñanza de los idiomas. Hablar de competencia, en opinión de Milanovic y Saville (1996), ha ayudado a centrar la atención de la evaluación en la actuación del examinando. En un examen basado en la actuación, al examinando se le requiere producir una muestra escrita o hablada de la lengua. Aunque la evaluación tradicional ha utilizado este método, el enfoque comunicativo ha introducido consideraciones de autenticidad y de contexto en el que se produce la actuación, con el fin de reproducir las situaciones en las que la comunicación tiene lugar en la vida real.

Bachman y Palmer (1996) recogen los pasos de la discusión en su modelo de habilidad comunicativa en la lengua (*Communicative Language Ability*), en el que se definen una serie de competencias que, todas juntas, componen la citada habilidad. Es decir, se puede desglosar la habilidad comunicativa en la lengua en dos: mecanismos "psicofisiológicos" y estrategias metacognitivas. Estas incluyen una competencia estratégica y una competencia lingüística. La competencia lingüística a su vez se desglosa en una competencia para la organización (gramatical y textual) y otra pragmática (funcional y sociolingüística). Los conocimientos temáticos

y el contexto de la situación completan, en una descripción muy escueta, lo que para estos autores es la habilidad comunicativa en la lengua y lo que debe ser el objeto de evaluación en una lengua.

El significado de la habilidad en la comunicación escrita fue examinado con minuciosidad por los autores del estudio internacional de redacción (*International Study of Written Composition*) llevado a cabo en los años 80 con escolares de 14 países y 11 lenguas diferentes. El aspecto de la habilidad o competencia en la escritura, su definición, fue considerado el factor decisivo en la elección de las tareas, así como del método y criterios de evaluación. Según Purves et al. (1988), esta competencia consiste en dos características discretas. Por un lado, la competencia para producir o codificar un texto, que se refiere tanto a la competencia motora para formar letras como a la lingüística para producir un texto con la gramática, ortografía y puntuación adecuadas. Y por otro, tenemos la competencia para la estructuración del discurso, que aglutina dos competencias, una cognitiva y otra social. La cognitiva es la capacidad de manejo y de organización de un tema; la competencia social se asocia a las funciones del discurso y a su contextualización.

Observamos que las definiciones del constructo en el caso del examen de Inglés de Selectividad se solapan. Es decir, las definiciones de la habilidad comunicativa en una lengua y de la habilidad para la escritura o la redacción en una lengua, junto con la adecuación del contenido de la respuesta, son factores a tener en cuenta en la elección del método y criterios de corrección.

La influencia del juez en las calificaciones de Selectividad

La investigación sobre la Selectividad, desde su implantación en el curso 1974-75, se ha interesado por diversos factores que inciden en la fiabilidad, investigación que ha propulsado sucesivos cambios y mejoras en las pruebas (Escudero, 1997). Si nos atenemos a la taxonomía de Bachman descrita en el apartado anterior, podríamos decir que la mayor parte de los trabajos han estudiado los factores más controlables del contexto examinador y las características de los examinandos. Por ejemplo, en la presente década se ha realizado un estudio de comparación entre las calificaciones de la enseñanza secundaria y de la Selectividad, así como una exploración de las diferencias entre grupos de alumnos utilizando variables tales como el distrito escolar, el tribunal examinador, tipo de escuela secundaria, opción, sexo y media de B.U.P./C.O.U. (Muñoz-Repiso, 1991).

Sans Martí (1991) examinó el efecto del tribunal examinador y tanda, hallando que las diferencias, especialmente en las asignaturas de humanidades, constituían una fuente de error importante. Bueno et al. (1990) también examinaron el papel del tribunal examinador, pudiendo señalar a unos tribunales como "duros" y a otros como "blandos", aunque todos mostraron una coherencia suficiente en la calificación.

Por lo que respecta al acuerdo entre correctores, aspecto cuyo estudio requiere una doble corrección, Nieda y otros (1984) realizaron un experimento utilizando criterios especialmente diseñados. Al aplicar criterios de puntuación por acierto (ver definición en Tipos de criterios) en un examen de Biología, el grupo de investigadores comparó sus puntuaciones con las dadas por los correctores oficiales. Los resultados fueron aceptables en tres de los cuatro tribunales, con correlaciones entre 0,784 y 0,996, mientras que la comparación con el cuarto tribunal dio correlaciones entre 0,199 y 0,48. Las razones de una actuación tan anómala, según los investigadores, fueron la no utilización de los criterios oficiales previamente acordados, que los correctores oficiales no eran especialistas en la asignatura y que la corrección en algunos casos había sido descuidada.

Un experimento a gran escala para estudiar la fiabilidad de las calificaciones en Selectividad fue llevado a cabo en 1992 con el establecimiento de un tribunal paralelo en Teruel (Escudero y Bueno, 1994). El citado estudio halló a través de la calificación doble que las distintas pruebas juegan un papel compensatorio en la estabilidad de los resultados globales, aunque las discrepancias observadas podrían haber afectado la calificación final de ciertos individuos, quienes habrían aprobado con un conjunto de jueces y no con otro. Aunque las discrepancias no eran estadísticamente significativas y se consideraron casi inevitables en pruebas abiertas como la Selectividad, las conclusiones señalaron la necesidad de diseñar criterios de corrección más precisos y de mejorar los sistemas de coordinación entre los tribunales examinadores y sectores especializados en evaluación.

Cuxart y otros (1997) ofrecen sus exploraciones sobre un sistema para controlar la calidad de la calificación en Selectividad. Exámenes de Matemáticas y de Filosofía de la convocatoria de junio de 1995 en Cataluña fueron objeto de una doble corrección. Los resultados demostraron el valor de la calificación por duplicado, a la vez que se puso de manifiesto claras diferencias en la concordancia entre los jueces de las dos materias. Hubo un mayor acuerdo entre los evaluadores de los ejercicios de Matemáticas que entre los de Filosofía. Sus conclusiones resaltan la baja calidad de la

corrección en las dos materias, lo cual podría afectar las notas cerca del límite del aprobado.

Cuxart (1998), en su estudio sobre la calidad de la corrección, añade que las razones para esa baja calidad podrían deberse a los exámenes en sí (formato abierto, distintos niveles de dificultad y de discriminación de las preguntas, etc.) y al sistema de corrección. Su impresión es que los correctores no son "suficientemente expertos", por lo que aboga por una mejor comunicación entre coordinadores y correctores, la preparación de pautas de corrección claras y la formación de los correctores.

Cuxart y Longford (1998) estudian la posibilidad de ajustar las calificaciones para reducir el impacto de la severidad e inconstancia de los jueces, así como de estandarizar las puntuaciones en la Selectividad. Su artículo ofrece bastantes sugerencias para mejorar las pruebas de acceso a la universidad, cuya puesta en práctica requeriría una investigación previa y una adaptación cuidadosa a circunstancias específicas. Los autores destacan la naturaleza dinámica de la Selectividad, por lo que son cautelosos a la hora de generalizar sus conclusiones estadísticas.

Quedan por investigar, como se ha visto, las medidas a tomar en una materia específica. El presente estudio tiene como objetivo examinar cómo mejorar la fiabilidad en la calificación de la prueba de Inglés como Lengua Extranjera. El hecho de que la prueba de idioma sea obligatoria y que el Inglés sea el idioma más estudiado pone en evidencia su papel ponderativo en el conjunto de las pruebas.

Tipos de criterios

Las escalas o criterios de corrección se hacen para lograr valoraciones sistemáticas pero a menudo fallan en su propósito. En apariencia, pueden tener intervalos equivalentes y en la práctica, asemejarse a una escalera con peldaños de distinta altura. Al interpretar la escala, los jueces pueden diferir; por ejemplo, en una escala de 1 a 10, lo que es 9 para un juez, puede ser 5 para otro. En los casos dudosos, algunos jueces serán sistemáticamente más severos que otros a la hora de elegir entre una puntuación u otra.

La elección del método y de los criterios de corrección debería hacerse teniendo en cuenta el propósito y contexto de la situación evaluadora. Si las pruebas se utilizan para medir el progreso del alumno, para determinar la salida de un curso o admisión a una institución, para hacer diagnosis o tomar decisiones sobre la ubicación de alumnos, para

determinar la competencia o llevar a cabo una investigación, todos ellos son objetivos que deben ser definidos junto con todos los elementos que componen el contexto evaluador.

Son muchos los autores que han escrito acerca de la evaluación tanto de lenguas como en general, pero no hay acuerdo entre ellos sobre la definición de los distintos tipos de criterios. La mayoría reconocen dos tipos o métodos de corrección, el global u "holístico" y el analítico. Nuestro estudio plantea cinco tipos de criterio: el global, el analítico, el global-definido, el descuento por error y la puntuación por acierto.

En la puntuación por acierto, el evaluador empieza desde cero y otorga puntos por la presencia de rasgos o contenidos preestablecidos. En el descuento por error, el evaluador resta puntos de un total por los errores hasta llegar a un mínimo de cero. En pruebas de idioma, la puntuación por acierto se puede utilizar, por ejemplo, en la evaluación de la destreza de la expresión oral. El descuento por error, que se utiliza cuando es primordial la precisión, es apropiado en la evaluación de traducciones.

El tipo global u holístico suele implicar un juicio rápido del conjunto de la prueba por parte del evaluador. Está especialmente indicado cuando los correctores tienen mucha experiencia y cuentan con una procedencia común en su formación, cuando existe cierta prisa en la ejecución de la calificación y/o cuando sólo es necesario ordenar las pruebas según el nivel de calidad. El tipo analítico requiere la puntuación de una tarea por distintos aspectos, puntuaciones que se suman para llegar a una calificación final. Este tipo obliga a los evaluadores a considerar aspectos que de otro modo serían ignorados, admite evaluadores de procedencia heterogénea y de menos experiencia y tiende a producir una calificación más fiable por contar con varias puntuaciones. El inconveniente principal es que presupone que los evaluadores podrían discriminar eficazmente entre los distintos aspectos y que el proceso de evaluación es más lento que el global.

Por último, el tipo de criterio que hemos convenido en llamar global-definido, como adaptación del término inglés *focused holistic*, es un híbrido de los tipos global y analítico. Los criterios presentan los perfiles de los varios niveles de corrección atendiendo a ciertos rasgos definidos. Las muestras que son objeto de evaluación se juzgan por su adecuación a dichos perfiles. Este tipo de criterios combina las ventajas de la rapidez del método global y la obligación de considerar diversos aspectos del método analítico, que en principio favorece a los jueces menos experimentados. El uso de este tipo de criterio parece apropiado en los exámenes a gran escala, como lo demuestra el hecho de que se utilicen en la evaluación de la producción

escrita en los renombrados exámenes de Inglés como Lengua Extranjera de la University of Cambridge Local Examinations Syndicate (el *Proficiency* y el *First Certificate*); del Educational Testing Service (el Test of Written English del TOEFL); y del Instituto de Lengua Inglesa de la Universidad de Michigan (también llamado *Proficiency*).

Materiales y Método

Los criterios globales que se emplean actualmente en la calificación de la prueba de inglés de la Selectividad en la provincia de Valencia (ver apéndice) explican que hay que evaluar esencialmente la capacidad de comprensión del alumno y valorar positivamente a los alumnos que demuestren variedad de vocabulario y corrección gramatical, así como una complejidad sintáctica y riqueza de estilo por encima de la media. A pesar de que se especifica el valor numérico de cada pregunta, no se facilitan descriptores de cada característica ni cómo evaluar positivamente los rasgos de vocabulario y gramática ni cuántos puntos hay que otorgar a las respuestas con diverso grado de corrección.

El presente estudio propone acotar los actuales criterios globales con la adición de descriptores de los diversos niveles de corrección y el consiguiente desglose de la puntuación que se otorga a cada respuesta, aplicando criterios global-definidos. El siguiente esquema muestra el desglose según los tipos de pregunta que componen la actual prueba de inglés. Se observa que para las preguntas de vocabulario y de gramática no se admiten niveles de corrección, es decir, la respuesta es correcta o incorrecta. En cambio, en las preguntas abiertas de respuesta corta que corresponden a la comprensión lectora, opinión y tema -representando el 70% de la puntuación total- sí se tiene en cuenta el grado de corrección.

Tipo de pregunta	*Nº de pregunta*	*Valor Parcial*	*Valor Total*
Comprensión lectora	1 y 2	1 0'5 0	2 puntos
Tema y Opinión	3 y 12	2'5 2 1'5 1 0'5 0	5 puntos
Vocabulario	4-7	0'25	1 punto
Gramática	8-11	0'5	2 puntos

Tabla 1 - Desglose de la puntuación

Los criterios propuestos distinguen cuatro características que diferentes estudios de investigación han identificado como inherentes a la buena escritura: el contenido; la organización y la conexión; la gramática y

el vocabulario y la ortografía y la puntuación. Los criterios de corrección propuestos son los siguientes:

CRITERIOS DE CORRECCIÓN
PARA LA PRUEBA DE INGLÉS DE SELECTIVIDAD

Sección A, nº 1 y nº 2 1 pto. Comprensión 1 punto cada pregunta	Respuesta correcta, sin errores gramaticales mayores. *	
	0´5	Respuesta correcta, entendible pero el dominio morfológico y/o sintáctico no es constante.
	0	Respuesta incorrecta o inexistente, o con una falta total de claridad o de dominio morfológico y/o sintáctico, o indescifrable
Sección A, nº 3 Opinión 2’5 puntos en total	2’5	Opinión plenamente desarrollada. Organización y conexión apropiadas. Utilización de una amplia gama de estructuras sintácticas Control morfológico correcto. Vocabulario apropiado. Ortografía y puntuación sin errores. Longitud adecuada.
	2	Opinión bien desarrollada. Estructura organizativa lógica. Conexión con pocos problemas. Sintaxis simple y compleja. Morfología casi siempre correcta. Vocabulario adecuado. Errores de ortografía y puntuación no distraen.
	1’5	Opinión desarrollada, aunque incompleta, falta de claridad o sin enfoque. Organización parcialmente adecuada. Conexión no siempre conseguida o ausente. Sintaxis simple y compleja pero con errores, o sintaxis sin errores pero restringida y simple. Dominio morfológico inconstante. Vocabulario a veces inapropiado. Errores de ortografía y puntuación que distraen a veces. Longitud excesiva o insuficiente.
	1	Opinión de desarrollo limitado, incompleto o confuso. Organización poco controlada. Conexión insuficiente o nula. Estructuras sintácticas simples pero con muchos errores, o sintaxis compleja que no refleja dominio. Dominio morfológico extremadamente limitado. Vocabulario poco variado y simple, de significado aproximado, a menudo inapropiado. Errores de ortografía y puntuación que a menudo distraen.
	0’5	Opinión sin desarrollar. Organización inexistente Conexión inexistente. Dominio morfosintáctico extremadamente limitado.

		Vocabulario poco variado y repetitivo. Errores de ortografía y puntuación que causan interferencias serias.
	0	Respuesta no es una opinión, es totalmente incoherente, indescifrable, o inexistente.
Sección B Vocabulario, nº 4 - nº 7, 0'25 punto cada una	0'25	Respuestas con la misma morfología. p.e. broken down, fallen to pieces = crumbled; (se acepta have crumbled pero no to crumble, crumbling, etc.).
Sección C Gramática nº 8 - nº 11, 0'5 punto cada una	0'5	Respuestas con el mismo significado y gramaticalmente correctas. Se aceptan todas las alternativas.
Sección D Expresión 2'5 puntos en total	2'5	Tema del texto desarrollado con riqueza y eficacia. Organización y conexión apropiadas. Utilización de una amplia gama de estructuras sintácticas Control morfológico correcto. Vocabulario apropiado. Ortografía y puntuación sin errores. Longitud adecuada.
	2	Tema bien desarrollado. Estructura organizativa lógica. Conexión con pocos problemas. Sintaxis simple y compleja. Morfología casi siempre correcta. Vocabulario adecuado. Errores de ortografía y puntuación no distraen.
	1'5	Tema desarrollado, aunque incompleto, falto de claridad o sin enfoque. Organización parcialmente adecuada. Conexión no siempre conseguida o ausente. Sintaxis simple y compleja pero con errores, o sintaxis sin errores pero restringida y simple. Dominio morfológico inconstante. Vocabulario a veces inapropiado. Errores de ortografía y puntuación que a veces distraen. Longitud excesiva o insuficiente.
	1	Tema de desarrollo limitado, incompleto o confuso. Organización poco controlada. Conexión insuficiente o nula. Estructuras sintácticas simples pero con muchos errores o sintaxis compleja que no refleja dominio. Dominio morfológico extremadamente limitado. Vocabulario poco variado y simple, de significado aproximado, a menudo inapropiado. Errores de ortografía y puntuación que a menudo distraen.
	0'5	Tema sin desarrollar. Organización inexistente. Conexión inexistente. Dominio morfosintáctico extremadamente limitado.

	Vocabulario poco variado y repetitivo. Errores de ortografía y puntuación que causan interferencias serias.
0	Respuesta sin relación con el tema, es totalmente incoherente o indescifrable, o inexistente.

* Ejemplos de errores gramaticales mayores:

It appears the question very clear. (Sujeto doble)
Is better to work for others. (Falta sujeto)
This people..., another things... (Falta de concordancia en el número – en algunos casos este error es menor.)
Importants papers (Adjetivos en plural)
The boy speaked to the girl (Forma o tiempo verbal incorrecto)

Errores gramaticales menores:

He come every day (No ha puesto -s en 3ª persona singular.)
The abortion is illegal (Artículo definido ante sustantivo incontable utilizado en general; si la falta de dominio en el uso de los artículos es total, este error es mayor.)
Better that (confunde el comparativo *than*).

~~La frecuencia de errores menores los puede convertir en errores~~ mayores.

El experimento para comprobar la consistencia de los nuevos criterios se llevó a cabo con dos grupos de cuatro jueces, un grupo empleó los nuevos criterios y el otro, los tradicionales. Los ocho jueces, sin experiencia previa en la calificación de pruebas de Selectividad, corrigieron las mismas 100 pruebas. Se escogieron aleatoriamente exámenes ya calificados en la convocatoria de Selectividad de junio de 1995, y posteriormente se pasaron a limpio para erradicar todo signo de la corrección anterior. Se contó con la ayuda de alumnos de la Universidad Politécnica de Valencia, quienes hicieron las 100 copias a mano en el mismo papel-formato que se emplea normalmente en la Selectividad. A continuación se hicieron fotocopias de las 100 pruebas para cada uno de los ocho jueces. Después de recoger las puntuaciones dadas por cada juez a cada respuesta, los datos fueron sometidos a análisis estadístico. Se hallaron las notas medias y los rangos, se hicieron diferentes análisis de variancia de medidas repetidas, se obtuvieron coeficientes de fiabilidad, con los valores de t asociados, y por último, se hallaron las z de comparación entre los distintos criterios.

Resultados

Los resultados que presenta la figura 1 muestran que los nuevos criterios son más consistentes que los tradicionales; demuestran una mayor concordancia entre jueces, y una mejor diferenciación en las calificaciones. Según los valores z, con los nuevos criterios hubo una significativa mayor fiabilidad en las calificaciones totales; al igual que la hubo en seis de las doce preguntas de la ejercicio. Asimismo, en tres de las otras seis preguntas hubo una tendencia muy marcada a favor de la calificación con los nuevos criterios. Sólo dos preguntas presentaron una tendencia favorable leve a los criterios tradicionales y una única pregunta mostró significativamente mayor fiabilidad en los criterios tradicionales.

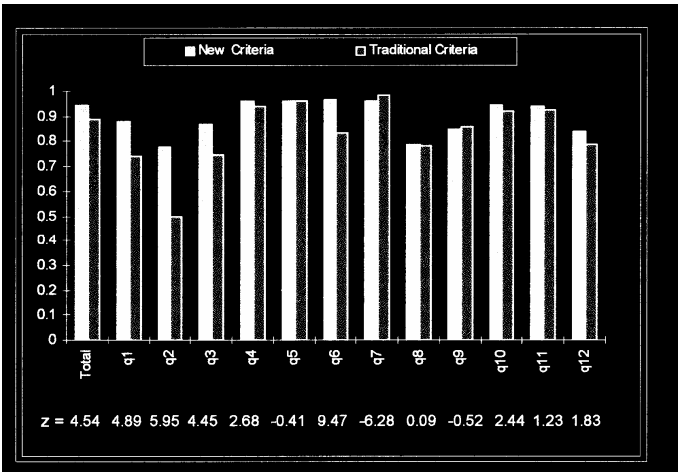


Figura 1 - Resultados obtenidos con criterios tradicionales y nuevos

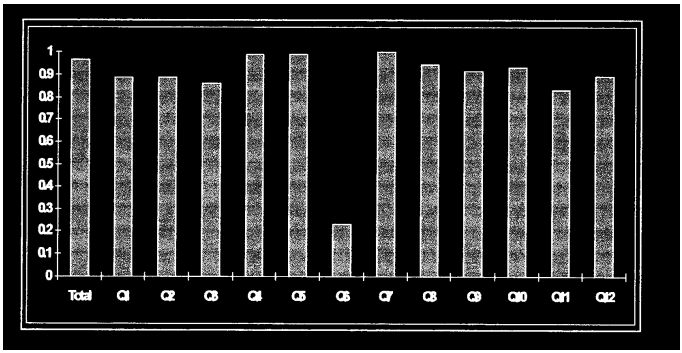
Los análisis de variancia de medidas repetidas arrojaron además información valiosísima sobre las fuentes de variación en las calificaciones analizadas. En casi todas las preguntas se observó una mayor variancia en las calificaciones obtenidas con los nuevos criterios y una menor variancia entre los jueces que los utilizaron. Esta observación indica que los nuevos criterios actúan sobre los jueces restringiendo su variabilidad y sobre las calificaciones diversificándolas. Es decir, con los nuevos criterios los jueces consiguen una mayor finura o precisión en la medición, que se complementa con una mayor diversidad de los resultados, teniendo como efecto una mejor situación de los sujetos. Todo ello tiene una enorme transcendencia en un contexto como el de la Selectividad, donde las décimas en la calificación tienen consecuencias decisivas.

Para comprobar si los nuevos criterios mantienen su consistencia en el tiempo, se hizo una prueba con dos jueces con experiencia en la calificación de pruebas de Selectividad. Los dos emplearon los nuevos criterios para corregir las mismas 20 pruebas, utilizando fotocopias limpias conseguidas del modo descrito en el experimento anterior. La corrección se repitió del mismo modo cuatro meses más tarde.

Las calificaciones que los jueces experimentados obtuvieron con los nuevos criterios mostraron una alta consistencia. Hubo, sin embargo, una pregunta en la que discreparon, debido a una aplicación diferente de los criterios nuevos. En la primera ocasión los jueces se ajustaron estrictamente a los nuevos criterios, mientras que en la segunda hubo un retorno a los criterios tradicionales en dicha pregunta. Esto se interpreta como indicador de la necesidad permanente de que los jueces estén coordinados. El resultado del ejercicio demuestra la consistencia de los nuevos criterios en el tiempo, tal y como se observa en la figura 2.

Figura 2 - Consistencia de los nuevos criterios en el tiempo

En conclusión, este estudio demuestra que con el método global-definido se consigue una muy significativa mayor concordancia entre jueces, y también una mayor precisión en las calificaciones. Se verifica que la aplicación de los criterios de modo diferente provoca un descenso agudo en la fiabilidad, por lo que también se constata el valor de la coordinación entre los jueces.



Conclusión

Sería ingenuo intentar compensar todas las deficiencias atribuibles a los jueces. Pero creemos poder aspirar a erradicar los desequilibrios más notables potenciando una mayor atención al método y al procedimiento de evaluación de las pruebas de Selectividad. Nuestra propuesta está concebida teniendo como referencia la prueba de Inglés, pero quizás sea de interés en

la evaluación de otras materias. Contiene tanto sugerencias para mejorar las pautas de corrección como para reforzar la formación de los jueces.

DISEÑO DE LOS CRITERIOS DE CORRECCIÓN. Se recomienda la adopción de criterios con descriptores de los diversos niveles o grados de corrección de respuesta. Este tipo de criterio combina las virtudes de dos clases de criterios: la rapidez de los criterios globales o de impresión que utilizan los jueces muy experimentados y la obligación a considerar aspectos específicos que conllevan los criterios analíticos, que tanto beneficia a los jueces con menos experiencia y entrenamiento. Dichos criterios serían ajustados a la prueba de Selectividad y consensuados entre el coordinador especialista, correctores con experiencia en la calificación de Selectividad, y expertos en evaluación; los criterios se publicarían antes de la convocatoria.

FIJACIÓN DE LA NORMA. Se recomienda la selección de ejercicios representativos de los niveles de corrección descritos en los criterios para que sirvan de modelo en la coordinación entre jueces.

COORDINACIÓN DE JUECES. Se recomienda llevar a cabo sesiones en las que los jueces tengan la oportunidad de calificar los ejercicios modelo y puedan resolver dudas y discrepancias. Estas sesiones supondrían un foro de discusión, con el objetivo final de aplicar los criterios de la forma más uniforme posible.

REUNIÓN DE COORDINACIÓN. La reunión preceptiva de coordinación que tiene lugar al terminar las pruebas de Selectividad seguiría con el propósito de comentar las incidencias de la prueba, a la vez que serviría para resolver posibles dudas puntuales sobre la aplicación de los criterios.

COMPROBACIÓN DE FIABILIDAD. Se recomiendan dos maneras de comprobar la fiabilidad de las calificaciones. Una, que todos los jueces califiquen un determinado número de ejercicios de la misma convocatoria, seleccionados y fotocopiados antes de comenzar la calificación general. Posteriormente, los ejercicios se analizarían estadísticamente para comprobar la fiabilidad. El otro modo sería la calificación aleatoria doble por parte de especialistas, seguido del correspondiente análisis estadístico de fiabilidad. Esta comprobación permitiría un seguimiento de la calidad de la calificación.

Esta propuesta está en la misma línea que la recomendación nº 8 del Informe de la Ponencia de estudio sobre la Selectividad que encargó el

Senado y publicó en su Boletín Oficial el 23 de septiembre de 1997. Los puntos 1- 4 de nuestra propuesta desarrollan el apartado b de dicha recomendación con respecto a los criterios de puntuación. Nuestro punto 5 es una alternativa al sistema de doble corrección recomendado por el Informe, sistema considerado por muchos como inviable.

En la prueba de Selectividad, de tanta repercusión para el alumno y por ende para el sistema educativo, la importancia de cada uno de los elementos que intervienen en el proceso de evaluación no debe subestimarse. Controlar la calidad de la calificación es posible. Confiamos que este estudio invite a considerar los Criterios Global-definido como un camino metodológico que aumenta significativamente la fiabilidad al conseguir calificaciones substancialmente más discriminatorias y una mayor concordancia entre jueces.

Referencias Bibliográficas

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

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

Bueno García, C., Escudero Escorza, T. y Palacín Gil, E. (1990). *Los resultados de la Selectividad: un modelo de análisis.* Zaragoza: Instituto de Ciencias de la Educación, Universidad de Zaragoza.

Cuxart i Jardi, A. (1998). Models estadístics en avaluació educativa: les proves de accés a la universitat. Resumen de tesis doctoral presentada en la Universitat Politècnica de Catalunya, Barcelona.

Cuxart i Jardi, A. y Longford, N. (1998). Monitoring the University Admission Process in Spain. *Higher Education in Europe,* Vol. XXIII, No. 3, 385-396.

Cuxart i Jardi, A., Martí Recober, M. y Ferrer Julià, F. (1997). Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las pruebas de aptitud de Acceso a la Universidad (PAAU). *Revista de Educación,* 314, 63-88.

Escudero Escorza, T. (1997). Investigación sobre el procedimiento de selección de universitarios en España: Revisión comentada. *Revista de Educación,* 314, 7-27.

Escudero Escorza, T. y Bueno García, C. (1994). Examen de Selectividad: el estudio del tribunal paralelo. *Revista de Educación,* 304, 281-298.

Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research,* 60(2), 237-263.

Marín Ibañez, R. (1980). Pruebas objetivas y de ensayo. Zaragoza: Edelvives.

Milanovic, M. y Saville, N. (1996). (eds.) Performance Testing, Cognition and Assessment. Cambridge: Cambridge University Press.

Muñoz-Repiso, M. (1991). Las calificaciones en las Pruebas de Acceso a la Universidad: diferencias de resultados según centro, opción y sexo; en Latiesa et al., (eds.) *La investigación educativa sobre la Universidad*. Madrid: Centro de Publicaciones, Ministerio de Educación y Ciencia, CIDE, 113-133.

Nieda, J., Díaz, M.V., García Barquero, P., Ortega, P., Bonilla, I. y Aguirre, I. (1984). La fiabilidad de las calificaciones en preguntas abiertas de Biología; en Aguirre, I. (ed.) *La Selectividad a debate*, Madrid: Universidad Autónoma, 268-276.

Purves, A. C., T. P. Gorman y S. Takala. (1988). The Development of the Scoring Scheme and Scales; en Gorman, T. P., Purves, A. C. y Degenhart, R. E. (eds.) *The IEA Study of Written Composition I: The International Writing Tasks and Scoring Scales.* Oxford: Pergamon Press. 41-58.

Rodríguez Neira, T., L. Álvarez Pérez, M. Cadrecha Caparrós, J. Hernández García, M. Luengo García, J. Ordoñez Álvarez & E. Soler Vázquez. (1995). *Evaluación de aprendizajes.* Oviedo: Instituto de Ciencias de la Educación de la Universidad de Oviedo.

Sans Martí, A. (1991). Fiabilidad y consistencia del proceso de Selectividad. Un gigante con los pies de barro; en Latiesa et al. (eds.). *La investigación educativa sobre la Universidad*, Madrid: Centro de Publicaciones, Ministerio de Educación y Ciencia, CIDE, 219-227.

Wolf, R. M. (1990). Rating Scales; en Keeves, J. P., (ed.) *Educational research, methodology, and measurement: an international handbook*, second edition. Oxford: Pergamon Press, 496-497.

Wood, R. (1993). *Assessment and Testing*. Cambridge: Press Syndicate of the University of Cambridge.

Apéndice - Criterios Tradicionales

CRITERIOS DE CORRECCIÓN PARA LA PRUEBA DE SELECTIVIDAD 1994-95. INGLÉS

Los criterios de corrección y calificación de la prueba de selectividad, inglés, para el curso 1994-95, elaborados y contrastados teniendo en cuenta el nivel general de conocimientos de los alumnos de C.O.U. y las características del examen, son los siguientes:

Hay que subrayar que se trata de un examen esencialmente de comprensión y que, por lo tanto, un aspecto importante de la prueba es comprobar que el estudiante ha entendido el texto. El objetivo fundamental de las preguntas 1 y 2, y el del resumen de las ideas fundamentales del texto, es el de valorar la capacidad de comprensión del estudiante.

Al valorar la capacidad de expresión, tanto en las preguntas anteriormente mencionadas como en el resto de la prueba, el especialista corrector de la prueba favorecerá a los alumnos que demuestren una riqueza léxica, corrección gramatical, complejidad sintáctica y variedad estilística superior a la media.

La puntuación de los distintos apartados sigue siendo la misma que en años anteriores:

Las preguntas 1 y 2 tienen el valor de 1 punto cada una. Se valorará en estas preguntas la capacidad de comprensión del texto.

La pregunta 3 se puntúa con 2'5 puntos. En esta pregunta el alumno tiene que expresar su opinión y valoraremos positivamente la claridad de expresión, el uso de estructuras complejas y la variedad de vocabulario.

Las cuatro expresiones de vocabulario tienen el valor de 0'25 cada una. En esta sección el estudiante sólo tiene que indicar que identifica correctamente el vocabulario.

La sección de estructuras gramaticales, cuatro apartados, puntuan 0'5 cada uno y, en este apartado, se valora exclusivamente la corrección gramatical. Si hay varias posibilidades se considerará válida cualquier alternativa.

El resumen del texto se puntúa con 2'5 puntos. Es importante en este apartado valorar la capacidad de síntesis del alumno y la indicación de que ha comprendido el texto en su totalidad. La expresión en sus propias palabras y en su propio estilo es fundamental.