Algunos factores que inciden en el rendimiento y en la evaluación de los alumnos en las PAAU

Anna Cuxart Jardì[1], Manuel Martí Recober[2], Ferran Ferrer Julià[3]

Agradecimientos

El presente artículo muestra los resultados de distintas investigaciones, financiadas en parte por una ayuda a la investigación *DGICYT PB93-0403* de la Generalitat de Catalunya y por el *Concurso Nacional de Proyectos de Investigación Educativa, 1995* del Ministerio de Educación y Cultura de España.

[1] Departament d'Economia, Universitat Pompeu Fabra.
[2] Catedrático de la Universidad Politécnica de Cataluña, Departamento de Estadística e Investigación Operativa.Coordinador del COU y de las PAAU de Cataluña.
[3] Departamento de Pedagogía Sistemática y Social, Universida Autónoma de Barcelona.

Abstract

The context where the university admissions exams are performed is presented and the main concerns about this exams are outlined and discussed from a statistical point of view. The paper offers an illustration of the use of random coefficient models in the study of educational data. The association between two individual scores (one internal and the other external to the school) and the effect of the school in the external exam is analized by a regression model with random intercept and fixed slope. A variance component model for the analysis of the grading process is also presented. The paper ends with an outline of the main findings and the presentation of some specific proposals to improve and control the equity of the system. Some pedagogic reflections are also included.

Key words: *PAAU exams, COU' scores, admissions process, random coefficient models, variance component models, rater reliability*

Journal of Economic Literature classification: C89, I19

1. Presentación y antecedentes

En junio de 1993 el Boletín Oficial del Estado publica la Orden Ministerial de 9 de junio sobre pruebas de aptitud para el acceso a las Facultades, Escuelas Técnicas Superiores y Colegios Universitarios, donde señala: "…Una vez finalizado todo el proceso de las pruebas, cuando se observe una elevada desviación entre las medias de los expedientes[4] de los alumnos y las calificaciones globales otorgadas por un Tribunal,… , procederán a estudiar las causa… y a proponer las oportunas medidas en relación con los Centros o Tribunales correspondientes". El mismo año, la Memoria de Actividades del Consejo de Universidades, insiste en la necesidad de controlar la diferencia entre la media de Expediente y la media de PAAU a nivel de centro. Ese mismo año y de manera paralela al trabajo sistemático que se realiza en la *Oficina de Coordinación del COU i les PAAU de Catalunya*, Anna Cuxart inicia un trabajo de investigación[5] dirigido por Manuel Martí-Recober. El objetivo es investigar aquellos modelos estadísticos que puedan ser de utilidad en el análisis de los resultados académicos de las PAAU, con una atención especial al estudio de las variables o factores que inciden en la evaluación de los alumnos que se presentan a dichas pruebas. Las preguntas que nos planteábamos eran:

- ¿Cual es el grado de asociación entre las puntuaciones que obtienen los estudiantes en el COU y en las PAAU? En cada materia y globalmente. Es decir, ¿cual es la capacidad predictiva de una puntuación respecto de la otra?.
- ¿Existe homogeneidad entre los Centros que imparten COU en cuanto a los resultados que obtienen sus alumnos en las PAAU? ¿Y en cuanto a la puntuación que ellos les otorgan? ¿Se puede hablar de puntuación uniforme (estandarizada) en COU?.
- Y en las PAAU, ¿existen diferencias apreciables entre correctores de una misma materia que desvirtuaría el principio de *prueba estándar*?

Las características de la investigación realizada y que a continuación resumimos son:

- La modelización estadística que se construye a partir de datos individuales
- La utilización de amplia información[6] sobre cada alumno: notas de COU y de PAAU por asignaturas -ésta es una de las diferencias a destacar respecto de

[4] La media de Expediente (o *nota de Expediente*) de cada alumno es la media aritmética entre los cuatro cursos de secundaria (tres cursos de bachillerato y el COU). La *nota de acceso a la Universidad* que servirá más tarde para ordenar los estudiantes y ubicarlos en los estudios superiores es el resultado de calcular la media aritmética entre la nota de Expediente y la nota PAAU (media ponderada del conjunto de Pruebas de Aptitud para el Acceso a la Universidad, PAAU).

[5] El resultado de dicha investigación constituye el núcleo de una tesis doctoral en curso, de la cual se han presentado resultados parciales en Cuxart, Graffelman i Martí (1995); Cuxart and Longford (1996); Martí y otros (1995).

[6] La Oficina de Coordinació del COU i les PAAU de Catalunya dispone de las notas por asignaturas de todos los alumnos que han cursado el COU. Puesto que el examen PAAU se basa en las materias cursadas en COU nos ha parecido más adecuado analizar la relación entre estas puntuaciones que entre la nota de Expediente y la de PAAU.

otros estudios publicados-, características personales y del Centro en que estudió el COU,…
- Una investigación empírica que se lleva a cabo para detectar y cuantificar las imperfecciones del proceso
- La propuesta de ajustes incorporados al proceso (autorevisión) con el objetivo de obtener una estimación más eficiente de la preparación de los alumnos. En este sentido se señalan ciertas perspectivas de automatización en el futuro.

Hemos estructurado el presente artículo en tres grandes apartados: La influencia del centro escolar en la predicción de la nota PAAU; La influencia del profesor-corrector; y, para terminar, un apartado de Conclusiones que incluye perspectivas de investigación así como consideraciones pedagógicas a la luz de las estadísticas.

2. La influencia del centro escolar en la predicción de la nota PAAU de cada alumno

2.1 Anteriores estudios

Muchos han sido los investigadores que han estudiado con profundidad las llamadas pruebas de Selectividad. Entre los pioneros, Tomás Escudero y I. Aguirre de Cárcer. En el documento anteriormente citado del Consejo de Universidades se incluye una amplia bibliografía que permite conocer cuáles han sido las preocupaciones de los investigadores, sus principales aportaciones y las sucesivas modificaciones introducidas en la normativa de las pruebas a la luz de dichas investigaciones. Nos limitaremos a mencionar aquellos trabajos que por sus objetivos y resultados están directamente relacionados con los estudios que presentamos. Antoni Sans (1990) analiza los resultados en las pruebas PAAU de los 12423 estudiantes matriculados de COU el curso 1986-87 y adscritos a la UAB, encontrando diferencias significativas en la nota PAAU entre tribunales y entre tandas (convocatorias de exámenes). Sans llega a proponer que no se tenga en cuenta la nota de Expediente en el cómputo de la nota de acceso a la Universidad a la luz de las discrepancias observadas entre nota de expediente y nota PAAU en algunos centros. Precisamente los centros con peores resultados en las PAAU eran los que habían puntuado más alto a sus alumnos. El manifiesto comportamiento heterogéneo entre centros le lleva a esta recomendación. El desequilibrio existente se manifiesta también entre centros privados y públicos: mientras que de los alumnos matriculados en los centros públicos, el 50% superan el COU y el 38% las PAAU, en los centros privados estos porcentajes ascienden a 78% y 66%, respectivamente.

Mercedes Muñoz *y otros* (1991) resumen en su trabajo los principales resultados de las investigaciones sobre las pruebas de acceso a la universidad. Al mismo tiempo estudian la incidencia de las modificaciones introducidas en su formato y organización a partir de los resultados de las convocatorias 1987, 88 y 89 relativos a una muestra estratificada de universidades de todo el Estado. Destacamos alguna de las conclusiones del trabajo de M. Muñoz en cuanto al sistema de acceso actual:

"Podrían resumirse sus defectos en que cumple mal y de forma desigual la función que se le asigna de ordenar de manera aquilatada[7] a los estudiantes en cuanto a su prioridad para obtener un puesto en la universidad"

2.2 Los datos. Una primera exploración

Para abordar el estudio de la asociación entre la *nota PAAU* y la *nota COU*[8] de cada estudiante, se escogió una muestra aleatoria[9] de 26 centros (1619 estudiantes) del distrito de Catalunya, entre los 400 centros (unos 25000 estudiantes en total) que concurrieron a las PAAU en junio de 1993. Nuestra hipótesis de trabajo era que esta relación varía de un centro a otro. Se trataba pues de escoger un modelo de variación de la *nota PAAU* que incluyera la *nota COU* como variable explicativa (entre otras) y que a la vez contemplara la posible diferencia entre centros.

En una primera exploración constatamos que la correlación intra-centros[10] para la *nota PAAU* era de 0.195. Es decir, aproximadamente, un 20% de la variación de la nota PAAU observada se debía a variación entre centros. Este hecho desaconsejaba la aproximación clásica, vía un modelo de regresión con una misma ecuación para los 1619 estudiantes[11].

El Gráfico 1 nos muestra:

a) La devaluación de nota (de COU a PAAU) que sufren la mayoría de estudiantes. Si la *nota COU* y la *nota PAAU* estuvieran midiendo la misma variable latente "aptitud del alumno para los estudios universitarios", sería de esperar que el gráfico de dispersión se situara alrededor de la recta bisectriz de este primer cuadrante. Pero no es así, los puntos se sitúan alrededor de una recta paralela a dicha bisectriz.

---

[7] La preocupación sobre el grado de homogeneidad existente entre los centros al puntuar a sus alumnos sigue siendo motivo de preocupación en uno de los estudios publicados recientemente. López, Mª del R. (1997) estudia los resultados de las pruebas de acceso a la Universidad Autónoma de Madrid del curso 95-96. La distribución por centros (132 centros de secundaria) de la diferencia entre la nota de Expediente y la nota de las PAAU presenta un promedio de -1.6 con una variabilidad que va desde -0.8 hasta -2.8. Para los estudiantes del batxillerato LOGSE el promedio de dicha diferencia por centros (14 centros) es de -1.9 variando entre -0.9 i -3. Luego, no parece que el nuevo sistema educativo vaya a reducir estas diferencias en la evaluación de los alumnos. Sigue pues siendo un tema pendiente.

[8] *Nota COU* es la media aritmética de las ocho asignaturas cursadas mientras que *nota PAAU*, como decíamos antes, es una media ponderada de los nueve ejercicios que componen dichas pruebas en Catalunya.

[9] La Oficina de Coordinació del COU i les PAAU de Catalunya suministró los datos.

[10] El coeficiente de correlación intra-grupos es una medida del grado de homogeneidad existente dentro de los grupos. La estimación que hemos tomado es la que propone Muthén (1994) a partir de la descomposición derivada de un análisis de la varianza clásico. Para más detalle ver Cochran (1977).

[11] Al no satisfacer la hipótesis de incorrelación entre residuos, los estimadores de cuadrados mínimos (MCO) subestiman los errores estándar de los coeficientes de la ecuación de regresión. Esta "subestimación" de los errores estándar conlleva que no puedan darse por válidos los test que suelen ofrecer los paquetes de software estadístico al realizar un análisis de la regresión. Al mismo tiempo, y puesto que nuestro interés se centraba en toda la población de centros no únicamente en los 26 de la muestra, desestimamos los modelos de análisis de la covarianza (ANCOVA) que ofrecen la estimación del efecto fijo debido a cada centro sin inferir sobre la población de los mismos.

Existe un sesgo entre una valoración y la otra. Está claro que esta diferencia se distribuye de manera desigual entre los estudiantes. Nos preguntamos, ¿cómo se distribuye entre centros?

b) Las distribuciones marginales de la *nota COU* y de la *nota PAAU*. Más de la mitad de los estudiantes tienen una *nota COU* comprendida entre 6 y 7. Cabe recordar que los alumnos que se presentan a las PAAU deben haber aprobado el COU, es decir tienen una *nota COU* superior a 5.5. Un 65% de los estudiantes obtiene una *nota PAAU* entre 4 y 6; 8.4% obtiene una nota inferior a 4; 20.7% entre 6 y 7; 6.2% entre 7 y 9; 0.1% superior a 9.

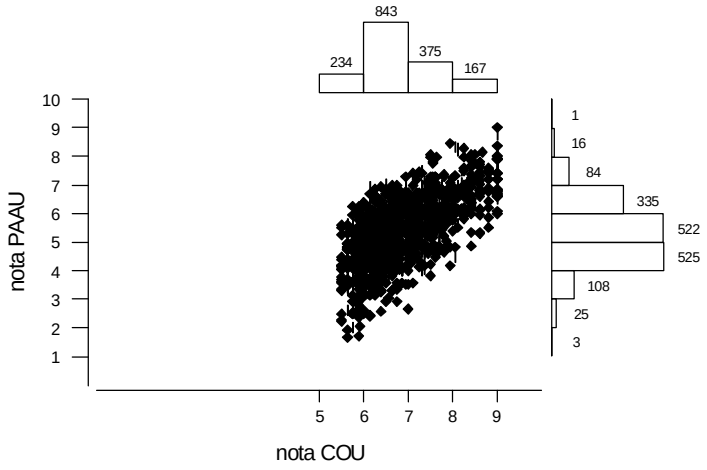


Gráfico 1. Nota PAAU versus nota COU (1619 estudiantes)

Si en el gráfico anterior nos centramos solamente en los resultados obtenidos por los estudiantes de dos escuelas concretas (la nº 2 y la nº 18 de nuestra muestra), observamos la siguiente situación:

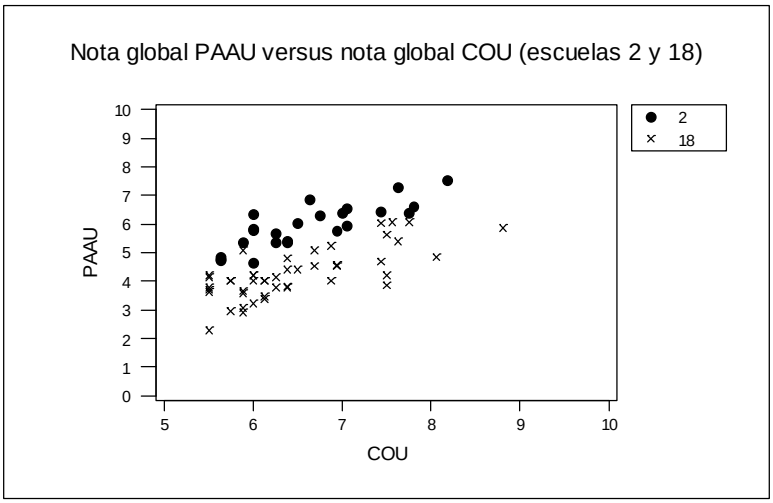


Gráfico 2 . Notas de los alumnos de los centros 2 y 18. Las notas medias de COU y de PAAU son, respectivamente: 6.62 y 5.91, el primero; 6.46 y 4.17, el segundo.

El Gráfico 2 sugiere la conveniencia de estimar un modelo de regresión con parámetros que puedan tomar valores diferentes[12] para cada escuela.

2.3 Asociación entre la nota de COU y la nota de las PAAU

Con el objetivo de poder hacer inferencias sobre la población de centros y a la vez estimar eficientemente los parámetros de interés, decidimos modelizar la variación de la *nota PAAU* vía un modelo de regresión de coeficientes aleatorios (parámetros que varían de una escuela a otra de acuerdo con una distribución de probabilidad) tomando la *nota COU* como variable explicativa. Aitkin and Longford (1986) proponen la utilización de dichos modelos para el estudio de datos de estructura jerárquica. En nuestro caso los estudiantes (o unidades del primer nivel) aparecen agrupados en centros escolares (o unidades de segundo nivel). El método de estimación seguido ha sido el de Mínimos Cuadrados Generalizados Iterativos (MCGI) con el software informático Mln (ver Goldstein, 1995).

Tabla 1. La variación de la *nota PAAU* individual con relación a la *nota COU,* el género, el ser o no repetidor de COU y la opción de COU cursada.

	Modelo 1 MCO	Modelo 2 MCGI	Modelo 3 MCGI	Modelo 4 MCGI
coeficientes (E.E.)				
constante	-0.84 (0.17)	5.24 (0.10)	-0.68 (0.17)	-0.36 (0.18)
NOTCOU	0.90 (0.03)		0.88 (0.02)	0.85 (0.02)
SEXO				0.20 (0.02)
REPCOU				-0.23 (0.05)
OPA				-0.27 (0.04)
OPB				-0.41 (0.05)
varianzas				
entre centros	-	0.22 (0.07)	0.18 (0.07)	0.18 (0.05)
entre estudiantes	0.706	1.03 (0.04)	0.52 (0.02)	0.48 (0.02)
coef. de correlación intra-centros	-	0.175	0.25	0.27

En la Tabla 1 presentamos los resultados que se derivan del ajuste de diferentes modelos a los datos de la muestra así como las variables incorporadas en cada uno de estos modelos. El Modelo 1 es un análisis de la regresión clásico en que no se tiene en cuenta la agrupación de los estudiantes en centros. El Modelo 2 es el más simple de los modelos de coeficientes aleatorios de dos niveles: descompone la varianza observada de la *nota PAAU* en varianza entre centros y varianza entre alumnos dentro de los centros; estima la media global de la *nota PAAU* (5.24) y permite calcular una estimación del coeficiente de correlación intra-centros. El Modelo 3, amplía el modelo anterior incluyendo la variable explicativa

[12] Si aplicáramos un modelo de regresión clásico por separado a estas dos escuelas se obtendrían dos rectas diferentes quizás paralelas, es decir con la misma pendiente y diferente ordenada en el origen.

nota COU . Como resultado de la estimación[13] podemos decir que el coeficiente de dicha variable (pendiente de la recta) es el mismo para todos los centros, no varía, y en cambio la constante (ordenada en el origen) varía entre centros. Esta constante que varia de un centro a otro está recogiendo el efecto debido al centro escolar[14]. El modelo 3 se amplió añadiendo diversas variables cualitativas como son el género, el ser repetidor de COU, la opción de COU, el tipo de centro (público o privado). El tipo de centro no resultó ser una variable significativa. No se observa diferencia entre la opción C y la D, que tomamos como opción base o referencia El modelo más completo que además de la variable *nota COU* incorpora otras variables que explican la variación de la *nota PAAU* es el Modelo 4. La exploración de los residuos del Modelo 4 no puso en evidencia ninguna violación de las hipótesis sobre los mismos con lo cual es un buen modelo parar explicar la variación de la *nota PAAU.* Según estos dos modelos (3 y 4), el efecto centro en la nota PAAU es significativo y se reduce a un término aditivo de media 0 y varianza 0.18. En consecuencia, para cada centro la recta de regresión de PAAU sobre COU tiene una constante propia, resultante de sumar la constante común a todos los centros (-0.68 en el Modelo 3) con el *efecto centro* correspondiente, y una pendiente fija (0.88 en el Modelo 3), la misma para todos los centros. Se confirma, pues, el patrón de comportamiento que sugería el Gráfico 2.

Es de destacar que al ir añadiendo variables explicativas (mod. 2, 3 y 4) se va reduciendo la varianza entre centros y entre estudiantes -esta última en mayor grado-, pero el coeficiente de correlación intra-centros aumenta. Al ajustar (corregir) la variación de *la nota PAAU* con las sucesivas covariantes se hace todavía más ostensible la homogeneidad interna de los centros frente a la heterogeneidad de los mismos en lo que se refiere a la puntuación en las PAAU.

Centrándonos en la interpretación del modelo 4, la *nota COU* es predictora[15] de la puntuación que se puede conseguir en las PAAU (matizada con una serie de características individuales) pero también el factor centro escolar tiene carácter predictivo. Es importante destacar ciertas observaciones al respecto:

- Por el hecho de haber cursado el COU en un centro u otro y respecto al comportamiento medio de la población, la *nota PAAU* esperada de cada estudiante sufrirá un incremento que puede oscilar entre -0.84 y 0.84 puntos[16].
- En las Pruebas PAAU a los chicos les va mejor (0.20 puntos en promedio) que a las chicas en igualdad de condiciones con respecto al resto de variables.
- El hecho de ser repetidor disminuye la predicción de nota en 0.23 puntos.

[13] Hemos desestimado el modelo de dos niveles que contempla la ordenada en el origen y la pendiente variando de un centro a otro ya que la varianza de este último coeficiente no resultó significativa (no se puede rechazar la hipótesis de que dicho coeficiente sea constante) y además dicho modelo ofrecía un peor ajuste (test de la razón de verosimilitud) en comparación con el Modelo 3.

[14] A partir de este momento, para abreviar, nos referimos al efecto debido al centro estimado en los modelos 2, 3 y 4 como el *efecto centro.*

[15] El coeficiente de correlación lineal entre la *nota COU* y la *nota PAAU* de cada estudiante es 0.66.

[16] $0.84 = 2DE = 2\sqrt{0.18}$

- A los estudiantes de Ciencias (0.27 por debajo de la media para los de la opción A y 0.41 para los de la opción B) les va, en general, peor que a sus compañeros de Letras (opciones C y D).

2.4 El efecto debido al centro escolar y las escalas de medida

Los resultados de la estimación del modelo 4 corroboran los encontrados por Muñoz-Repiso (1991) *y otros* a nivel estatal. Nuestro enfoque, vía modelización estadística, ofrece, además, la posibilidad de obtener un identificador para cada centro[17]. Los centros que presentan un asociación *extrema* entre la *nota PAAU* y la *nota COU* deberían ser motivo de estudio. El análisis de las características de dichos centros y la discusión con los responsables de los mismos aportará un mayor conocimiento sobre la diversidad[18] de centros.

Mientras la *nota PAAU* presenta una variación entre centros considerable no ocurre lo mismo con la *nota COU*. El coeficiente de correlación intra-centros para esta nota es prácticamente 0. Se diría que las distribuciones de puntuaciones (aprobados) en COU no varían de un centro a otro. Cada centro ha "ordenado" sus alumnos con una media y una variabilidad muy similares. En cambio, según acabamos de ver, al proponer el mismo examen a todos los alumnos se desvelan las diferencias entre los resultados de los alumnos de un centro a otro. Nuestra conclusión es que los centros están utilizando escalas de medida diferentes y distintos estándares en la preparación de los alumnos.

2.5 La información a los centros. Indicadores y información

Los estudiantes concurren a las PAAU a través del centro en que han cursado el COU. Cada centro está adscrito a una universidad. En Cataluña, la *Oficina de Coordinació del COU i les PAAU* coordina y administra la realización de estas pruebas (los exámenes son los mismos para todas las Universidades) y , a la vez, es

[17] Dicho identificador, al tener asociada, según el modelo, una distribución de probabilidad, permite distinguir los valores extremos de los considerados más comunes. En los modelos de regresión de nivel múltiple (Goldstein, 1995) nos encontramos con residuos asociados a cada una de las variables o coeficientes aleatorios introducidos en el modelo. El modelo que hemos estimado para nuestros datos contempla la existencia de residuos a nivel individuo (tantos como individuos) y residuos a nivel centro (tantos como centros). Cada residuo no es más que una realización (o predicción de realización) de una variable aleatoria. El residuo correspondiente a cada centro (*efecto centro*) se estima a partir de los residuos individuales, el número de alumnos y el valor del coeficiente de correlación intra-centros. Se trata de *estimadores encogidos* (subestiman la información sobre un subgrupo a favor de la información sobre todo el grupo cuando el número de elementos del subgrupo es muy pequeño y puede conducir a estimaciones poco eficientes). Una vez identificado cada centro con su *efecto* podemos establecer una banda de efectos no distinguibles (que se separan como máximo una desviación estándar de la media) y considerar aquellos centros que presentan un efecto superior a 0.42 o inferior a -0.42.

[18] Los centros que impartirán el bachillerato LOGSE (la nueva enseñanza secundaria postobligatoria que emana de la LOGSE) proceden de institutos de bachillerato, institutos de formación profesional o centros de enseñanza secundaria creados *ex profeso*. Cabe esperar, pues, diferencias importantes entre los centros, no solamente en cuanto al conjunto de materias sino también en cuanto al estilo de la enseñanza que impartirán.

la encargada de informar a los centros sobre los resultados de sus alumnos. La información que suministra se concreta en la nota media de COU, la nota media de PAAU y la diferencia entre ambas para los alumnos de su centro y la media del conjunto de alumnos de la convocatoria. Estos son, en la actualidad unos de los pocos indicadores[19] que maneja nuestra comunidad educativa. Como tales han sido utilizados desde la Administración y desde cada uno de los centros para realizar comparaciones. Nos preguntamos sobre la "oportunidad" de tales comparaciones[20] dado que en muchos casos las diferencias observadas no sean significativas. Es más, debido a la diversidad existente en cuanto al número de alumnos[21] que cada centro presenta a las PAAU, algunas comparaciones carecen incluso de sentido.

Para abordar esta cuestión decidimos estudiar la estructura de covarianza entre la *nota COU* y la *nota PAAU* descomponiendo la variación global en variación entre centros y variación dentro de los centros. La metodología seguida consiste en modelizar de manera conjunta (bivariante) la variación de la *nota COU* y la *nota PAAU* de cada estudiante, descomponiendo la variación de cada nota en efecto debido al centro y efecto específico del estudiante y admitiendo la posible covarianza entre los efectos de un mismo nivel. Hemos aplicado a los datos de la muestra del apartado 2.2 un modelo de componentes de la varianza para el estudio de datos multivariantes con estructura jerárquica[22]. Los resultados[23] de la estimación de la variación entre centros, matriz $\sum_C$, y dentro de los centros entre estudiantes, matriz $\sum_E$, son:

$$\sum_C = \begin{pmatrix} 0.022 & 0 \\ 0 & 0.201 \end{pmatrix} \qquad \sum_E = \begin{pmatrix} 0.662 & 0.582 \\ 0.582 & 1033 \end{pmatrix}$$

Conclusiones

Una primera conclusión que se desprende del tratamiento estadístico señala que la media global de la *nota COU* y de la *nota PAAU* es, respectivamente, 6.74 y 5.24. La descomposición de la variación total en variación entre centros y dentro de los centros está justificada. La varianza entre centros es significativa para ambas

[19] En este sentido, hay que celebrar la labor que está realizando el INCE en su "Proyecto de sistema estatal de indicadores de la educación". Otra referencia imprescindible es el documento de la OCDE "Education at a glance. OECD indicators" de 1995.

[20] Estos datos no se hacen públicos en Cataluña -tampoco nos consta que se haga en el resto de España- hasta el momento. En otros paises, como Inglaterra y el Pais de Gales en que tradicionalmente se publican estas informaciones, y a la vista del uso no siempre adecuado que se ha hecho de estos indicadores, los científicos (ver artículo de Goldstein y Spiegelhalter, 1996) se han visto obligados a recordar la incertidumbre inherente a este tipo de medidas y la necesidad de contextualizarlas.

[21] Los centros con un número reducido de estudiantes son más susceptibles de verse afectados en sus medias por valores excepcionales. El tamaño del centro es una característica que se debería tener en cuenta al comparar centros.

[22] Para más detalle sobre este tipo de modelos ver Goldstein, 1987.

[23] La diferencia con el modelo del apartado anterior radica en que, en aquel caso, tomábamos una única respuesta, la *nota PAAU*, y estudiábamos su variación previo ajuste con la *nota COU*. Ahora consideramos ambas notas como respuestas y estudiamos su variación conjunta. Los detalles sobre este enfoque bivariante de nivel múltiple y su aplicación a la muestra de 26 centros se encuentran en la tesis que A. Cuxart está elaborando.

notas. Los resultados de estas estimaciones corroboran lo que a nivel descriptivo ya se apuntaba anteriormente: una mayor variabilidad entre *efectos centro* en la *nota PAAU* que en la *nota COU*. Mientras el *efecto centro* en la *nota PAAU* toma valores en un rango que va de -0.9 a 0.9[24], en la *nota COU* varía entre -0.3 y 0.3 . En el Gráfico 3 puede observarse como el rango de variación de la nota media de PAAU es mucho mayor (más del doble) que el de la nota media de COU. Tanto a nivel individuo como a nivel centro, la *nota PAAU* presenta más variabilidad[25] que la *nota COU*. No solamente la *nota PAAU* 'separa' mejor los centros sino que dentro de un mismo centro la *nota PAAU* introduce más discriminación que la *nota COU*.

Una segunda conclusión igualmente importante es que a nivel centro la covarianza no es significativa. La covarianza nula entre el efecto centro en la nota COU y el correspondiente en la nota PAAU nos diría que el hecho de tener una media alta de la nota COU no siempre va acompañado de una media alta también en la nota PAAU. En cambio, a nivel individuo, la correlación entre efectos es 0.704 (superior a 0.66, la correlación obtenida entre las dos notas a nivel global). Existe mayor coherencia entre ambas puntuaciones dentro de cada centro que si consideramos todos los alumnos de la muestra. Este hecho tiene una clara explicación: cada centro ha ordenado a los alumnos que superan el COU a través de sus puntuaciones y el resultado son distribuciones de notas COU con una media similar de un centro escolar a otro. En PAAU se realiza una nueva ordenación. Aunque a nivel interno de cada centro exista una coherencia entre ambas ordenaciones, parte de la misma se pierde al agrupar todos los centros ya que la escala o baremo utilizado por cada uno de los centros en COU no es exactamente la misma.

Por último, destacamos hasta qué punto es fiable la comparación de centros a través sus respectivas notas medias de COU y PAAU. A este respecto y a tenor de nuestros resultados podemos decir que:

- No tiene sentido comparar centros por sus medias en COU[26] (en el caso de que dos centros se diferenciaran en una cantidad apreciable, esta diferencia solamente sería *observable* si el tamaño de los centros fuera superior a 197.
- Para centros de tamaño inferior a 30 alumnos tampoco se recomienda establecer ordenaciones a partir de las medias de PAAU.
- En cambio, la ordenación 'más informativa', la que puede ser utilizada para tamaños incluso de 16 alumnos, es la diferencia entre las dos medias. Esta ordenación es la que puede resultar más útil para la Inspección educativa en el

---

[24] $0.90 = 2DE = 2\sqrt{0.201}$

[25] La similitud de medias de COU entre centros y dentro de los centros es una consecuencia, en parte, del reducido rango de valores con que se puntua cada asignatura: 5.5, 6.5, 7.5 y 9. Como decíamos en el apartado anterior, la nota PAAU evidencia que los centros estan puntuando con escalas diferentes en COU pero dando como resultado distribuciones de notas similares. En el bachillerato LOGSE se ha corregido esta deficiencia y la puntuación es más 'fina'.

[26] En el gráfico 3 observamos la similitud en las medias de COU de los centros de la muestra. Cabe observar que los centros públicos ocupan posiciones centrales en cuanto a la media que obtienen en las PAAU. Estos resultados coinciden con los obtenidos por Muñoz-Repiso *y otros* (1997) en un estudio relativo a los centros adscritoa a la Universidad Autónoma de Madrid.

estudio de los casos que se separan considerablemente del comportamiento promedio.

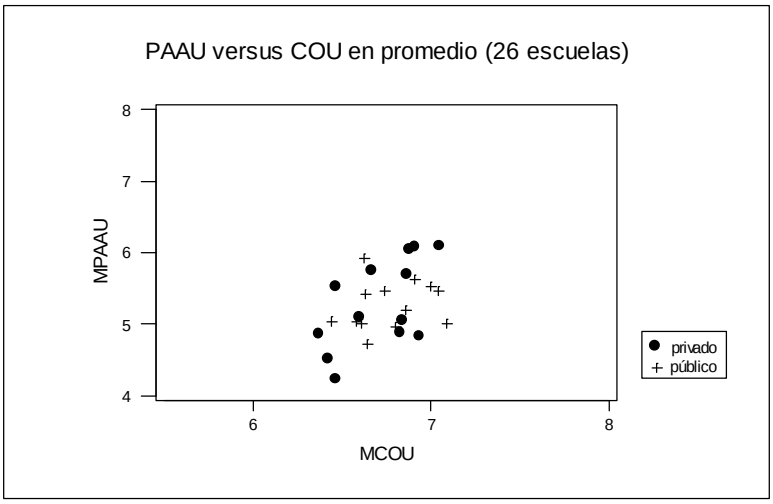


Gráfico 3. Medias en COU y en las PAAU de los 26 centros.

3. La influencia del profesor-corrector. Análisis del proceso de corrección de las pruebas PAAU[27]

3.1 Estudios anteriores y motivación

En educación - en general en todo el ámbito de las ciencias sociales- las variables suelen ser difíciles de medir. Por ejemplo, si pretendemos evaluar la *habilidad en matemáticas* de una población de estudiantes, tendremos que definir previamente, qué entendemos por *habilidad en matemáticas*, qué tipo de prueba prepararemos para provocar que se manifieste tal *habilidad* -si se trata de un examen, si va a ser de preguntas abiertas o cerradas, oral o escrito,...- qué respuestas esperamos obtener y cuáles de ellas daremos por válidas, cómo se administrará la prueba, cómo se puntuará, y por fin una vez tengamos la puntuación final qué interpretación corresponderá a los posibles resultados. Cada uno de los elementos que integran el proceso de medida conlleva arbitrariedad e incertidumbre. Ante esta situación, digamos de imperfección del instrumento de medida, y puesto que erradicarla es imposible, lo aconsejable es avanzar en el conocimiento de sus causas para limitar al máximo su impacto.

Sans (1991) entre otras consideraciones apunta la necesidad de medir la fiabilidad del proceso de corrección de las PAAU a la vista de las diferencias observadas entre tribunales. Muñoz-Repiso *y otros* (1991) también abordan el tema de la corrección y, por ejemplo, al intentar explicar porqué los estudiantes de Ciencias (opciones A y B) sufren una disminución de nota -de la nota de expediente a la nota PAAU- superior a sus compañeros de Letras (opciones C y D)

[27] Este apartado ha sido redactado en base a una investigación ya presentada que dirige el Sr. Manuel Martí Recober y que está financiada en parte por el CIDE (Proyectos de investigación, 1995).

sugieren que una de las razones podría ser que las asignaturas específicas del área de Ciencias permiten una mayor discriminación entre los alumnos que las asignaturas específicas del área de Letras, a la vez que en las asignaturas comunes a todas las opciones -como Filosofía o Comentario de Texto- se tienda a discriminar poco otorgando puntuaciones que en su mayoría van del 4 al 7. Ante la instauración del tribunal único para el curso 1991-92, Muñoz-Repiso *y otros* (1991) planteaban la necesidad de criterios de homologación y garantías de mayor objetividad en la corrección de las pruebas.

Para medir la fiabilidad de la corrección es necesario disponer de algún tipo de réplicas. Escudero y Bueno (1994) realizan un experimento con un tribunal paralelo que evalúa los exámenes puntuados a su vez por el tribunal oficial correspondiente. Al comparar las puntuaciones que otorgan los dos tribunales, no encuentran diferencias significativas entre las medias de la nota final de PAAU (tampoco son significativas las diferencias entre la mayor parte de notas agregadas[28]). Desde nuestro punto de vista uno de los resultados más relevantes del trabajo de Escudero y Bueno (1994) es la constatación de que si se tomaran las puntuaciones de la segunda corrección (experimental) en lugar de las oficiales, para un 10% de los estudiantes se hubiera invertido la resolución (obtener o no un aprobado en la nota de acceso a la Universidad).

Aunque de manera global los resultados de un tribunal no sean significativamente diferentes de los de otro tribunal, ésta, creemos, no es una razón suficiente para pensar que el proceso en sí es "justo". Deberían ser los expertos los que se pronunciaran sobre este tema. Y a partir de sus observaciones, es una labor de política educativa el definir qué se entiende por una "diferencia aceptable". Además, los datos de interés en educación deberían considerarse a nivel de individuo, no solamente a nivel de agrupación de individuos, sea tribunal o centro escolar.

En junio de 1991, Albert Satorra y Frederic Udina de la UPF, llevaron a cabo un experimento[29] de control de la calidad en la corrección de los exámenes de Matemáticas I de las PAAU. De los resultados se pudo estimar que la varianza inducida por la corrección en la nota de Matemáticas I era del 10%. Este estudio no podía ser considerado como concluyente, si no más bien una invitación a la reflexión -insistían sus autores-, dado el carácter voluntario de las respuestas.

En la línea de los trabajos realizados a nivel estatal, era necesario, pues, abrir una vía de investigación en un tema insuficientemente estudiado[30] en nuestro país:

[28] En el trabajo de Escudero y Bueno no se comparan los resultados de la doble corrección en las pruebas específicas por asignaturas. Se estudian los resultados a nivel global de la primera obligatoria, segunda obligatoria, primera optativa y segunda optativa.

[29] A los 73 correctores de dicha asignatura de la ciudad de Barcelona se enviaron por correo dos exámenes fotocopiados (al azar se escogió 20 exámenes de uno de los tribunales y se fotocopiaron antes de ser corregidos oficialmente) pidiéndoles que los corrigieran con el mismo criterio que días antes habían aplicado en la corrección oficial. De los 73, respondieron 39.

[30] Lo cierto es que a pesar del valioso estudio realizado por el Consejo de Universidades en 1993 y de las interesantes recomendaciones que en el mismo ya se incluían, poco se ha avanzado en estos cuatro años para implementarlas. Una de las recomendaciones era la de utilizar un mayor número

la necesidad de medir eficientemente la variabilidad de la corrección en cada una de las pruebas así como indagar sobre las componentes de dicha variabilidad.

En el trabajo que hemos resumido en la sección anterior se constató la existencia de un efecto debido al centro -centro en que cursó el COU- en la *nota PAAU* de cada estudiante. El efecto debido al centro también aparecía al analizar la asociación entre la nota PAAU de cada materia y la correspondiente nota en COU. El paso siguiente consistió en analizar[31] las posibles causas de dicho efecto. Por cuestiones organizativas, en Cataluña un corrector corrige los exámenes de todos los alumnos de un mismo centro. Luego, el efecto centro podía estar confundido con el efecto debido al corrector. Aplicando modelos de regresión de nivel múltiple pudimos separar (Cuxart, 1995) el efecto centro del efecto corrector en la materia de Matemáticas I, estimando una varianza significativa entre centros del mismo orden que la varianza entre correctores[32]. Cabe destacar que la magnitud de la incidencia del efecto centro en la nota PAAU resultó ser muy próxima a la obtenida por Satorra y Udina (1991). Pero, el hecho de que cada examen fuera corregido por una sola persona no nos permitía investigar sobre los posibles grados de severidad de los correctores ni tampoco abordar temas de fiabilidad en la corrección.

Con la intención de poder estudiar con más profundidad el proceso de corrección en la calificación de los estudiantes que se presentan a las PAAU, emprendimos el diseño de un experimento que permitiera evaluar la calidad de la corrección en dos materias, Matemáticas I y Filosofía (consideradas de diferente dificultad en la concreción y aplicación de criterios de corrección). El objetivo principal del estudio era la obtención de medidas de la calidad de la corrección que nos permitieran iniciar un proceso de seguimiento y control del sistema en posteriores convocatorias. Un segundo objetivo, ligado al anterior y que adquiere sentido en función de éste, era la detección de las posibles fuentes o factores de la variabilidad de la corrección.

En evaluación educativa en general y en las pruebas PAAU en particular, interesa que el examen sea válido[33], es decir mida aquello que ha de medir y para lo que ha sido construido y que la puntuación que otorga el proceso de corrección sea fiable. La fiabilidad tiene sentido en un contexto de réplicas y se refiere a la precisión del instrumento de medida. La puntuación será más fiable o precisa cuanto menor sea el error de medida introducido en el proceso de corrección.

---

de preguntas con la intención de abarcar mejor el programa y así incrementar la fiabilidad de las pruebas.

[31] Es evidente que para estudiar el efecto centro se debería investigar en el proceso de evaluación en las escuelas. Este sería tema de otra investigación muy interesante, por cierto. Nuestra esfuerzos se centraron en conocer el funcionamiento del instrumento de medida de la nota PAAU, es decir, el proceso de corrección.

[32] Los detalles de este trabajo se encuentran en un informe preparatorio.

[33] No abordamos el tema de la validación del examen. Nos centraremos en el proceso de corrección y en su fiabilidad. No obstante, como se verá más adelante, al analizar las causas de las discrepancias observadas entre correctores, se apunta la posibilidad de que los exámenes propuestos no estén midiendo adecuadamente la preparación de los alumnos. De ser así, tendríamos que una insuficiente validación del examen conllevaría al mismo tiempo una fuente de discrepancia en la corrección, añadiendo más elementos de injusticia al proceso.

En el estudio de Satorra y Udina (1994) la selección de correctores no fué aleatoria. La corrección se llevó a cabo con posterioridad a las pruebas y sin la presión del volumen de exámenes a corregir. El modelo de componentes de la varianza de estos autores no distinguía entre posibles fuentes de error en la estimación. El trabajo de Escudero y Bueno (1994), en el cual se respetaron las condiciones de realización que acabamos de citar, involucraba pocos correctores. Cada tribunal, en aquellos momentos, solía tener solamente un corrector para cada asignatura. Al comparar dos tribunales se estaban comparando dos correctores.

Las características de nuestro diseño pueden resumirse en:

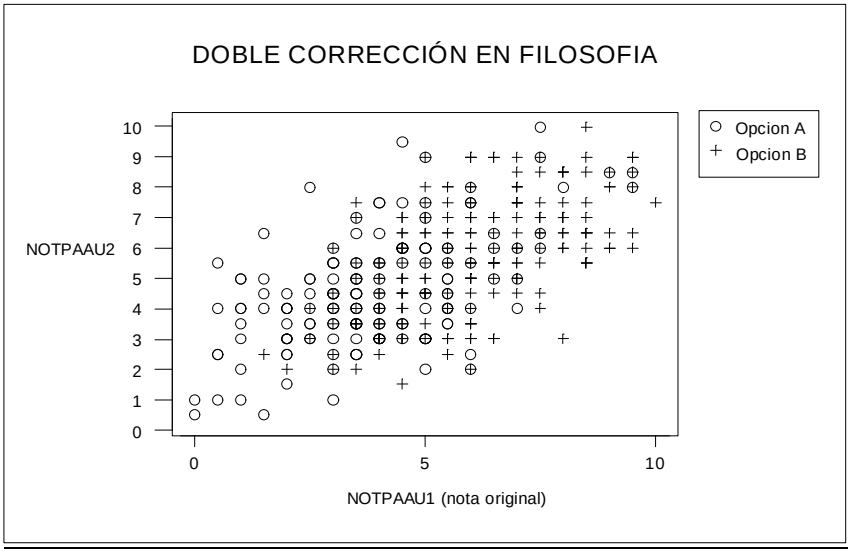
- Se hizo una corrección doble. Cada examen debía ser corregido dos veces, una sería la corrección oficial y la otra la realizaría un corrector adscrito a otro tribunal de las PAAU. Esta, llamémosle, segunda corrección -aunque se hiciera al mismo tiempo y sin conocimiento de la corrección oficial- se haría a partir de una fotocopia[34].
- Se asignó el segundo corrector al azar, no de manera voluntaria.
- Se corrigieron los exámenes en las fechas oficiales, no con posterioridad.
- Participaron un número importante de estudiantes (187 en Matemáticas y 363 en Filosofía) permitiendo un abanico de situaciones suficientemente amplio. Intervinieron un número también grande de correctores (10 en Matemáticas y 20 en Filosofía, adscritos a los 18 tribunales de Barcelona, junio de 1995) lo cual garantizó una cierta representatividad de la muestra respecto a la población de correctores.
- Los correctores desconocían el uso que se haría de la puntuación (según la información que se dió a los correctores, éstos no podían saber si su nota sería la oficial o si sería utilizada solamente con finalidades estadísticas).

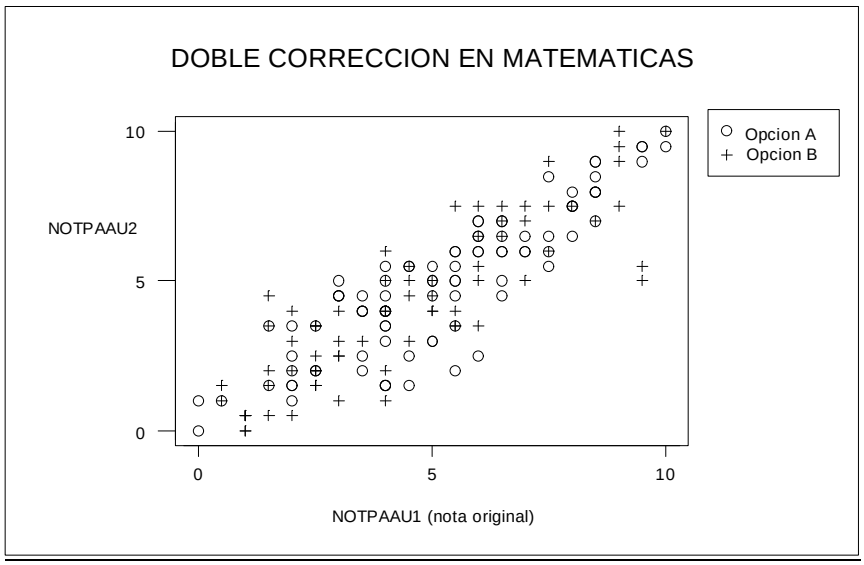
En los Gráficos 4 y 5 se representa la puntuación que obtuvo cada examen según el corrector oficial y según el segundo corrector (NOTPAAU1 y NOTPAAU2, respectivamente). Un símbolo especial informa de la modalidad (opción A o B) escogida por el alumno . Dichos gráficos muestran una discrepancia entre correctores considerable, mayor en Filosofía que en Matemáticas. Al distinguir por opciones, parece ser que en Filosofía existe mayor concentración de puntos de la opción A en el extremo inferior izquierdo mientras que la opción B predomina en el extremo superior derecho. En general, los exámenes de la opción B de Filosofía han obtenido, tanto en la primera como en la segunda corrección, notas superiores a los de la opción A. En Matemáticas no se aprecia, a primera vista, una tendencia tan clara como en Filosofía.

La Tabla 2 nos ofrece, para cada asignatura por separado, la correlación entre la corrección oficial, la segunda corrección y la correspondiente puntuación en COU de cada alumno, confirmando la dispersión manifestada en los Gráficos 4

[34] Se fotocopiaron todos los exámenes de estas asignaturas de dos de los tribunales. Se repartieron las fotocopias aleatoriamente entre el resto de correctores. Los correctores de las fotocopias recibieron en el sobre que contenía los exámenes de su tribunal veinte fotocopias, aproximadamente, de exámenes sin corregir junto con una carta en que se les pedía que corrigieran estos veinte exámenes con los mismos criterios y al mismo tiempo que el resto de exámenes.

y 5. Es natural que la segunda corrección correlacione con COU más débilmente que la corrección oficial. Se trata de un efecto debido al diseño: el número de correctores es mayor en la segunda corrección que en la primera.





Gráficos 4 y 5. Diagramas de la doble corrección. NOTPAAU1 es la puntuación que dió el corrector oficial y NOTPAAU2 la que dió el segundo corrector.

Tabla 2. Correlaciones Pearson entre las notas de Filosofía y entre las notas de Matemáticas

FILOSOFIA			**MATEMÁTICAS**		
	COU	NOTPAAU1		COU	NOTPAAU1
NOTPAAU1	0.316		NOTPAAU1	0.614	
NOTPAAU2	0.311	0.600	NOTPAAU2	0.572	0.874

Según la Tabla 3, el grado de concordancia entre correcciones no es muy alto[35]. Como era de esperar, se observa más coincidencia entre correctores en Matemáticas que en Filosofía. No deja de sorprendernos, sin embargo, el 28% de casos en que la diferencia entre las dos correcciones de Matemáticas supera la unidad. Otro dato preocupante es el 13% de casos en Filosofía con una discrepancia superior a los tres puntos.

Tabla 3. Frecuencias de valores de la variable diferencia (en valor absoluto) entre las dos correcciones. Para 185 estudiantes la nota de Filosofía otorgada por el primer corrector difiere en menos de un punto de la nota que le habría otorgado el segundo.

	Dif ≤ 1	1 < Dif ≤ 3	3 < Dif	total de estudiantes
Filosofía	185 (51.0%)	130 (35.8%)	48 (13.2%)	363
Matemáticas	134 (71.7%)	49 (26.2%)	4 (2.1%)	187

Como decíamos anteriormente, estudios ya publicados en los que se planteaba la comparación entre la corrección de dos tribunales se basaban en análisis de la varianza. Hemos realizado tales análisis con nuestros datos y podemos decir que, al igual que en el estudio de Escudero y Bueno (1994)[36], no ha resultado ser significativa la diferencia entre la primera y la segunda corrección, ni en el caso de Filosofía ni en el de Matemáticas. Ahora bien, aunque, en promedio la diferencia entre correcciones no sea estadísticamente significativa, puede muy bien ocurrir que, para un número importante de estudiantes, el hecho de tener un corrector u otro modifique sus posibilidades futuras (sabemos que el acceso a algunos estudios universitarios depende de décimas). Naturalmente, hay que empezar por saber cual es la variabilidad en que nos movemos y a partir de ahí definir el plan de trabajo: especificación de objetivos, oportuno seguimiento, control e intervención.

Si, antes, decíamos que al considerar la primera y la segunda corrección no se podía concluir que las medias fueran significativamente diferentes, al introducir el factor opción de examen ello cambia. La opción de examen es un factor diferenciador en el sentido que la media de notas que obtienen los estudiantes de Filosofía que escogen la opción A es significativamente diferente de la media de notas que obtienen aquellos que escogen la opción B. Este hecho se da tanto en la corrección original como en la segunda corrección y en las dos materias, si bien con un grado de significación más alto en Filosofía que en Matemáticas. Estos análisis corroboran lo que a nivel descriptivo apuntaban los gráficos de doble corrección

[35] Cabe destacar que en aquellos momentos, junio de 1995, ya se había realizado un esfuerzo considerable para adecuar los programas de COU y modificar los formatos de examen y criterios de corrección en aras de una mayor objetividad.

[36] Cabe recordar que en el trabajo de Escudero y Bueno (1994) la nota de Matemáticas no se estudiaba propiamente ya que aparecía mezclada con el resto de primeras obligatorias de opción. Sí se estudiaba la nota de Filosofía.

3.2 Descomposición de la variación observada

Con el objetivo de profundizar un poco más en el estudio de la discrepancia observada decidimos modelizar nuestros datos, descomponiendo el error de medida introducido en la corrección en sus diferentes fuentes de variación. Nuestro enfoque se enmarca en la teoría de la generalizabilidad (Cronbach, 1972) y en la adaptación que Longford (1994 y 1995, Chap.2) propuso para el estudio de datos relativos a exámenes y correctores. Al intentar explicar porqué dos correctores al corregir el mismo examen dan puntuaciones diferentes podríamos distinguir entre dos posibles fuentes de discrepancia (Longford, 1994): la *severidad* y la *inconsistencia*.

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 sobras es sabido que la discrepancia no se debe solamente a los diferentes grados de severidad. Un mismo examen al ser corregido por un corrector puede obtener una puntuación diferente si se trata de uno de los primeros exámenes que corrige o si el corrector ya lleva corregidos un buen número de ellos. El cansancio puede influir en la agudeza y en la atención. También el hecho de haber visto el contenido de muchos exámenes puede modificar[37] el criterio haciéndolo, a partir de un cierto momento, más indulgente o más exigente que al principio. Esta segunda fuente de error, que engloba una serie de imperfecciones presentes en el proceso de corrección, la llamaremos *inconsistencia* o "error no sistemático". 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". El modelo concreto de componentes de la varianza que proponemos para explicar la variación de la puntuación de un examen es el modelo aditivo (Longford 1994):

$$y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \qquad \text{(MDV)}$$

siendo $i = 1, 2,..., I$ el índice del examen o estudiante; $j=1, 2,..., J$ el del corrector. El número de puntuaciones que entran en el estudio es $2I$; y_{ij} es la puntuación que el corrector j ha dado al examen i ; α_i es la puntuación *verdadera*

[37] Uno de los hechos observados es el de la adaptación del corrector al grupo de exámenes. Parece ser que algunos profesores distribuyen sus alumnos como harían en su propia aula o grupo-clase sin tener en cuenta que deben aplicar unos criterios universales y prescindir del particular grupo de estudiantes que están corrigiendo. Este fenómeno de adaptación genera injustícias. En función del conjunto de centros que van a parar a un mismo tribunal puede repercutir en una ventatja o inconveniente para cada alumno en particular.
Este fenómeno se limitaría si: 1) los correctores no fueran adscritos a tribunales, separando vigilancia de corrección; 2) cada corrector recibiera un bloque aleatorizado de exámenes, con desconocimiento de las escuelas de procedencia; 3) se repartieran normas consensuadas de corrección de cada examen.
Sobre las normas de corrección caldría distinguir entre las generales, elaboradas a priori y aplicables a cualquier examen, y las específicas, elaboradas por el equipo que propone los enunciados de examen en el momento de su confección y revisadas, por este mismo equipo, a partir de la corrección de una muestra aleatoria de los exámenes una vez que se dispone de los mismos.

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 seria 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.

3.3 Estimación

El método que hemos seguido para estimar el modelo del apartado anterior es el de los momentos, una adaptación de los clásicos estimadores de la varianza (ADEVA) al caso de datos *no balanceados* y para más de un factor. Una vez determinadas las fórmulas algebraicas[38] para los estimadores hemos utilizado el software estadístico MLn (Goldstein, 1986a) para la confección de los programas que calculan las estimaciones y permiten la revisión de las hipótesis del modelo.

Tabla 4. Estimaciones de las componentes de la varianza de la puntuación observada . Entre paréntesis la proporción de varianza respecto de la varianza total.

Componentes de la varianza total	**Matemáticas**	**Filosofía**
$\hat{\sigma}_a^2$, var. entre *notas verdaderas* α_i	5.350(86.5%)	2.475(60.2%)
$\hat{\sigma}_b^2$, var. de la *severidad* β_j	0.011 (0.2%)	0.248(6.0%)
$\hat{\sigma}_e^2$, var. de la *inconsistencia* ε_{ij}	0.827 (13.3%)	1.386(33.7%)
varianza total (suma) estimada	6.188	4.109
varianza muestral	*6.189*	*4.065*

En una primera comparación de las varianzas de las dos asignaturas, constatamos como en Filosofía se da una concentración de notas alrededor de su media mucho mayor que en Matemáticas. Este hecho se da tanto en la varianza total como en la varianza debida a la nota verdadera. La prueba de Filosofía en las PAAU discrimina menos que la de Matemáticas I. En cambio, la varianza de estas dos asignaturas en COU es muy similar (1.46 en Matemáticas frente a 1.36 en Filosofía)[39].

[38] Los detalles técnicos se encuentran en la tesis que A. Cuxart está elaborando.

[39] No podemos olvidar que las puntuaciones en las asignaturas de COU son en cuatro categorías (5.5, 6.5, 7.5 o 9) , hecho que provoca una excesiva similitud entre estudiantes y entre distribuciones por asignaturas en COU.

Resultados

- La primera conclusión a que llegamos es que la calidad de la corrección es baja. Para una tercera parte de los exámenes de Filosofía y un 18% de Matemáticas I, la diferencia entre las dos correcciones resultó ser superior o igual a dos puntos. Los indicadores de la calidad obtenidos a partir de la modelización de los datos -léase, varianzas de la *nota verdadera*, de la *severidad* y de la *inconsistencia*, así como correlaciones- confirman las primeras observaciones derivadas de la simple comparación de puntuaciones.

- En las dos asignaturas analizadas la principal fuente de error en la corrección es la *inconsistencia*. La varianza debida a la *inconsistencia* representa un 13% de la varianza total en Matemáticas y un 34% en Filosofía. En Matemáticas no se aprecia una diferencia entre la severidad de los correctores, en cambio en Filosofía parecen coexistir diferentes grados de severidad entre correctores -la varianza de la *severidad* en Filosofía representa un 6% de la varianza total. El Gráfico 6 ilustra la participación porcentual de cada fuente de error en la variación total.

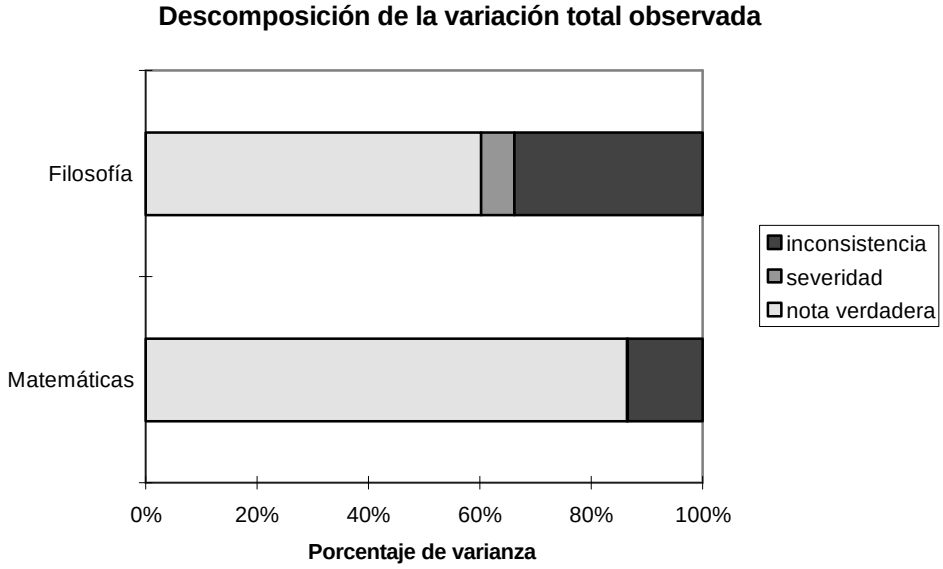


Gráfico 6. Descomposición de la varianza total de la nota observada en Filosofía y en Matemáticas I, según aplicación del modelo (MDV)

- El análisis de la *inconsistencia* apunta la posibilidad de que ciertos exámenes o preguntas conlleven mayor probabilidad de discrepancia entre correctores. El hecho más remarcable, que sustenta nuestra conjetura, es la evidencia que en Filosofía una de las opciones ha generado más discrepancia que la otra y, además, en sentido inverso, de tal manera que, la diferencia entre las puntuaciones del primer y segundo corrector en promedio es más del doble en la opción A (-0.714) que en la B (0.33). Los diagramas de caja (*box-plot*) del Gráfico 7 ilustran esta situación. En ellos, la línea que divide cada caja se sitúa en la diferencia mediana . El diagrama relativo a la opción B se desplaza hacia

valores más positivos, evidenciando que para esta opción la corrección oficial tendió a ser superior a la segunda corrección, mientras que para los exámenes de la opción A fue la segunda corrección la que tendió a ser superior. Una de las razones de esta situación -según se desprende de un estudio más pormenorizado sobre los exámenes de los alumnos y las anotaciones de los correctores- parece ser el comportamiento diferenciado de los dos correctores oficiales que participaron en el estudio. Parte de la diferencia observada entre tribunales se debe, pues, al comportamiento diferenciado de sus respectivos correctores (oficiales).

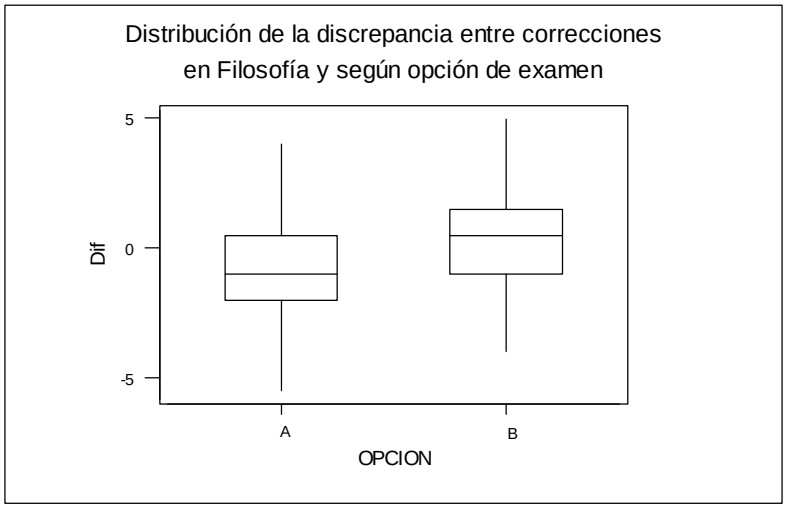


Gráfico 7. Discrepancia entre correcciones en Filosofía y opción de examen.
Dif = corrección oficial -segunda corrección

- En cuanto a los efectos en el acceso a la Universidad, aproximadamente un 3% de los estudiantes de la muestra, 11 sobre un total de 362, habrían sido ubicados de manera diferente (obtener o no un aprobado en la nota de acceso) si en lugar de tener los correctores oficiales hubieran tenido los segundos correctores.

En condiciones normales solamente se dispone de una corrección por examen. Es de destacar la mejora que se introduciría, sobretodo en Filosofía, si para cada examen se pudiera contar con dos correcciones y tomar como nota definitiva la media de ambas. Según se desprende de la Tabla 5, el valor del coeficiente de correlación entre la puntuación observada y la *nota verdadera*, se vería incrementado en un 12% al tomar la media entre las dos correcciones (de 0.78 a 0.87). En Matemáticas tan sólo representaría una mejora del 3%.

Tabla 5. Indicadores de la calidad de la corrección, calculados a partir de las estimaciones que se derivan de la aplicación del modelo (MDV)

Coeficientes de correlación	**Matemáticas**	**Filosofía**
r (entre las dos puntuaciones)	0.86	0.60
r_{a1} (entre *nota verdadera* y puntuación observada)	0.93	0.78
r_{a2} (entre *nota verdadera* y puntuación media)	0.96	0.87

A pesar que los resultados de este estudio no contradicen la intuición, no pueden generalizarse a otras asignaturas, ni a anteriores o futuras convocatorias, ni tampoco a diferentes formatos de examen.

3.4 Algunas reflexiones y propuestas

El Consejo de Universidades en su documento del año 1993: " Las Pruebas de Aptitud para el acceso a la Universidad: Problemática actual y propuesta de modificación", concretamente en la propuesta 3.3 relativa a la "Realización de una doble corrección del Comentario de Texto", señala:

> *"Dadas la características especiales de este examen, sería conveniente utilizar para el Comentario de Texto una doble corrección. La doble corrección es utilizada en distintos tipos de prueba y ejercicios como garantía de ponderación y equilibrio en la calificación final que se otorga.*
> .....
> *En el caso de la prueba de Comentario de Texto, que no se refiere a una materia específica, sino que tiene un carácter general, y dado que los profesores que han de corregirlo, al ser de cualesquiera de las otras materia que integran la prueba, pueden tener apreciaciones y criterios dispares, se hace más aconsejable el uso de este procedimiento.*
> *La calificación final será la media de la otorgada por ambos correctores, siempre que no se diferencies en más de 2 puntos. Si la diferencia es mayor, los correctores procederían, en cada caso, a la revisión de las calificaciones efectuadas siguiendo meticulosamente los criterios de corrección específicos establecidos para el examen. La corrección de este examen se dará por concluida una vez que todas las calificaciones se hayan obtenido promediando puntuaciones con diferencias que no superen el límite establecido (Modificación legal del R.D. 406/83. /Art. 4.1.(nuevo) y de la Orden 3-9-37 /Art.nuevo).*

Teniendo en cuenta lo apuntado hasta el momento cabe plantearse, a nuestro entender, tres observaciones relevantes:

- Esta propuesta no se está teniendo en cuenta en las actuales pruebas de acceso, al menos en todas las administraciones que hemos contactado.
- Si para una materia como el Comentario de Texto se considera no admisible una diferencia superior a 2 puntos entre las dos correcciones, ¿cual sería -nos

preguntamos- la diferencia que se podría admitir en Filosofía o en Matemáticas en que se considera que los correctores son expertos en la materia?

- Según los datos del presente estudio[40], en Matemáticas 14 de los exámenes (un 7%) recibieron puntuaciones que diferían en más de 2 puntos, mientras que en Filosofía fueron 77 tales exámenes (un 21%).

A la vista de estas reflexiones y en aras de incrementar la precisión en la corrección de las pruebas PAAU, presentamos, a continuación, una serie de propuestas:

a) Considerar en cada materia de examen la posibilidad de substituir el examen actual, o una parte del mismo, por una prueba con preguntas de respuesta cerrada.
b) De manera sistemática, y para los exámenes con preguntas de respuesta abierta, en cada convocatoria se debería seleccionar una muestra de exámenes de cada materia y realizar una doble corrección de los mismos con el objetivo de medir la fiabilidad o precisión en su corrección y detectar posibles fuentes de discrepancia.
c) Para aquellas asignaturas con una precisión baja realizar una doble corrección de todos los exámenes, como ya recomendaba el Consejo de Universidades en 1993.
d) Incorporar un mecanismo de revisión automático de todas las puntuaciones PAAU que, al comparar con las de la misma prueba en COU u otra variable indicativa, permita destacar aquellas puntuaciones que se separan demasiado de las previsiones. Realizar una doble corrección (si se trata de preguntas de respuesta abierta) de estos casos y introducir los ajustes que se deriven. Aquellos centros (y correctores) cuyos alumnos(exámenes) en un porcentaje alto han sufrido revisión deberían ser analizados.

Para poder realizar de manera eficaz las propuestas anteriores conviene tener una infraestructura y unos medios que lo permitan.

4. Conclusiones

Destacamos a continuación las principales conclusiones que se derivan de la aplicación de modelos de regresión con coeficientes aleatorios para el estudio de la asociación entre la nota de COU y la nota de PAAU de cada estudiante:

- Existe una variación significativa de la nota PAAU entre centros escolares. Un 20%, aproximadamente, de la variación total de la nota PAAU corresponde a variación entre centros.
- La influencia del centro escolar en la predicción de la nota PAAU individual se concreta en un término aditivo, común a todos los estudiantes del mismo centro y que hemos llamado *efecto centro*. Tales efectos tienen asociada una distribución de probabilidad y permiten identificar los centros escolares que presentan una asociación entre nota COU y nota PAAU extrema.
- Las distribuciones de la nota COU, a diferencia de la nota PAAU, varían muy poco de un centro a otro. De ahí que el coeficiente de correlación intra-centros

[40] La distribución de estos exámenes por opciones A / B fue de 10/ 4 en Matemáticas y 33/ 44 en Filosofía, respectivamente.

para la nota COU sea prácticamente 0. Este hecho sugiere que los centros están utilizando escalas de puntuación propias, diferentes de un centro a otro, según pone en evidencia el examen PAAU.

- El modelo de regresión de coeficientes aleatorios de la nota PAAU versus la nota COU que contempla género, posible repetición de COU, opción de COU y tipo de centro, nos lleva a una serie de conclusiones en cuanto al valor predictivo de estas variables coincidentes con anteriores estudios realizados a nivel estatal. La novedad de nuestro enfoque estriba en la determinación del papel predictivo de cada centro en la nota individual de PAAU de manera conjunta con el resto de variables citadas.

El análisis del proceso de corrección de las pruebas PAAU ha puesto en evidencia:

- La baja calidad de la corrección en las dos asignaturas estudiadas, incluso en Matemáticas. La principal fuente de error ha sido la *inconsistencia* (la varianza debida a la *inconsistencia* representa un 13% de la varianza total en Matemáticas y un 34% en Filosofía). En Filosofía, además parecen coexistir diferentes grados de severidad entre correctores (un 6% de la varianza total corresponde a *severidad*)
- Las consecuencias que para algunos estudiantes se pueden derivar de esta imperfección del sistema: un 3% de los estudiantes de la muestra habrían sufrido una ubicación distinta de tener en cuenta la segunda corrección en lugar de la oficial. Los estudiantes más afectados por la baja fiabilidad son, naturalmente, los que se encuentran cerca de la frontera (*borderline)* del aprobado.
- El valor que tiene el *monitorizar* una investigación conectada a la ejecución[41], tanto por la información que suministra como por la posibilidad de intervenir para realizar un *ajuste* a tiempo
- La necesidad de *interpenetrar* correctores y exámenes si queremos comparar los resultados de diferentes tribunales, centros, comarcas,...
- La conveniencia de contrastar empíricamente la dificultad de las preguntas de cada examen y materia.
- La existencia de diversas fuentes de variación en la corrección, algunas de ellas relacionadas con el diseño y contenido de los exámenes (de ahí la necesidad de mejorar el procedimiento de elaboración de los exámenes); otras relacionadas con la organización de las pruebas (por este motivo recomendamos la separación entre la labor de vigilante y la labor de corrector).
- La importancia que adquieren todos estos temas en la discusión de las nuevas PAAU. En la actualidad, el hecho de utilizar una puntuación que es una "media de medias" diluye, en gran parte, los efectos de un sistema imperfecto. Una ponderación que diera un peso mayor a algunas de las materia requiere de un mayor grado de fiabilidad en la corrección de las mismas y justifica la incorporación de pruebas de respuesta cerrada.

4.1 Perspectivas del trabajo de investigación

[41] En (Cuxart and Longford, 1996) se incluyen una serie de reflexiones y propuestas sobre el efecto de la *elección*, la posibilidad de realizar *reajustes*, de comparar resultados, y de realizar *pretests* de las preguntas para conocer su dificultad.

Los modelos de descomposición de la varianza han demostrado ser de utilidad en la investigación realizada hasta el momento y consideramos interesante ahondar en sus posibilidades. En una de las investigaciones en curso hemos abordado el estudio de la variación conjunta del vector de notas PAAU de cada estudiante con el objetivo de detectar la asociación existente entre las diversas materias; el poder discriminador del primer y segundo ejercicio y de la nota global de PAAU; así como la capacidad evaluadora de ambos ejercicios. Para empezar, las correlaciones entre las diferentes pruebas son muy débiles, incluso al calcularlas para los estudiantes de una misma opción. Es un hecho conocido que el error de medida en cada evaluación tiene el efecto de *atenuar* los coeficientes que miden la relación entre variables y que las preguntas de respuesta abierta conllevan mayor subjetividad e imprecisión en su corrección. Si, además, los formatos de examen de dos asignaturas, aunque sean propias de la opción, son muy diferentes pueden estar evaluando habilidades distintas a la vez que conocimientos. Por otro lado, los actuales exámenes no cubren de manera exhaustiva el programa de las asignaturas. De ahí que pueda hablarse de un factor *suerte* en cuanto a los temas que aparecen cada año a examen. La *suerte* de una asignatura a otra puede variar y nos encontramos con otra fuente de variabilidad. Todas estas consideraciones hacen referencia a la validación del examen y a la fiabilidad del mismo.

En el futuro creemos que sería interesante la utilización de modelos estadísticos que tengan en cuenta el error de medida. Para ello necesitamos disponer de réplicas (doble corrección, por ejemplo) para al menos una muestra de cada materia. Pensamos que los modelos LISREL nos permitirían introducir un poco más de luz en el complejo sistema de relaciones entre las materias y entre los factores que influyen en su evaluación.

Del análisis de la corrección en las dos asignaturas estudiadas por opciones (A o B) se desprende que el nivel de puntuaciones no es el mismo en las dos opciones. En general, y tanto en la primera como en la segunda corrección, la opción A de Filosofía recibió notas inferiores a la opción B, mientras que en Matemáticas ocurrió la situación inversa. Dado que los estudiantes fueron quienes eligieron la opción de examen es imposible separar el factor opción de la preparación del estudiante. Se plantea pues la necesidad de conocer la dificultad[42] de las preguntas planteadas y, a la vez, recomendar la limitación al máximo de la opcionalidad en estos exámenes.

La implantación de la LOGSE y las nuevas PAAU es una oportunidad para, a la luz de la experiencia acumulada en el sistema anterior, introducir cambios estructurales. Consideramos un hecho clave la incorporación de la investigación estadística en el seguimiento del proceso y la evaluación empírica de las modificaciones que se vayan introduciendo.

En una situación de transición como la que estamos viviendo en estos momentos, con los dos sistemas de pruebas PAAU vigentes, es interesante conocer

[42] Este curso se estan analizando los datos de un nuevo experimento de doble corrección para un número mayor de asignaturas. El objetivo de este experimento es calcular la fiabilidad de la corrección, conocer la dificultad de las preguntas y recoger la opinión de los correctores en cuanto al enunciado de examen y a los criterios de corrección específicos.

y comparar porcentajes de superación de las diferentes fases educativas en relación a la población de jóvenes de una misma edad o cohorte. De ahí también la oportunidad de investigar la aplicación de modelos de regresión logística.

4.2 Algunas consideraciones pedagógicas a la luz de las estadísticas

Entre las consideraciones pedagógicas que se desprenden del presente estudio, destacamos las siguientes:

1. Lo importante que es para todo sistema educativo el disponer de datos fiables. En este sentido las pruebas PAAU, como examen externo a los centros y estándar, se revela como un instrumento de gran utilidad.
2. La necesidad de comparar resultados con rigor y teniendo en cuenta el contexto. Si se dan estos dos requisitos, la prevención a la comparación entre centros carecerá de sentido.
3. Se debería avanzar en la "cultura" de realizar estudios que sean útiles para la Administración y que, al mismo tiempo, sirvan de referencia y contraste para los centros.
4. La ampliamente argumentada necesidad de proponer exámenes con preguntas lo más cerradas posible plantea un cambio en la pedagogía. Los profesores de secundaría deberían incorporar en su docencia este tipo de pruebas[43].
5. El examen PAAU no debería ser el primer examen global de la asignatura con que se enfrentan los estudiantes. Por ello es importante que en la secundaria hayan preparado y realizado exámenes que abarquen una parte importante de la programación, toda a ser posible.
6. Las habilidades en comunicación escrita son de gran importancia en las pruebas de respuesta abierta -hemos hablado de limitarlas en las PAAU, no de eliminarlas-. Puesto que dichas habilidades necesitan de aprendizaje, está claro que debería ser una de las prácticas prioritarias en secundaria.

5. Referencias

Aguirre de Càrcer, I. , ed. (1984b): *La Selectividad a debate.* Madrid, Universidad Autónoma.

Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies. *J. R. Statistical Society* A **149**, *Part 1, pp* 1-43.

Cochran, W.G. (1977). *Sampling Techniques.*3rd ed. Roronto: Wiley.

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, A., Graffelman, J. y Martí, M. La nota PAAU y su relación con la nota COU: un modelo de regresión con coeficientes aleatorios para el estudio del efecto centro en la nota PAAU. *Actas de la 5ª Conferencia Española de Biometría.* Valencia. 1995.

Cuxart i Jardí, A. and Longford, N.T. (1996) *Equity in the university admissions process in Spain* . Estudio -en prensa- presentado al 10th European Congress of Psychometric Society, Santiago de Compostela, julio de1997 y al 5th European Research Congress on Education, Francfurt, Alemania, set. 1997.

[43] Existen múltiples experiencias en este ámbito: Las pruebas *Canguro* de matemáticas que se realizan simultáneamente en varios paises europeos, las pruebas MIR de nuestro país, el examen SAT americano,....

Escudero, T. (1987) Investigaciones y Experiencias: Buscando una mejor selección de universitarios. *Revista de Educación*, nº 283 Ministerio de Educación y Ciencia.

Escudero, T. y Bueno García, C. Investigaciones y Experiencias: Examen de Selectividad. El estudio del tribunal paralelo. *Revista de Educación*, nº 304 Ministerio de Educación y Ciencia. 1994.

Goldstein,H. (1986a) Multilevel mixed linear model analysis using iterative generalised least squares. Biometrika, 73, 43-56.

Goldstein, H. (1987) Multilevel models in Educational and Social Research. Oxford University Press, New York.

Goldstein, H. and Spiegelhalter, D. J. (1996) 'League tables and their limitations: statistical issues in comparisons of institutional performances (with discussion)', *Journal of the Royal Statistical Society*, Ser. A, 159, 385-443.

Goldstein, H. (1995) *Multilevel Statistical Models*. 2nd ed. Kendall's Library of Statistics 3 (London, Edward Arnold).

Longford, N.T. (1993) *Random Coefficient Models* .Oxford Science Publications, Clarendon Press, Oxford

Longford, N.T. (1994) Reliability of essay rating and score adjustement. Journal of Educational and Behavioral Statistics, Vol 19, No. 3, pp 171-200

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

López, Mª del R. Algunos resultados sobre las Pruebas de Acceso a la Universidad Autònoma de Madrid en Alvarez, J.B. y Arroyo, F. (comp) (1977) Acceso a la Universidad y Marco Educativo, *Tarbiya*, número extraordinatio, Junio.

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

Memoria de actividades del Consejo de Universidades. Junio 1991- Julio 1993.

Muñoz-Repiso Izaguirre, M., et al (1991) Las calificaciones en las pruebas de aptitud para el acceso a la universidad, nº 61 *colección INVESTIGACIÓN*. Madrid: CIDE.

Muñoz-Repiso, M y otros (1997) *El sistema de acceso a la Universidad en España: tres estudios para aclarar el debate*. Madrid: CIDE.

Sans, A. (1990) *Selectivitat universitària. Anàlisi a Catalunya*. Tesi doctoral. Bellaterra: Publicaciones de la UAB.

Satorra, A. and Udina, F. (1994). Comunicación personal.