Revista de Investigación Educativa, 2002, Vol. 20, n.º 1, págs. 209-223

LA CALIDAD EN EL PROCESO DE CORRECCIÓN DE LAS PRUEBAS DE ACCESO A LA UNIVERSIDAD: VARIABILIDAD Y FACTORES

Rosa Grau, Anna Cuxart
Departament d'Economia i Empresa, Universitat Pompeu Fabra

Manuel Martí-Recober
Universitat Politècnica de Catalunya

RESUMEN

El informe que presentamos resume una investigación realizada en 1997 en el marco de un proyecto más amplio de seguimiento y mejora de las Pruebas de Acceso a la Universidad (PAU) iniciado dos años antes en Catalunya. El objetivo concreto del presente trabajo es el estudio de la calidad de la corrección en cuatro asignaturas que constituyen sendas pruebas en las PAU: Filosofía, Biología, Matemáticas I y Literatura Catalana. El estudio confirma las diferencias existentes entre correctores, siendo mayor la magnitud de dichas diferencias en las asignaturas de Filosofía y de Literatura catalana, así como la necesidad de introducir mejoras en el sistema de elaboración y corrección de las pruebas. Al mismo tiempo, desde el punto de vista metodológico, el estudio pone en evidencia las limitaciones de algunos indicadores clásicos de la calidad, confirmando la idoneidad del modelo de descomposición de la varianza (Longford, 1994) ya especificado por los autores en un estudio anterior. El interés de este tipo de estudios en un momento de transición del sistema PAU-COU al sistema PAU-LOGSE es innegable: aportan información útil en la toma de decisiones.

Palabras claves: *pruebas PAU, nota observada y nota verdadera, fiabilidad en la corrección, modelos de descomposición de la varianza.*

E-mail: rosa.grau@econ.upf.es anna.cuxart@econ.upf.es

ABSTRACT

The educational system in Spain is undergoing a reorganisation. At present, high-school graduates who want to enrol at a public university must take a set of examinations Pruebas de Acceso a la Universidad *(PAU). This paper is related to one of the main educational measurement issues associated with the examinations: reliability of grading. It summarises an experiment designed and conducted to evaluate the quality of grading in four subjects: Philosophy, Biology, Mathematics and Catalan Literature. Elementary summaries, such a cross-tabulations of operational and experimental grades, provide a clear evidence of between-raters differences. A statistically more profound way of summarising the quality of grading is in terms of the variance due to ability, severity and inconsistency (Longford, 1994). The study demonstrates the feasibility of an on-going scheme of partial duplicate grading which would be solely used for quality control and for understanding the inconsistencies occurring in the grading process.*

Key words: *PAU exams, observed and true scores, rater reliability, variance components models.*

I. INTRODUCCIÓN

La fiabilidad de la corrección en las pruebas de Selectividad y la necesidad de medir su magnitud ha sido una preocupación constante entre investigadores y responsables de las pruebas. Véanse, al respecto[1], los trabajos de Sans (1989), de Escudero y Bueno (1994) y el documento del Consejo de Universidades de 1991-93.

En 1991, el Centro de Investigación y Documentación Educativa (CIDE) bajo la dirección de M. Muñoz-Repiso[2] inició una investigación, con alumnos de COU, para medir la calidad de los exámenes elaborando una prueba experimental alternativa en la que la mayoría de las preguntas eran de respuesta cerrada, o bien, de respuesta abierta, pero breves y siempre con pautas de corrección muy precisas. El estudio confirmó el aumento de fiabilidad respecto a las pruebas tradicionales.

En Catalunya, la *Coordinació del COU i de les PAAU*[3] inició en 1995 un proyecto de seguimiento de la calidad del proceso de corrección de las Pruebas de Acceso a la Universidad (PAU). La investigación[4] realizada hasta el momento ha permitido obtener un conocimiento más profundo de todo el proceso, señalando aquellos aspectos que

---

1 El artículo *Las calificaciones en las Pruebas de Aptitud para el Acceso a la Universidad,* de M. Muñoz-Repiso *et al.* de 1991 resume de manera clara y exhaustiva los estudios realizados hasta esa fecha analizando, al mismo tiempo, la incidencia de las modificaciones introducidas en la estructura de los exámenes y la organización de las pruebas.

2 *Estudio experimental de pruebas objetivas para el acceso a la Universidad.* Servicio de Evaluación, CIDE, Ministerio de Educación, Madrid 1991. En este estudio se elaboraron 15 pruebas objetivas y se pasaron a una muestra de 1.307 voluntarios de 16 centros del distrito de Madrid analizando los resultados y comparándolos con los obtenidos, posteriormente, en las pruebas de acceso a la universidad (oficiales).

3 Institución creada en 1989 por el Consejo Interuniversitario de Catalunya (CIC) con la participación del Govern de la Generalitat. En la actualidad, el CIC integra a las siete universidades catalanas.

4 Para más detalle sobre la investigación realizada hasta el momento, véase M. Martí-Recober *y otros* (1998), Cuxart (1998), Cuxart and Longford (1997) y Cuxart *y otros* (1997).

requieren de un control o mejora. Entre dichos aspectos, cabe señalar la variabilidad confirmada entre centros de secundaria en cuanto a los estándares aplicados en la evaluación del COU y la discrepancia existente, entre correctores, en la puntuación de las preguntas de respuesta abierta de las PAU. Al mismo tiempo, se han podido detectar una serie de factores, la mayoría de ellos relacionados con el diseño y elaboración de las exámenes, que repercuten posteriormente en la validez de las pruebas y en la precisión o fiabilidad de la corrección.

En el marco de dicho proyecto de mejora de las PAU, A. Cuxart realizó un experimento de doble corrección con una muestra de exámenes de Matemáticas y Filosofía, en el que participaron todos los correctores de estas asignaturas de 18 tribunales de las PAU de junio de 1995. El experimento se llevó a cabo simultáneamente a la realización de las pruebas, para garantizar al máximo las condiciones habituales de la corrección oficial. El estudio permitió medir la precisión (o fiabilidad) de la corrección y reveló que algunas de las causas de la variabilidad de la corrección estaban relacionadas con la elaboración y formato de los exámenes. Al mismo tiempo, el estudio de Cuxart (1998) confirmó la necesidad de realizar de manera sistemática estudios empíricos y la posibilidad de llevar a cabo experimentos ligados a la realización de las pruebas.

En sus estudios, tanto Muñoz Repiso (1991) como Cuxart (1997) insisten en la necesidad de mejorar algunos aspectos del proceso de las PAU: en la selección y ordenación de los alumnos que pueden acceder a determinados estudios universitarios se utiliza una prueba de cuyo comportamiento se tiene un total desconocimiento empírico. Se carece, por ejemplo, de pruebas piloto que puedan informar sobre el nivel de dificultad y sobre la variabilidad que origina la corrección de los exámenes. Hasta el momento, la mayor parte de los exámenes propuestos en las Pruebas de Acceso a la Universidad se experimentan, por primera vez, en el momento del examen y, sin embargo, se utilizan para ordenar y facilitar o negar el acceso a la Universidad de los estudiantes. Dada la repercusión de este examen en el futuro de cada estudiante, es justo avanzar en la creación de una base de datos de exámenes ya experimentados que pueda ser utilizada en el futuro para confeccionar y elaborar nuevas pruebas.

Con este objetivo como meta, en 1997 se llevó a cabo una segunda experiencia que pretendía ampliar el estudio de la variabilidad de la corrección a nuevas materias, comparar los resultados con el anterior estudio de 1995, e iniciar, al mismo tiempo, el estudio de la dificultad y poder discriminador de las preguntas. El informe que presentamos a continuación se refiere a la primera parte de dicha experiencia: el estudio de la variabilidad de la corrección en las pruebas PAU-COU de junio 1997.

Hemos estructurado el informe como sigue: una breve introducción en la primera sección. En la Sección 2 se especifican los objetivos del estudio así como las características de su diseño. En la Sección 3, tras una breve presentación de la metodología seguida en el análisis de los datos, se resume la exploración de los notas globales del examen. En la sección cuarta se comparan los resultados del presente estudio con los obtenidos en junio 95, incluyendo los que se derivan de la aplicación de un modelo de descomposición de la varianza observada. La Sección 5 incluye las conclusiones que emanan del estudio.

2. OBJETIVOS Y DISEÑO DEL EXPERIMENTO

El estudio que presentamos, se basa en el análisis de cuatro muestras de exámenes de las asignaturas de Filosofía, Matemáticas I, Literatura Catalana y Biología de las pruebas PAU de los alumnos[5] de COU de Junio de 1997. Se trataba de un experimento de triple corrección en el que participaron un grupo de correctores, alrededor de 10 correctores por asignatura. El objetivo del experimento era doble. En primer lugar se trataba de ampliar el estudio de 1995 a nuevas asignaturas, de las cuales se quería obtener un conocimiento en profundidad de la variabilidad de la corrección. Se escogieron las dos asignaturas ya estudiadas en 1995 (Matemáticas y Filosofía) y se optó[6] por Biología (con un formato de muchas preguntas cortas con pautas de corrección muy precisas) y Literatura catalana (examen de dos preguntas de cinco puntos cada una y sin pautas de corrección específicas) por ser dos exámenes de características muy diferenciadas. Así el segundo objetivo del presente estudio era la comparación de la calidad de la corrección entre asignaturas y la posible asociación con los factores formato y existencia de pautas de corrección.

Para llevar a cabo este estudio se seleccionó, al azar, una muestra[7] de, aproximadamente, 100 exámenes de cada una de las citadas asignaturas. De cada ejemplar de examen se hicieron dos fotocopias. Se eligió al azar una muestra de correctores, y cada uno de ellos recibió aproximadamente 20 réplicas o fotocopias junto con los exámenes oficiales que les correspondía corregir. Estas 20 réplicas se asignaron al azar a la variable *réplica1* o bien a la variable *réplica2*. De esta forma se disponía de tres correcciones para cada examen, la del corrector oficial y las dos réplicas. Brevemente, las características del diseño del experimento se pueden resumir en:

- Corrección triple de cada examen
- Asignación al azar de exámenes a los correctores
- Corrección simultánea de exámenes originales y fotocopias, para poder garantizar uniformidad en las condiciones de corrección
- Intervención de un número suficiente de exámenes para garantizar un amplio abanico de situaciones entre el alumnado
- Participación de un número suficientemente grande de correctores que permitiese la representatividad de la muestra de correctores
- Participación de los coordinadores[8] de materia como correctores de fotocopias
- En ningún momento se informó a los correctores sobre si la nota de la fotocopia serviría para evaluar o no al alumno.

---

5 Los exámenes pertenecían a alumnos de 6 centros de secundaria adscritos a la Universitat Pompeu Fabra de Barcelona. Los centros eran de la misma zona geográfica.

6 Por un tema de presupuesto no se pudo extender el estudio a un número mayor de asignaturas.

7 El número de exámenes originales fue inferior al utilizado en 1995. Las necesidad de no obstaculizar la realización de las pruebas obliga a limitar el número de originales que deben ser fotocopiados. Se optó por incrementar el número de fotocopias para asegurar, de este modo, la participación de un número adecuado de correctores.

8 En Catalunya, los coordinadores (uno por materia) son las personas responsables de la preparación del examen.

- Además de la nota global de cada examen, se tomó nota de las puntuaciones por preguntas tanto en los exámenes originales como en las fotocopias.

3. VARIABILIDAD EN LA CORRECCIÓN DEL EXAMEN. PRIMEROS RESULTADOS

Para la exploración de las notas globales de cada una de las cuatro pruebas, se definieron las variables *Diferencia* entre puntuaciones de cada par de correctores y *Amplitud* en la corrección del examen de cada estudiante. El análisis de las distribuciones de ambas variables así como el cálculo de coeficientes de correlación entre cada bloque de puntuaciones ofrecen elementos suficientes para una primera medida (exploratoria) del grado de concordancia entre correctores. Los gráficos de dispersión de cada bloque de puntuaciones respecto de la *nota mediana* de cada examen complementan dicha información. En la Sección 4 se recogen los estimadores de la *fiabilidad* de la corrección derivados de un tratamiento más formal, vía modelización estadística de la varianza observada.

Diferencias

En cada asignatura calculamos la diferencia entre las notas que otorgan dos correctores al mismo examen, en valor absoluto, es decir, sin tener en cuenta el signo. Así para una muestra de tamaño 100, obtuvimos 300 valores al hacer las diferencias, obviando el signo, entre los pares de valores: «corrección oficial-réplica1», «corrección oficial-réplica2», «réplica1-réplica2». Esta variable *Diferencia* toma, pues, el valor cero cuando las dos correcciones del mismo examen coinciden en su puntuación. La Figura 1 muestra la distribución de la variable *Diferencia* en cada una de las asignaturas, destacando la existencia de diferencias superiores a 3 puntos. Según la Tabla 1, las diferencias de dicha magnitud representan un 10.2 % de las diferencias calculadas en Filosofía, un 4.76% de las de Literatura catalana y tan sólo un 0.67 % de las de Matemáticas. En Biología no se encontró ninguna diferencia superior a 3 puntos.

Los histogramas de la Figura 1 muestran patrones de comportamiento diferenciado: mientras que en Matemáticas y Biología se da una concentración de diferencias en valores bajos, inferiores o iguales a 1, las distribuciones de Filosofía y Literatura catalana, por el contrario, presentan mucha más dispersión siendo altamente probables valores de la diferencia superiores a dos y tres puntos.

TABLA 1
RESULTADOS MUESTRA 1997. DISTRIBUCIÓN DE LA VARIABLE DIFERENCIA *(DIF)*

Año 1997	Matemáticas I	Filosofía	Biología	Literatura Catalana
Número de diferencias	300	294	300	210
Dif≤1	81.67%	44.2%	75%	44.76%
1< Dif≤3	17.67%	45.5%	25%	50.48%
3< Dif	0.67%	10.2%	0%	4.76%

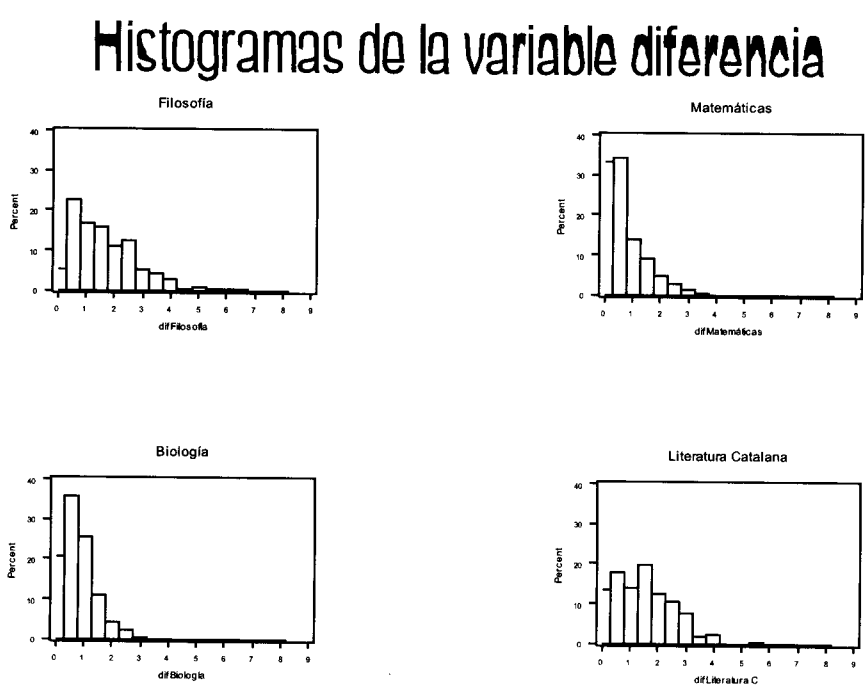

Figura 1
Distribución de frecuencias de la variable Diferencia *por asignaturas.*

TABLA 2
ESTADÍSTICOS DESCRIPTIVOS DE LA VARIABLE DIFERENCIA

Diferencia	N	Media	Mediana	St.dev.	Mínimo	Máximo	Q1	Q3
Matemáticas I	300	0.67	0.5	0.73	0	3.5	0	1
Filosofía	294	1.59	1.45	1.20	0	6.3	0.5	2.25
Biología	300	0.73	0.6	0.59	0	2.9	0.25	1.05
Lit. Catalana	210	1.45	1.5	1.06	0	5.5	0.5	2

La Tabla 2 resume las características numéricas de las distribuciones por asignaturas de la variable *Diferencia.* Las diferencias entre dos correcciones de un mismo examen en Filosofía y Literatura Catalana son, en promedio, de punto y medio aproximadamente, mientras que en Matemáticas y Biología la media de estas diferencias es de aproximadamente 0.7 puntos. Esta variable *Diferencia* toma valores superiores a los dos puntos en una cuarta parte de los exámenes de Filosofía y Literatura Catalana (tercer cuartil *Q3* igual a 2.25 y 2, respectivamente). En cambio, en Matemáticas I y Biología el tercer cuartil se sitúa en un punto. Como casos extremos mencio-

namos los siguientes: en Filosofía un alumno llega a obtener 6.3 puntos de diferencia entre dos de las puntuaciones asignadas; en Literatura Catalana la mayor diferencia observada es de 5.5 puntos. Es importante destacar, también, que en Matemáticas una cuarta parte de las puntuaciones coincidieron totalmente (primer cuartil *Q1* igual a 0). Por otro parte, al estudiar separadamente los exámenes por opciones se pudo observar que la variabilidad generada por la corrección no era similar en las dos opciones de examen (A/B).

Amplitud

Hemos definido la variable *Amplitud* = nota máxima – nota mínima entre las otorgadas por los tres correctores al mismo examen. Esta variable mide la diferencia entre la corrección más generosa y la más severa de cada examen. Según se desprende de la Tabla 3, en Biología y Matemáticas, la variable *Amplitud* es, en promedio, de un punto, mientras que en Filosofía y Literatura Catalana es, en ambos casos, superior a los dos puntos. En estas dos asignaturas y para un 25% de los alumnos, la diferencia entre la puntuación del corrector más benévolo y la del más severo de los tres correctores es superior o igual a tres puntos. En Matemáticas y Biología sólo hay un 5% y un 2% respectivamente de alumnos para los cuales, el hecho de cambiar el corrector más generoso por el más severo de los tres, representa una variación en la puntuación superior o igual a 3 puntos.

TABLA 3
ESTADÍSTICOS DESCRIPTIVOS DE LA VARIABLE AMPLITUD

Diferencia	N	Media	Mediana	St.dev.	Mínimo	Máximo	Q1	Q3
Matemáticas	100	1.01	0.75	0.88	0.00	3.5	0.50	1.50
Filosofía	98	2.39	2.25	1.31	0.00	6.30	1.50	3.00
Biología	100	1.10	1.00	0.63	0.20	2.90	0.61	1.49
Lit. Catalana	70	2.18	2.00	1.06	0.00	5.5	1.50	3.00

Correlaciones

El coeficiente de correlación lineal de Pearson es una medida de la precisión o *fiabilidad* de la corrección. La Tabla 4 muestra los resultados por asignaturas y bloques de correcciones sin distinguir la opción de examen escogida por el alumno. Los coeficientes de correlación, así como los gráficos de dispersión de la Figura 2, nos confirman nuevamente que la concordancia entre las correcciones es muy buena en Matemáticas y Biología. Por el contrario, es necesario recabar esfuerzos para conseguir una mejora en la homogeneidad de la corrección en las asignaturas de Filosofía y Literatura catalana.

TABLA 4

COEFICIENTES DE CORRELACIÓN ENTRE LAS VARIABLES CORRECTOR OFICIAL, RÉPLICA1 Y RÉPLICA2

Coeficientes de correlación	Matemáticas	Filosofía	Biología	Literatura Catalana
N	100	98	100	70
Oficial, réplica1	0.953	0.601	0.840	0.687
Oficial, réplica2	0.911	0.524	0.867	0.633
Réplica1, réplica2	0.907	0.603	0.874	0.555

El alumno puede escoger en cada asignatura una de las dos opciones de examen y, tal como se aprecia en la Tabla 5, las dos opciones no son, en general, comparables en cuanto a la fiabilidad de la corrección. Aunque la opción B de Filosofía, sólo ha sido escogida por un 20% del alumnado preocupa considerablemente que el coeficiente de correlación entre el corrector oficial y la réplica 2 sea, tan sólo, del **0.29**.

TABLA 5

COEFICIENTES DE CORRELACIÓN ENTRE LAS VARIABLES CORRECTOR OFICIAL, RÉPLICA1 Y RÉPLICA2. RESULTADOS POR OPCIONES DE EXAMEN (OPCIÓN A/OPCIÓN B)

	Matemáticas		Filosofía		Biología		Literatura cat.	
Opción de examen	A	B	A	B	A	B	A	B
Número de exámenes	46	54	78	20	72	28	62	8
Oficial, réplica1	.95	.95	.61	.59	.87	.80	.63	—
Oficial, réplica2	.91	.91	.61	**.29**	.92	.72	.65	—
Réplica1, réplica2	.84	.94	.66	.43	.87	.88	.56	—

Dispersión respecto de la mediana

La *nota verdadera*[9] que le correspondería a cada alumno a partir del examen realizado es una variable no observable que estimamos mediante la *nota mediana* de las tres correcciones (ordenadas las tres puntuaciones de menor a mayor valor, la *mediana* es la que ocupa la posición central). Los gráficos de la Figura 2, doce gráficos

9 La *nota verdadera* que le corresponde a cada examen y alumno no es directamente observable. Cada puntuación de examen es una estimación o aproximación de dicha nota. En cualquier caso, la mejor estimación se obtendría si todos los correctores corrigieran cada examen. Al tomar la *mediana* de las tres puntuaciones estamos estimando la *nota verdadera*.

en total, tres por cada asignatura, muestran la dispersión (o la proximidad) de las notas otorgadas por los correctores oficiales y los correctores de las réplicas versus la *nota verdadera*.

El primero de los gráficos de la Figura 2 (nota del corrector oficial de Filosofía versus nota mediana) muestra gran dispersión, la mayoría de los puntos se sitúan por debajo de la bisectriz del primer cuadrante, indicando que el corrector oficial de Filosofía ha puntuado sistemáticamente por debajo de la nota *mediana* de las tres correciones. A modo de ejemplo, si nos situamos en la franja de nota *mediana* igual a 7, encontramos dos alumnos a los que el corrector oficial ha asignado un 3.5 y un 4, respectivamente. Para estos alumnos la valoración del corrector oficial ha diferido (por defecto) de la valoración de sus compañeros en como mínimo 3.5 y 3 puntos, respectivamente. En Literatura Catalana (cuarta columna) la situación es a la

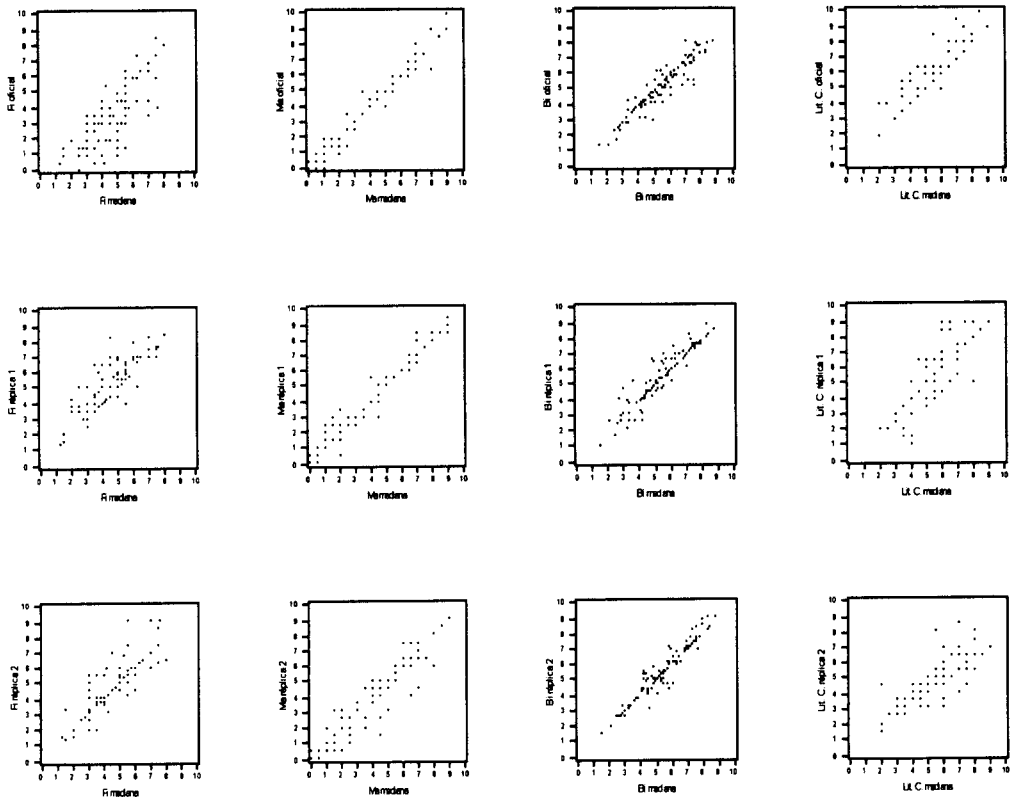

Figura 2

Diagramas de dispersión de las tres puntuaciones (Oficial, réplica1 y réplica2) respecto de la nota mediana *de las tres. Por columnas y en este orden, Filosofía, Matemáticas I, Biología y Literatura catalana.*

inversa: el corrector oficial sistemáticamente ha puntuado los exámenes por encima de las puntuaciones de sus compañeros. En la asignatura de Matemáticas las notas del corrector oficial oscilan alrededor de un punto por encima o por debajo de la nota *mediana* del alumno. El gráfico de Biología es más fino y presenta menos dispersión.

4. INDICADORES DE LA CALIDAD DE LA CORRECCIÓN. COMPARACIÓN CON LOS RESULTADOS DE 1995

La Tabla 6 permite comparar los resultados del presente estudio con los que se derivaron del estudio de 1995 (asignaturas de Filosofía y Matemáticas). Un hecho importante (y estadísticamente significativo) es el aumento de la coincidencia en la corrección de Matemáticas: el porcentaje de exámenes para los cuales la diferencia entre las dos correcciones no supera un punto fue de 71.7 % en 1995 y de 81.7 % en 1997. Es obligado citar que ninguno de los estudiantes que había escogido la opción A en esta materia respondió correctamente a la pregunta 1A (un ejercicio de probabilidad), la mayor parte obtuvieron un 0 sobre un máximo de 2 puntos en esta pregunta. Podría ser que este hecho (no deseable) haya contribuido al aumento de la concordancia. Resta pendiente estudiar a fondo este posible efecto. La conclusión es que en Matemáticas mejoró la fiabilidad en 1997 pero tenemos nuestras dudas sobre el poder discriminador y la pertinencia de algunas de las preguntas. En Filosofía no se aprecia ninguna mejora: la magnitud de la discrepancia entre correctores sigue siendo importante. Cabe señalar que los avances en el estudio de la doble corrección iniciado en 1995 se comunicaron a los coordinadores de las dos materias y que a raíz de los resultados del estudio se elaboraron a partir de junio de 1997 pautas de corrección[10] ajustadas al examen propuesto de Matemáticas I.

TABLA 6
DISTRIBUCIÓN DE LA VARIABLE DIFERENCIA *(DIF). COMPARACIÓN DE RESULTADOS*

	Matemáticas		**Filosofía**	
Año	1995	1997	1995	1997
Número de diferencias	374	300	726	294
Dif≤1	71.7 %	81.7 %	51 %	44.2 %
1< Dif≤3	26.2 %	17.7 %	35.8 %	45.5 %
3< Dif	2.1 %	0.7 %	13.2 %	10.2 %

10 También a partir de 1997, la elaboración de pautas específicas de corrección se extendió a todas las materias LOGSE.

Al mismo tiempo, los datos recogidos para el presente estudio han permitido confirmar la validez del modelo de descomposición de la varianza y su aplicación al cálculo de indicadores de la calidad de la corrección. Con el objetivo de profundizar en el estudio de la discrepancia observada, en Cuxart *et al.* (1997) se especifica un modelo de variación que descompone el *error de medida* introducido en la corrección en sus diferentes fuentes de variación. Dicho enfoque se enmarca en la adaptación propuesta por Longford (1995, Chap.2) de la teoría de la Generalizabilidad (Cronbach *et al.*, 1972) para el estudio de datos relativos a exámenes y correctores. Longford (1995) distingue entre dos posibles fuentes de discrepancia en la corrección[11]: la *severidad* y la *inconsistencia*. El modelo concreto de componentes de la varianza propuesto para explicar la variación de la puntuación de un examen es el modelo aditivo:

$$y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$$

siendo $i = 1, 2, \ldots, I$ el índice del examen o estudiante; $j=1, 2, \ldots, J$ el del corrector. El número de puntuaciones que entran en el estudio es $3I$; y_{ij} es la puntuación que el corrector j ha dado al examen i; α_i es a puntuación *verdadera* y no observable del examen i; β_j es la *severidad* del corrector j; ε_{ij} representa la *inconsistencia específica* de cada corrección. Suponemos que estos tres últimos términos están mutuamente no correlacionados con medias iguales a μ, 0 y 0 y varianzas σ_a^2, σ_b^2 y σ_e^2, respectivamente. α_i sería la media que obtendríamos si todos los correctores corrigieran el examen i, mientras que β_j sería la diferencia entre la media global μ (todos los exámenes corregidos por todos los correctores) y la media correspondiente al corrector j (todos los exámenes y este corrector); ε_{ij} recogería la separación del corrector j en el examen i respecto de su comportamiento medio.

Una buena corrección requiere que las componentes de la varianza relativas a la *severidad* y a la *inconsistencia* sean pequeñas con relación a la varianza de la nota verdadera. De ahí que las estimaciones de las componentes de la varianza así como otros estadísticos[12] que de ellas se derivan puedan considerarse como indicadores de la calidad de la corrección.

La Tabla 7 ofrece las estimaciones de las tres componentes de la varianza observada así como la proporción de varianza total correspondiente a cada componente. La tabla incluye los resultados de las dos ediciones: 95 y 97. El método de estimación uti-

---

11 Por *severidad* de un corrector, entenderemos la diferencia entre dos cantidades no observables: «la media del corrector (que conoceríamos si dicho corrector corrigiera todos los exámenes) y la media global» (calculable si todos los exámenes fueran corregidos por todos los correctores de la población). La *inconsistencia* o «error no sistemático» es una amalgama de imperfecciones presentes en el proceso de corrección. La *inconsistencia específica* de cada examen y corrector sería la «desviación de la puntuación otorgada respecto a la puntuación que en promedio dicho corrector otorgaría al examen en cuestión».

12 En la tesis de A. Cuxart (1998) se proponen diversos estadísticos que permiten la revisión de las hipótesis del modelo, ofreciendo, al mismo tiempo, elementos de diagnosis y detección de comportamientos extremos (correctores que discrepan ostensiblemente de sus compañeros, correctores que adjudican notas excesivamente similares,...).

lizado ha sido el de los momentos. De ahí que el parámetro correspondiente a la *severidad* sea el que acusa la menor eficiencia[13] en la estimación, debido al reducido número de correctores que participaron en el experimento (alrededor de 10 por asignatura) en comparación con el número de exámenes. En 1997, la calidad de la corrección en Literatura catalana es baja, similar a la de Filosofía. En cambio, el examen de Biología ha dado valores de los indicadores cercanos a los de Matemáticas. La *inconsistencia* es presente en la corrección de las cuatro asignaturas, mientras que solamente se aprecian diferentes grados de *severidad* entre los correctores de Filosofía y de Literatura catalana. En estas dos asignaturas la contribución del *error de medida* introducido en la corrección es similar a la contribución de la *nota verdadera* e, incluso, superior a este último en Filosofía 1997 (58.8 % frente a 41.2 % de la variación total observada).

En el caso de Filosofía[14] puede que, en un principio, sorprenda la diferencia existente entre los valores de la correlación muestral de la Tabla 4, y la estimación del *coeficiente de fiabilidad*[15] $r = 0.412$ que se deduce de la Tabla 7. Cabe recordar que la correlación muestral es un buen estimador de la *fiabilidad* siempre que $\sigma_b^2 = 0$. En el caso que $\sigma_b^2 > 0$, el valor de la correlación muestral depende del diseño (Longford, 1995), es decir de cómo se han asignado los exámenes a los correctores. El presente ejercicio ha puesto en evidencia las limitaciones del estimador correlación muestral como estimador de la *fiabilidad*, aportando argumentos a favor de la utilización, para un segui-

TABLA 7
ESTIMACIÓN DE LAS COMPONENTES DE LA VARIANZA DE LA PUNTUACIÓN OBSERVADA. EDICIÓN DE 1997

Por columnas, $\hat{\sigma}_a^2$, varianza entre *notas verdaderas* α_i;
$\hat{\sigma}_b^2$, varianza de la *severidad* β_j;
$\hat{\sigma}_e^2$, varianza de la *inconsistencia* ε_{ij}.

	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\sigma}_e^2$	**Var. total**
Matemáticas	5.738 (92.1%)	0.163 (2.6%)	0.329 (5.3%)	6.230
Filosofía	1.390 (41.2%)	0.641 (19.0%)	1.342 (39.8%)	3.374
Biología	2.462 (84.8%)	0.143 (4.9%)	0.299 (10.3%)	2.905
Literatura cat.	2.134 (57.0%)	0.528 (14.1%)	1.085 (29.0%)	3.463

13 En la edición de 1995 se realizaron simulaciones por el método BOOTSTRAP para estimar los errores estándar de los estadísticos calculados. Según dichas simulaciones, el único parámetro no significativo resultó ser el correspondiente a la *severidad* en Matemáticas, los cinco restantes resultaron significativos.

14 Para el resto de asignaturas las correlaciones muestrales calculadas se mantienen alrededor del coeficiente de fiabilidad r o proporción de varianza total explicada por la nota verdadera.

15 En este contexto, al hablar de *coeficiente de fiabilidad* nos referimos al cociente entre la varianza de la *nota verdadera* y la varianza total observada. Este coeficiente representa la proporción de varianza total (observada) explicada por la variabilidad de la *nota verdadera* (o nivel de preparación demostrada por los alumnos en el examen).

miento adecuado de la calidad de la corrección, de los indicadores que se derivan de la modelización vía descomposición de la varianza.

5. CONCLUSIONES. DISCUSIÓN

Destacamos a continuación las principales conclusiones que se desprenden del estudio sobre la calidad de la corrección de las PAU 97.

- Se observa un comportamiento similar en las asignaturas de Matemáticas y Biología en claro contraste con Filosofía y Literatura Catalana. Los coeficientes de correlación son del orden de 0.9 en el primer grupo de asignaturas mientras que en el otro grupo son del orden de 0.6. La distribución de *Diferencias* se encuentra más concentrada en los valores bajos en el primer grupo. La media de la variable *Amplitud* entre las tres correcciones es de, aproximadamente, 1 punto en el primer grupo y superior a 2 puntos en el segundo grupo.
- Se observa un aumento de la concordancia en la corrección de Matemáticas respecto al estudio realizado en el año 1995. Dado que en este periodo de tiempo se han hecho esfuerzos para concretar las pautas específicas de corrección, podríamos inferir que estas pautas ayudan a reducir las discrepancias entre correcciones y sería menester incorporarlas en aquellas asignaturas que aun no disponen de ellas.
- La asignatura de Biología muestra una fiabilidad muy alta. El formato del examen, acompañado de una puntuación muy fina con pautas precisas, podría explicar la concordancia existente entre correctores reflejada tanto en los diagramas de dispersión como en los estadísticos calculados.
- Al finalizar el examen y previa a la corrección, los correctores de las asignaturas de Matemáticas y Biología disponen de pautas específicas de corrección del examen propuesto. En las asignaturas de Filosofía y Literatura Catalana los correctores sólo disponen de los criterios generales de corrección. Este hecho podría explicar, en parte, el comportamiento diferenciado de un grupo de asignaturas respecto al otro.
- Entendemos que una parte importante de la variabilidad en la corrección de las asignaturas de Filosofía y Literatura catalana se explica por estos dos factores: pocas preguntas y de respuesta abierta y ausencia de pautas de corrección específicas del examen. Estos dos factores aportan *incertidumbre* al proceso de evaluación.
- Otras fuentes de *incertidumbre* aparecen relacionadas con el sistema de puntuación, que obliga a redondear sin admitir puntuaciones intermedias, o a introducir penalizaciones, el efecto de las cuales podrían limitarse con la preparación minuciosa de pautas de corrección y con el entrenamiento de los correctores.
- Nuestra impresión es que los correctores no son «expertos». También es cierto que los coordinadores deberían esforzarse en saber comunicar a los correctores qué es lo que quieren medir o valorar en el examen y en cada pregunta en con-

creto. En este sentido, el conocimiento empírico *a priori* sobre la dificultad y poder discriminador de las preguntas, facilitaría la elaboración de pautas de corrección específicas de cada examen. Al mismo tiempo, la experimentación de las preguntas previamente a su utilización en el examen PAU permitiría conocer las posibles respuestas de los alumnos y consensuar la valoración de las mismas.

Defendemos la realización sistemática de estudios empíricos, que permitan conocer la calidad del proceso y, a partir de ahí, realizar un seguimiento adecuado del mismo. Entendemos, como se ha podido ver en el presente estudio, que el tratamiento estadístico de los datos no debería obviar un apartado previo de descripción de variables de interés que permita conocer, a través de las distribuciones, el patrón de comportamiento y las características numéricas de las mismas y que sugiera el estudio cualitativo de los casos extremos. El coeficiente de correlación muestral, utilizado con frecuencia para estimar la fiabilidad, se ve afectado en el presente estudio por la existencia de diferentes grados de severidad entre correctores. De ahí que, en su lugar —o como complemento ineludible— propongamos la aplicación del modelo de descomposición de la varianza de la Sección 4 que, tanto en su planteamiento como en el posterior cálculo de estimadores aparece libre del diseño que constriñe al cálculo del coeficiente de correlación muestral. El modelo aplicado admite que cada examen no sea corregido por el mismo número de correctores. En el futuro se podría ampliar el modelo incluyendo en el mismo otras fuentes de variabilidad o incertidumbre como el hecho de que el corrector haya participado o no en un curso de aprendizaje-consensus,...

6. BIBLIOGRAFÍA

Cronbach, L. J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: John Wiley.

Cuxart Jardí, A. (1998). *Models estadístics en avaluació educativa: les proves d'accés a la universitat.* Tesis doctoral. Universitat Politècnica de Catalunya. Barcelona.

Cuxart Jardí, A. and Longford, N.T. (1998). Monitoring the university admissions process in Spain. *Higher education in Europe.* UNESCO. Vol. XXIII, No. 3, pp. 385-396.

Cuxart, A., Martí, M. y Ferrer, 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. *Revista de Educación,* 314: 63-88.

Escudero, T. y Bueno García, C. (1994). Investigaciones y Experiencias: Examen de Selectividad. El estudio del tribunal paralelo. *Revista de Educación,* 304.

Longford, N.T. (1995). *Models for uncertainty in Educational Testing.* Springer Series in Statistics. New York.

Martí Recober, M. y otros (1998). *Los sistemas de corrección de las pruebas de Selectividad en España. Análisis y propuestas.* Concurso nacional de Proyectos de Investigación Educativa, 1995. Ministerio de Educación y Ciencia, CIDE. Documento inédito.

Muñoz-Repiso, M., Muñoz, F., Palacios, C. y Valle, J. (1991). *Las calificaciones en las Pruebas de Aptitud para el Acceso a la Universidad,* colección *INVESTIGACIÓN,* nº 61. Madrid: CIDE.

Muñoz-Repiso, M. *Estudio experimental de pruebas objetivas para el acceso a la Universidad.* CIDE (1991). Documento inédito.

Sans, A. (1989). Fiabilidad y consistencia del proceso de selectividad. *La investigación educativa sobre la universidad,* pp. 201-208. Madrid: CIDE.

Fecha de recepción: 3 de mayo de 2001.
Fecha de aceptación: 15 de febrero de 2002.