# Modelos estadísticos y evaluación: tres estudios en educación[1]

Anna Cuxart Jardí [2]

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

*Journal of Economic Literature classification:* C89, C99, I29

[1]The paper has been accepted in *Revista de Educación*. The research was partially supported by research grant DGICYT PB93-0403, DGES PB96-0300, and Concurso Nacional de Proyectos de Investigación Educativa, Spanish Ministry of Education.

[2]Department d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona, Spain

## Abstract

The educational system in Spain is undergoing a reorganization. At present, high-school graduates who want to enroll at a public university must take a set of examinations *Pruebas de Aptitud para el Acceso a la Universidad* (PAAU). A "new formula" (components, weights, type of exam,...) for university admission is been discussed. The present paper summarizes part of the research done by the author in her PhD. The context for this thesis is the evaluation of large-scale and complex systems of assessment. The main objectives were: to achieve a deep knowledge of the entire university admissions process in Spain, to discover the main sources of uncertainty and to promote empirical research in a continual improvement of the entire process. Focusing in the suitable statistical models and strategies which allow to highlight the imperfections of the system and reduce them, the paper develops, among other approaches, some applications of multilevel modeling.

# MODELOS ESTADÍSTICOS Y EVALUACIÓN: TRES ESTUDIOS[1] EN EDUCACIÓN

ANNA CUXART JARDÍ

El presente informe resume tres estudios integrados en la tesis doctoral[2] de la autora. Dicho trabajo ha consistido en el desarrollo y aplicación de técnicas estadísticas orientadas al estudio de las Pruebas de Acceso a la Universidad (PAU) en Catalunya. Se pretendía profundizar en el conocimiento del proceso, experimentar técnicas de seguimiento y obtener conclusiones, sobre la base de la investigación empírica y de la aplicación de modelos estadísticos.

El objetivo de este informe es ofrecer una visión general de la investigación llevada a cabo, destacando la metodología desarrollada -modelos estadísticos y estrategias de análisis-, así como los principales resultados y las líneas de investigación abiertas. Es también un objetivo de este informe el intentar contribuir a la divulgación de los modelos de coeficientes aleatorios -también conocidos como modelos de nivel múltiple-, en el ámbito de educación, en su calidad de instrumentos de análisis para el estudio de datos con estructura jerárquica. La metodología desarrollada, basada en gran parte, en este tipo de modelos, ha confirmado las diferencias existentes entre los estándares aplicados por los centros de secundaria en el COU. Los indicadores propuestos para evaluar el *efecto centro* son más eficientes y estables que los actuales (diferencia entre el promedio de cada centro en las PAU y en el COU), según se desprende del análisis de una muestra de centros del distrito de Catalunya a lo largo de tres años.

La modelización propuesta para el análisis y seguimiento de la calidad de la corrección ha permitido evaluar su impacto en términos de la varianza debida a la severidad y de la varianza generada por la inconsistencia. Las técnicas de revisión y diagnosis del modelo han sido especialmente útiles en la detección de "fuentes de discrepancia" entre correctores.

La aplicación de un modelo multivariante multinivel para explicar la variación conjunta de las notas del primer ejercicio de las PAU revela que dentro de los centros se dan comportamientos diferenciados, de manera que aunque algunos centros globalmente destaquen por conseguir en las PAU resultados por encima del promedio, este hecho no conlleva que en cada prueba -ni tan sólo en buena parte de ellas- hayan obtenido también resultados por encima del promedio general.

Resumiendo, en la investigación realizada se han tratado con especial atención los aspectos relativos a la validez y fiabilidad de los exámenes COU y PAU, así como dos de las principales fuentes de variación en el proceso de admisión a la universidad: los centros de secundaria y la corrección de los exámenes.

[1] Esta investigación ha sido en parte financiada por DGES PB96-0300 y Concurso Nacional de Proyectos de Investigación Educativa 1995, Ministerio de Educación.

[2] *Models estadístics en avaluació educativa: les proves d'acces a la universitat*, dirigida por Manuel Martí Recober y presentada en la Universidad Politècnica de Catalunya, en noviembre de 1998. Con motivo de la edición de un número monográfico dedicado a las Pruebas de Acceso a la Universidad en España, se publicó en esta revista (Cuxart *et al*., 1997) un avance de la investigación realizada.

## Introducción

Las pruebas PAU son una etapa clave en el proceso de transición de la enseñanza secundaria a la universidad[3]. Los exámenes de las pruebas PAU[4] se basan en las materias cursadas en el COU[5]. El estudiante concurre a las PAU una vez que ha aprobado todas las asignaturas del COU en el centro de secundaria en que las ha seguido. Las pruebas PAU son, en este sentido, una segunda evaluación de la preparación del alumno. Tratándose, en este caso, de una evaluación *externa* al centro y con un alto grado de homogeneidad: se trata de la misma prueba para todos los alumnos del distrito –en el caso de Catalunya, la misma prueba para todos los alumnos de los, aproximadamente, 400 centros (unos 25.000 alumnos en 1993).

El primero de los estudios que se presentan analiza la asociación entre las notas medias individuales de COU y de las PAU y la variación entre centros de dicha asociación.

El segundo estudio se centra en la investigación de la calidad de la corrección en las pruebas PAU. El objetivo es evaluar el impacto de los correctores y detectar los puntos débiles del proceso de corrección. Consta de un estudio empírico que permite experimentar una metodología de análisis, y de un estudio cualitativo de ámbito estatal que complementa las conclusiones y preguntas surgidas en el anterior. Esta investigación formó parte de un proyecto[6] de investigación financiado por el Ministerio de Educación. Se incluye, en este informe, un resumen de las entrevistas realizadas en el marco de dicho proyecto a los responsables de las PAU de seis distritos universitarios.

El último estudio se ha dedicado, en un enfoque multivariante, a la exploración del vector de notas PAU. En un intento de desvelar la estructura interna y el papel de cada una de las materias, se estudia la correlación y la estructura de covarianza a nivel estudiante y a nivel centro del conjunto de materias de las PAU.

La metodología seguida se ha basado en la exploración de datos y la posterior modelización estadística. Los datos[7] se han obtenido a partir del muestreo aleatorio de la población de centros de Catalunya y mediante el diseño de experimentos adecuados. En

---

[3] Para un tratamiento más profundo sobre el tema de la transición secundaria-universidad y en relación a la Reforma de las pruebas, véase el artículo de Martí *et al.* (1997). El desarrollo de la LOGSE: las nuevas Pruebas de Acceso a la Universidad. *Revista de Educación,* 314, pp. 89-114.

[4] Para superar las pruebas el estudiante ha de obtener una nota de acceso superior a 5 puntos. La nota de acceso es la media aritmética entre la *nota Expediente* (media aritmética de cuatro notas globales correspondientes a los tres cursos de bachillerato y al COU) y la *nota PAU* (media ponderada de las puntuaciones de ocho pruebas, nueve en las comunidades autónomas con lengua propia), calculándose dicho promedio siempre que la *nota PAU* no sea inferior a 4.

[5] La investigación que presentamos se refiere a estudios realizados entre 1994 y 1998 para las promociones del COU. Los resultados de la investigación puedan dar luz sobre temas también de interés en el sistema educativo LOGSE. De hecho, en 1997 iniciamos la aplicación de la modelización que aquí presentamos a las primeras promociones del bachillerato LOGSE que realizaban las pruebas de acceso a la universidad. En este curso 1999-2000 se examinará en Catalunya la primera promoción al completo que ha seguido el nuevo bachillerato (junto con un reducido número de estudiantes que habrán repetido el COU). Los resultados sobre estas primeras prmociones del bachillerato LOGSE serán motivo de una publicación específica.

[6] *Los sistemas de corrección de las pruebas de Selectividad en España. Análisis y propuestas.* Proyecto de Investigación Educativa. Convocatoria 1995. BOE 13-06-1995. Memoria final presentada en noviembre de 1998.

[7] Datos obtenidos gracias a la colaboración de la *Oficina de Coordinació del COU i les PAAU de Catalunya.*

consecuencia, las conclusiones que se derivan de la investigación empírica deben limitar su alcance al distrito de Catalunya .

En la fase de análisis de datos el software estadístico utilizado han sido las aplicaciones estadísticas MINITAB y SPAD. En la fase de estimación de modelos y diagnosis se ha combinado la utilización de la aplicación Mln[8] con programas elaborados por la autora.

## Primer estudio
## Asociación entre la *nota COU* y la *nota PAU* individuales: El efecto centro

En primer lugar se plantea el estudio de la variabilidad de la nota media que cada estudiante obtiene en las PAU –a la cual nos referiremos como *nota PAU*–, así como la determinación de las posibles fuentes de variabilidad asociadas. En especial, puesto que tanto la *nota COU* como la *nota PAU* sirven para evaluar la preparación del alumno para su ingreso en la universidad, se aborda el estudio de la variación conjunta entre ambas puntuaciones. Las preguntas planteadas al inicio de la investigación se concretaron en:

- ¿Todos los centros obtienen los mismos resultados en el COU y las PAU? ¿Existen diferencias significativas desde el punto de vista estadístico?

- En el caso que la respuesta sea afirmativa, además de conocer cual es la magnitud de la diferencia entre centros y de identificar los centros que en un sentido u otro se aparten del comportamiento medio, nos interesará conocer a qué se deben las diferencias observadas: ¿a la composición del alumnado?, ¿a la diferente preparación que ofrecen los centros? ¿a la aplicación de criterios de evaluación no uniformes?,... Pero, ¿es estadísticamente posible discernir una causa de otra con los datos de que disponemos? En todo caso, ¿qué información adicional necesitaríamos para poder discernir?

- ¿Existe asociación positiva entre la *nota COU* y la *nota PAU* de cada centro, es decir, los centros de secundaria que en COU se mantienen por encima de la media, también hacen lo mismo en las PAU?

- En el caso que el comportamiento de los centros varíe, ¿cómo recoger las diferencias entre centros? Concretamente, en cual de los siguientes indicadores que actualmente se proporcionan a los centros vale la pena poner más énfasis, en el sentido que reflejan diferencias significativas para un número importante de centros:

    - media de la *nota COU* de los alumnos del centro?
    - media de la *nota PAU* de los alumnos del centro?
    - diferencia entre ambas medias?

[8] Mln es un programa creado por *The Multilevel Project* del Institute of Education, University of London. Para más detalles sobre su funcionamiento, se puede consultar http://www.ioe.ac.uk./multilevel/

La vertiente empírica de esta primera investigación se ha basado en su mayor parte[9] en una muestra aleatoria de 26 centros y 1.619 estudiantes. La muestra fue extraída a partir de la población de centros de Catalunya (PAU de junio de 1993) y ha sido estudiada a lo largo de tres años.

El análisis exploratorio de los datos ilustra las diferencias entre los resultados de los exámenes COU y las pruebas PAU, sentando las bases para el estudio de la asociación entre ambas puntuaciones. Dicha asociación varía de un centro a otro. La modelización de la *nota PAU* individual por medio de modelos de regresión de coeficientes aleatorios permite evidenciar (y medir) las diferencias existentes entre centros de secundaria, diferencia que se materializa en el que se ha dado en llamar *efecto centro*. El primer capítulo de la tesis contiene una detallada introducción a los modelos de coeficientes aleatorios, también llamados modelos de nivel múltiple[10]. La aplicación de estos modelos complementada por su diagnosis confirma la variabilidad entre los estándares aplicados en COU.

El modelo de variación de dos niveles, en que los alumnos o unidades de primer nivel aparecen agrupados en centros o unidades de segundo nivel, especificado para la *nota PAU* y que llamaremos modelo (1), es el siguiente:

$$y_{ig} = a + bx_{ig} + u_g + e_{ig}$$

En este modelo, $y_{ig}$ es la *nota PAU* individual, $x_{ig}$ la *nota COU* correspondiente, $u_g$ es el residuo específico del centro $g$ (común a todos los alumnos de dicho centro) y $e_{ig}$ es el residuo específico del alumno $i$ del centro $g$. Las hipótesis sobre $u_g$ y $e_{ig}$ es que varían según una distribución de probabilidad de media 0 y varianza $\sigma_u^2$ y $\sigma^2$, respectivamente.

La diferencia entre este modelo y el modelo habitual de regresión de un solo nivel es que el primero admite la posibilidad de diferencias entre centros y permite la estimación de dos tipos de residuos (los debidos a cada centro y los debidos a cada alumno). El residuo $u_g$ del centro $g$ es una medida de la desviación de dicho centro respecto del comportamiento promedio. De ahí que se proponga llamar a $u_g$ *efecto centro.* El *efecto centro* $u_g$ es el valor añadido que debemos sumar a la predicción general de *nota PAU* a partir de la nota individual de COU por el hecho de proceder de un centro en concreto. El modelo especificado (1), en el cual se distingue una primera parte "fija" seguida de una parte aleatoria constituida por la suma de los dos residuos, admite una formulación alternativa como modelo de coeficientes aleatorios:

$$y_{ig} = a_g + bx_{ig} + e_{ig}$$

[9] Todos los modelos introducidos en este estudio han sido posteriormente validados con una segunda muestra de 53 centros y 3.500 estudiantes (PAU de junio de 1995).

[10] Con frecuencia, los datos en Educación –también, en Ciencias Sociales- presentan una estructura jerárquica (cada estudiante pertenece a un centro donde comparte profesores, métodos de enseñanza,...). Las observaciones o unidades del nivel inferior se agrupan en unidades del nivel superior, existiendo mayor homogeneidad entre los datos de un mismo grupo que entre un grupo y otro. Los modelos estadísticos para este tipo de datos han sido objeto de un intenso desarrollo en los últimos años. Véase Aitkin y Longford (1986); Goldstein (1995); Plewis (1997) y Kreft and De Leeuw (1998). En especial, el artículo de Aitkin and Longford (1986), en el cual los autores comparan la efectividad de un conjunto de centros utilizando diferentes modelos estadísticos representa un punto de partida en la investigación de modelos adecuados para el estudio de la variabilidad entre centros.

en que $a_g = a + u_g$ es un coeficiente aleatorio (no es fijo, compartido por todos los centros, sino que varía de un centro a otro según una distribución de probabilidad de media $a$ y varianza $\sigma_u^2$). Esta última formulación permite interpretar que, en el caso de que la varianza de $u_g$ sea significativamente distinta de cero, para cada centro existe una recta de regresión identificada por su ordenada en el origen $a_g$ . La desviación de cada ordenada respecto de la ordenada media $a$ es, precisamente, el *efecto centro* $u_g$.

Tabla 1
Modelos de regresión de la variable *nota PAU* en relación a la variable *nota COU* y a una serie de variables binarias. Entre paréntesis las estimaciones de los errores estándar. Muestra de 26 centros, datos de junio de 1993

| | **modelo A** MCO | **modelo B1** IMCG | **Modelo B2** IMCG | **modelo B3** IMCG |
|---|---|---|---|---|
| *Variables explicativas* | | | | |
| Constante | -0.84 (.17) | 5.24 (.10) | -0.68 (.17) | -0.36 (.18) |
| nota COU | 0.90 (.03) | | 0.88 (.02) | 0.85 (.02) |
| GENMAS | | | | 0.20 (.02) |
| REPCOU | | | | -0.23 (.05) |
| OPA | | | | -0.27 (.04) |
| OPB | | | | -0.41 (.05) |
| *Varianzas* | | | | |
| entre centros, $\sigma_u^2$ | _ | 0.22 (.07) | 0.18 (.07) | 0.18 (.05) |
| entre estudiantes, $\sigma^2$ | 0.706 | 1.03 (.04) | 0.52 (.02) | 0.48 (.02) |
| *coef. de correlación intra-centros* $r$ | _ | 0.175 | 0.25 | 0.27 |

El análisis y aplicación de estos modelos a los datos de la muestra confirmó que el coeficiente $b$ es fijo, común a todos los centros, mientras que el coeficiente $a_g$ es aleatorio, varía de un centro a otro.

La Tabla 1 resume las estimaciones derivados de la aplicación de diversos modelos de variación de la *nota PAU* a los datos de la muestra citada. En la columna de la izquierda se encuentran las estimaciones de los coeficientes del modelo de regresión ordinario (modelo A en la Tabla 1) estimado por el método de Mínimos Cuadrados Ordinarios (MCO). A su derecha, las estimaciones derivadas del modelo (modelo B1 en la Tabla 1) de descomposición de la varianza de la *nota PAU* en dos niveles (variación entre centros y variación entre estudiantes dentro de los centros). El modelo B2 es el modelo de regresión de la *nota PAU* sobre la *nota COU* de dos niveles. El modelo B3 es una ampliación del modelo anterior que incorpora las variables explicativas que identifican el género (GENMAS vale 1 para los hombres y 0 para las mujeres), si el estudiante ha repetido o no el COU (REPCOU vale 1 en caso afirmativo y 0 en caso negativo) así como la opción[11] cursada en COU (por ejemplo, OPA vale 1 si el estudiante ha cursado la opción A y 0 en caso contrario). En la estimación de los

[11] En un primera estimación del modelo B3 de la Tabla 1 y tomando como referencia base la opción C del COU se encontró que los resultados para la opción D no presentaban diferencias significativas. Los resultados que aparecen en la Tabla 1 son las estimaciones generadas por el modelo B3 después de prescindir de la variable OPD.

modelos de coeficientes aleatorios se ha utilizado el método Iterativo de los Mínimos Cuadrados Generalizados, IMCG (Goldstein, 1995).

Un enfoque complementario, basado en el estudio de la covarianza de las medias de COU y de PAU de cada centro a través de un modelo bivariante de descomposición de la variación total en variación entre centros y variación en los centros permite, una vez estimado el modelo, discutir la eficiencia de algunos indicadores educativos de los centros[12].

Entre las conclusiones de este primer estudio, cabe destacar:

- Mientras que la media de la *nota PAU* varía significativamente entre centros, la media de la *nota COU* apenas varía. Un 20%, aproximadamente, de la variación total[13] de la *nota PAU* corresponde a variación entre centros. Esta diferenciación entre centros que presenta la *nota PAU* se incrementa al hacer la regresión de la *nota PAU* respecto de la *nota COU*.

- En consecuencia, las diferencias existentes entre centros en cuanto a los resultados en las PAU no pueden atribuirse solamente a la composición de su alumnado. Una posible explicación sería que los centros se estuvieran rigiendo por diferentes estándares en la preparación y en la aplicación de criterios de evaluación de sus estudiantes, ordenando a sus alumnos sin tener en cuenta un referente externo, introduciendo cada profesor (y centro) su propio sesgo. Los centros estarían puntuando con criterios y escalas diferentes a pesar de que, como resultado, se obtengan distribuciones de aprobados en COU similares de un centro a otro.

- El modelo de regresión de dos niveles de la *nota PAU* versus la *nota COU* que contempla el género, si el alumno ha repetido o no el COU, las opciones de COU y el tipo de centro (público o privado) nos lleva a una serie de conclusiones en cuanto al papel predictor de estas variables que son coincidentes con otros estudios realizados a nivel estatal (Muñoz-Repiso et al., 1991). La novedad de nuestro enfoque se encuentra en la utilización de un modelo que permite determinar el efecto debido al centro en la nota PAU individual, y que tiene en cuenta al mismo tiempo la nota COU del estudiante así como otras características individuales y del centro.

- Se comprueba[14] que los estudiantes repetidores de COU obtienen en las PAU resultados, comparativamente, por debajo de sus compañeros. Las mujeres obtienen resultados en las PAU inferiores a lo que sería de esperar a partir de expediente de secundaria. Ambos factores (género y repetición de COU) se mantienen significativos a lo largo de los tres años estudiados, 1993-95. No se aprecian diferencias significativas en la *nota PAU* entre el conjunto de centros públicos y centros privados, en dicho período. En cambio, el factor opción de COU no mantiene dicha estabilidad. Una posible explicación de este último hecho se encontraría en que la dificultad de las pruebas puede diferir de un año a otro (la

[12] Para más detalles sobre los análisis y resultados de esta primera parte de la investigación, véase (Cuxart *et al*., 1997).

[13] En 1993 y para la muestra de centros en estudio, la media y la varianza de la *nota PAU* fueron 5.30 y 1.23, respectivamente. A su vez, para la *nota COU*, dichos valores fueron 6.75 y 0.68.

[14] Se estima, a partir de los datos, que los estudiantes repetidores de COU obtienen en promedio una *nota PAU* inferior en dos décimas de punto al resto de la población. Al mismo tiempo, en igualdad de condiciones en cuanto al resto de variables estudiadas, las mujeres obtienen en las PAU una nota inferior en dos décimas a la de sus compañeros, en promedio.

variabilidad observada de las medias por materias de toda la población a lo largo de los años pone en duda, a su vez, la constancia en el grado de dificultad de los exámenes de cada una de las materias).

- El grado de asociación entre notas medias de COU y de las PAU de cada centro es muy débil (no podemos rechazar, desde el punto de vista de la significación estadística, que la correlación entre dichas medias sea 0), indicando que los centros que en COU presentan una nota media alta, en relación a la población de centros, no siempre la mantienen en las PAU, sino que pueden incluso pasar a obtener resultados por debajo del promedio.

- En cuanto a la "posibilidad" de ordenación de los centros a partir de los resultados académicos de sus alumnos, se concluye que la ordenación más *informativa*, la que permite incluir todos los centros se obtiene a partir de la diferencia entre la media de la *nota COU* y la media de la *nota PAU* de cada centro. Sin embargo, este indicador del centro, que se ha venido utilizando en muchas administraciones, no es -como hemos podido comprobar a lo largo de los tres años-, tan estable como el indicador[15] que se deriva de la aplicación del modelo de regresión multinivel y que hemos llamado *efecto centro* ($u_g$ en el modelo (1)).

## Segundo estudio
## La calidad de la corrección en las pruebas PAU: Experimentación de un sistema de seguimiento

En el segundo estudio se analiza la calidad del proceso de corrección de las pruebas PAU. La investigación pretende evaluar la calidad de la corrección mediante el cálculo de indicadores adecuados, desvelar los puntos débiles del proceso de corrección y conocer el impacto de los mismos en el acceso a la universidad.

Estudios anteriores[16] habían apuntado la necesidad de evaluar la fiabilidad de dichos exámenes. A partir de un experimento de doble corrección[17] en el que participaron los correctores de Matemáticas y Filosofía de 18 tribunales de las PAU de junio de 1995 se generaron los datos necesarios para un estudio sobre la *fiabilidad* de la corrección en ambas materias. Un primer análisis de los datos ofrecía una clara evidencia de la discrepancia existente entre correctores así como de una mayor coincidencia en los exámenes de Matemáticas que en los de Filosofía. Por ejemplo, para un 72% de los exámenes de Matemáticas, la diferencia entre los dos correctores fue inferior o igual a un punto[18], mientras que en Filosofía este porcentaje fue tan sólo del 51%. La diferencia entre las dos correcciones superó los dos puntos en 77 exámenes de Filosofía (21 %) y en 14 exámenes de Matemáticas (7 % del total de dicha asignatura).

[15] Según dicho modelo, la varianza estimada del *efecto centro* es de 0.18. Los valores estimados del *efecto centro* para los 26 centros de la muestra varían entre 0.71 para el centro mejor situado y - 0.86 para el peor situado (aquel que obtiene los peores resultados en las PAU en relación a las notas de sus alumnos en el COU).
[16] Sans (1989), Muñoz-Repiso y otros (1991), *Memoria del C. de U.* (1993), Escudero (1994).
[17] Véase en Cuxart *et al.* (1997) los detalles del diseño y ejecución del experimento.
[18] La escala de puntuación era de 0 a 10.

La exploración inicial de los datos ofrecía también indicios de la existencia de una componente sistemática en el *error de medida*, componente que correspondía a diferencias entre correctores en cuanto al grado de severidad.

El modelo de variación para la nota observada[19] que se propone a continuación permite evaluar el impacto de los correctores en términos de la varianza debida a la *severidad* y de la varianza generada por la *inconsistencia*, ratificando las conclusiones de la exploración basada en la simple comparación de las notas de cada par de correcciones.

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 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 se propone para explicar la variación de la puntuación de un examen es el modelo aditivo (2):

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

siendo $i = 1, 2, ..., I$ el índice del examen o estudiante y $j=1, 2, ..., J$ el índice 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 $i$ ; $\alpha_i$ es la puntuación *verdadera* y no observable del examen $i$ ; $\beta_j$ es la *severidad* del corrector $j$ ; $\varepsilon_{ij}$ representa la *inconsistencia específica* de cada corrección. Se supone que estos tres últimos términos están mutuamente no correlacionados con medias iguales a $\mu$ , 0 y 0 , y varianzas $\sigma_a^2$ , $\sigma_b^2$ y $\sigma_e^2$ , respectivamente. Según este modelo la varianza total de las notas observadas deberían igualar a la suma de las tres varianzas componentes.

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*.

El estudio sobre la fiabilidad se amplió en 1997 a dos materias más: Biología y Literatura catalana. El estudio de esta segunda muestra ha corroborado los resultados de 1995 validando la modelización adoptada y permitiendo, al mismo tiempo, el inicio del estudio de dos nuevos temas: la dificultad y la capacidad discriminadora de las preguntas

[19] El modelo propuesto aparece documentado en Longford (1995, Cap. 2) en estudios sobre la fiabilidad de la corrección de preguntas de respuesta abierta. Para más detalles sobre su aplicación véase el artículo citado de Cuxart *et al.* (1997) donde se incluyen el análisis y los resultados relativos a 1995.

Los resultados de la estimación se incluyen en la Tabla 2, donde se puede ver como la varianza debida a la *inconsistencia* en el año 95 representa un 13% de la variación total en Matemáticas y un 34 % en Filosofía. La *severidad* no se aprecia en Matemáticas pero en Filosofía corresponde al 6% de la varianza total.

En la tesis se han desarrollado diversas técnicas de diagnosis que permiten, mediante la comparación con la distribución global, la identificación de correctores con influencia en el cálculo de la inconsistencia, correctores que adjudican notas muy dispares, correctores que adjudican notas demasiado similares entre sí, correctores que discrepan ostensiblemente de sus parejas,… La investigación ha permitido detectar *puntos débiles* en el proceso de elaboración y corrección de las pruebas PAU. La opcionalidad, existente en la mayoría de exámenes (A o B), es uno de ellos. Se ha comprobado que el grado de discrepancia entre correctores puede variar, de forma notable, entre opciones de examen.

Tabla 2

Estimación de las componentes de la varianza de la puntuación observada: $\hat{s}_a^2$, varianza entre *notas verdaderas*; $\hat{s}_b^2$, varianza de la *severidad* , $\hat{s}_e^2$, varianza de la *inconsistencia.*

| | $\hat{s}_a^2$ | $\hat{s}_b^2$ | $\hat{s}_e^2$ | Var. total |
|---|---|---|---|---|
| Junio 95 | | | | |
| Matemáticas | 5.350 (86.5%) | 0.011 (0.2%) | 0.827 (13.3%) | 6.188 |
| Filosofía | 2.475 (60.2%) | 0.248(6.0%) | 1.386 (33.7%) | 4.109 |
| Junio 97 | | | | |
| 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 |

Entre las conclusiones que se derivan del estudio de 1997 en cuanto a la variabilidad en la corrección, cabe destacar:

- Se observa un comportamiento similar en las asignaturas de Matemáticas y Biología en claro contraste con Filosofía y Literatura Catalana.

- En relación al estudio de 1995, se observa un aumento de la concordancia en la corrección de Matemáticas. 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 aún no disponen de ellas.

- Aunque la prueba de Filosofía consta de 5 preguntas valoradas en dos puntos cada una y la de Literatura Catalana consta tan sólo de dos preguntas valoradas en 5 puntos cada una, ambas materias presentan un patrón de descomposición de la variabilidad en la corrección muy similar.

- El examen de Biología muestra una fiabilidad muy alta, hecho que podría estar relacionada con su formato de preguntas de respuesta muy cerrada con criterios de corrección muy precisos.

- En las asignaturas de Filosofía y Literatura Catalana los correctores sólo disponen de los criterios generales de corrección. Este hecho podría explicar el comportamiento diferenciado de un grupo de asignaturas respecto al otro.

Se incluye a continuación, por su interés como soporte cualitativo al estudio empírico presentado, un resumen de las entrevistas realizadas a los responsables de varios distritos universitarios[20]. Las entrevistas tomaron como punto de partida un cuestionario relativo a las pruebas PAU-COU de junio de1996, al que previamente habían dado respuesta.

## Los sistemas de corrección de las pruebas de Selectividad en España. Análisis y propuestas.

***Datos generales***

Todas las universidades, a excepción de Catalunya, dedican tres días a la realización de las PAU. El número de alumnos por tribunal suele ser superior a 500 en promedio, excepto en las universidades UPC y UPF de Catalunya donde el número habitual es 200. Los correctores disponen de 5 o 6 días para corregir y el período de reclamaciones suele ser también de 5 días tal como establece la normativa. La publicación de las notas definitivas tiene lugar la segunda semana de Julio. El número de alumnos que se examinaron en la convocatoria 96 de las PAU-LOGSE fueron: 577 en Galicia; 410 en la UAM; 700 en la UCM; 2538 en Catalunya; y 810 en Zaragoza. El número de tribunales fue, respectivamente: 1, 1, 1, 14 y 3.

[20] Agradecemos la colaboración de los responsables de los distritos universitarios que a continuación señalamos: Comunidad Autónoma de Madrid (que incluye la U. de Alcalá, UAM, Carlos III, UCM y UPM, con un total de 41.155 estudiantes); Catalunya (que incluye la UAB, U. de Girona, UPC, UPF, URiV, U. de Lleida con un total de 29.850 estudiantes); Zaragoza (los centros de Teruel y Huesca también pertenecen a la U. de Zaragoza que en total son 6.461 estudiantes); Comisión Interuniversitaria de Galicia, CiUG, (A Coruña, Santiago y Vigo con 14.616 estudiantes. El cuestionario también se envió a otras universidades pero no se obtuvo respuesta.

***Elaboración de las pruebas***

En todos los distritos y para cada materia se nombra un profesor universitario como coordinador responsable de la elaboración de la prueba, excepto en el distrito de Madrid donde, para cada asignatura, se forma un equipo con dos responsables de cada universidad. Antes de darlos por definitivos, los exámenes son resueltos por personas que, en su mayoría, han participado también en la elaboración de los mismos. La valoración del grado de dificultad de las preguntas se basa en el criterio del coordinador. No se utilizan referencias empíricas[21]. En algunos distritos y con posterioridad al examen, antes o después de su corrección, se recoge la opinión sobre su dificultad. En Galicia se reúne el Seminario Permanente antes de la corrección para establecer los criterios de específicos. En Catalunya se recoge la opinión de los centros de secundaria a través de una encuesta posterior a la entrega de notas. Tan sólo la UAM y para las asignaturas de Lengua y Literatura, Historia y Filosofía, utiliza una base de datos con preguntas y modelos de examen en la elaboración de los mismos.

***Sobre el futuro (sistema LOGSE)***

En referencia a las novedades que presenta la elaboración de los exámenes de las pruebas PAU-LOGSE, la mayoría de los entrevistados coinciden en destacar el hecho que el temario de las pruebas no esté fijado por la Universidad, la falta de definición en muchas materias, la dificultad de incluir en las pruebas contenidos relativos a procedimientos, en definitiva, la complejidad del nuevo sistema y la escasa información acerca de las enseñanzas impartidas en los centros.

En algunos distritos se han formado grupos mixtos universidad-secundaria en aras de una mayor coordinación entre los dos ámbitos y con la intención de concretar los contenidos y objetivos de las pruebas de acceso a la Universidad. Entre las propuestas de los entrevistados, señalamos las más comunes:

- Establecimiento de un grupo estable de profesores elaboradores de las pruebas. Coordinación de este grupo de profesores elaboradores con profesores que están impartiendo clases. Previamente, debería garantizarse un nivel mínimo de homogeneidad en la interpretación de los temarios a impartir en el bachillerato.
- Reducción del número de asignaturas objeto de examen. Adecuar las materias de examen al Acceso Universitario.
- Avanzar hacia un formato de examen, al menos en parte, de preguntas de respuesta cerrada. Abarcar en el examen la casi totalidad del temario exigido. Limitar la opcionalidad o, al menos, racionalizarla.
- Realizar pruebas piloto para conocer la dificultad de las preguntas. Elaborar unos criterios de corrección más precisos. Para asegurar una uniformidad de criterios en la corrección de las preguntas de respuesta abierta se propone la elaboración de pautas de corrección aplicadas al examen concreto así como la realización de reuniones con los correctores. Debería arbitrarse un mayor seguimiento de las actuaciones de cada corrector. Se propone separar la labor de vigilancia de la labor de corrección -en algunos distritos, como CiUG, ya se está realizando.

En conclusión, según se desprende de las opiniones de los entrevistados, los últimos años del COU cierran un periodo en el que se ha hecho un avance importante en:

[21] Debe observarse que, puesto que los notas no se introducen en el ordenador teniendo en cuenta la opción de examen ni la puntuación de cada pregunta, es imposible en la actualidad hacer un seguimiento empírico, ni tan siquiera para una muestra, de la dificultad de cada pregunta y cada opción de examen.

- **estandarización**: la mayoría de los distritos universitarios elaboran un único examen para todas las universidades y tribunales
- **anonimato**: se han arbitrado sistemas que garantizan el anonimato del estudiante y del corrector (algunos sistemas son mejores que otros)
- **informatización**: poco a poco se han ido informatizando todos los sistemas con una ganancia considerables en tiempo, seguridad y información

Los responsables de las PAU, sin embargo, son conscientes de las imperfecciones que permanecen en el sistema de evaluación: discrepancia entre correctores de una misma materia, falta de homogeneidad en los planteamientos y evaluaciones de las diferentes pruebas, diferencias en los criterios y escalas de evaluación en aquellas puntuaciones que son responsabilidad de los centros de secundaria,… A la preocupación por reducir las imperfecciones citadas se añade la entrada en vigor de una nueva secundaria (la ESO y el bachillerato LOGSE) y la necesaria adecuación de las pruebas de acceso a la universidad.

Coinciden los responsables de las PAU en señalar la necesidad de fijar los contenidos de las materias evaluables en las PAU-LOGSE. Al mismo tiempo, querrían sacar más partido de la información generada por estas pruebas. En general, se acusa la falta, en mi opinión, de un plan de control de la calidad del sistema de evaluación así como de seguimiento de las novedades y mejoras que se vayan introduciendo.

## Tercer estudio
## Estructura de covarianza del conjunto de puntuaciones PAU. Elementos de análisis alrededor de la nueva fórmula de acceso

En el momento en que se inició este estudio, las pruebas de acceso a la universidad para los estudiantes del bachillerato LOGSE habían sido reguladas de manera provisional. Las propuestas sobre cuál debería ser su diseño (materias, contenidos, ponderaciones, criterios de evaluación,..) apuntaban hacia una doble prueba: una prueba común, genérica, de *madurez académica* y una prueba *específica* o de *contenidos* sobre los conocimientos adquiridos en el bachillerato y que estaría relacionada con los estudios universitarios que el alumno deseera cursar. En las actuales pruebas PAU que siguen los estudiantes que han superado el COU ya se observaba esta doble orientación. El primer ejercicio con un carácter más general y el segundo determinado por las materias específicas de la opción de COU cursada.

La discusión generada sobre las funciones[22] y estructura del nuevo examen era una invitación a la reflexión las funciones de las PAU del COU y el análisis del comportamiento del conjunto de las pruebas a la luz de los datos existentes. Este último ha sido el motivo del estudio que se presenta: explorar la estructura del vector de notas PAU del COU, conocer la capacidad discriminadora de cada prueba y cada ejercicio, así como el grado de asociación entre las diferentes pruebas. El interés se ha centrado no solamente en la variación total de las notas sino también en las diferencias entre centros.

Una de las funciones de las pruebas de acceso es ubicar a los estudiantes adecuadamente, ordenarlos en función de sus aptitudes y conocimientos demostrados en el examen y reflejados en el expediente académico. Una de las "virtudes" que ha de satisfacer una prueba como ésta es la de ser discriminadora en el sentido que separe

[22] Véase Martí *et al.* (1997) y Muñoz-Repiso *et al.* (1997).

correctamente los estudiantes. No es bueno que una prueba "iguale"' alumnos con niveles diferentes de conocimientos y/o aptitudes. El sistema debería "aprovechar al máximo la información disponible". Por ejemplo, si dos asignaturas están midiendo las mismas habilidades o conocimientos y además presentan una correlación alta pero una de ellas está calificada con mas fiabilidad que la otra sería recomendable evaluar únicamente la de mayor fiabilidad o, en todo caso, evaluar ambas asignándoles pesos diferentes al calcular una nota agregada. La idea subyacente es clara: el proceso de evaluación ha de ser el más eficiente posible. Está claro que previamente deberíamos saber qué se pretende evaluar, comprobar que las pruebas realmente evalúen las habilidades o conocimientos especificados (validar las pruebas) y conocer la fiabilidad del proceso.

La actual *nota de acceso* a la universidad es la media aritmética de la *nota PAU* y la *nota Expediente* del alumno. La *nota PAU e*s la semisuma de dos notas agregadas que llamaremos, respectivamente, *nota primer ejercicio* y *nota segundo ejercicio*.

La *nota primer ejercicio* es una media ponderada de las pruebas que componen el primer ejercicio: Comentario de Texto, Lengua catalana, Lengua castellana, Lengua extranjera y Filosofía. Las pruebas del primer ejercicio tienen como finalidad evaluar la madurez y formación general del alumno. Es de interés observar si todas estas pruebas correlacionan entre sí y por un igual con la *nota primer ejercicio*, como sería de esperar si todas ellas estuvieran midiendo el factor *madurez* del estudiante.

La *nota segundo ejercicio* es la media aritmética de las cuatro pruebas que componen el segundo ejercicio. Las pruebas de este ejercicio tienen como finalidad evaluar la formación específica del alumno en las materias de la opción escogida. Estas cuatro pruebas corresponden a las cuatro asignaturas específicas cursadas por el alumno en el COU: dos son las materias obligatorias de la opción y las otras dos las escogió el alumno entre las optativas de la opción de COU.

Las preguntas planteadas al inicio de este estudio se concretaron en:

- ¿Cuál es la dimensionalidad de las pruebas? ¿Qué información aportan los dos ejercicios? ¿Son diferentes?

- ¿Cuáles son las materias o agregaciones que mejor discriminan (separan los estudiantes) a nivel global? ¿Qué materias o agregaciones presentan diferencias entre centros? ¿Qué materias o agregaciones presentan diferencias entre estudiantes dentro de los centros?

- ¿Cómo es la relación entre los resultados de las diferentes materias? ¿Es homogénea dicha relación, o más bien, se dan diferencias según género, opción de COU, tipo de centro, haber repetido o no el COU, …?

- ¿Qué ponderaciones son "razonables" en la situación actual?

Para el estudio de la variación total se han aplicado técnicas clásicas de análisis exploratorio como el Análisis de Componentes Principales (ACP), que han permitido poner en evidencia la capacidad "separadora" de cada prueba y de cada nota agregada (nota de las pruebas comunes, nota de las pruebas específicas de la opción, *nota PAU*,...), destacando las diferencias que se observan en cuanto al papel de cada bloque de pruebas en los cuatro grupos de estudiantes que se derivan de la opción de COU

escogida. Para el estudio de la variación entre centros se especificó un modelo multivariante que distingue la variación a nivel centro de la variación a nivel estudiante.

Los datos utilizados en este estudio proceden de la muestra de 26 centros del distrito de Catalunya presentada en el primero de los estudios y analizada a lo largo de tres años.

### *Análisis exploratorio*

Entre las conclusiones que se derivan de los Análisis[23] de Componentes Principales (ACP) por opciones de las notas PAU, cabe destacar:

- Un hecho común a los cuatro análisis es la pobreza de la representación (el porcentaje de varianza que recogen los dos primeros ejes ronda el 50%) y el papel del primer eje o *factor rendimiento* que separa los estudiantes con mejores resultados de aquellos que, en general, los obtienen peores. El primer eje está relacionado con las asignaturas comunes y con las obligatorias de la opción (sobretodo con las tres Lenguas). El segundo eje separa, tanto en la opción A como en la B, las materias específicas de las tres Lenguas. En la opción C el segundo eje viene definido por Matemáticas II, prueba que también presenta una correlación alta con el primer eje. El segundo eje en la opción D no presenta ninguna asociación relevante.

- Las pruebas del primer ejercicio correlacionan débilmente entre sí (valores alrededor de 0.3). En particular, la prueba de Comentario de Texto[24] presenta correlaciones muy bajas (alrededor de 0.2), incluso negativas, con el resto de pruebas de dicho ejercicio. Una explicación para este hecho podría ser que esta prueba (nos referimos exclusivamente al distrito de Catalunya) esté valorando aptitudes muy diferentes del resto de exámenes. Nosotros nos inclinamos a pensar que por un lado la prueba se puede mejorar tanto en su formato como en sus criterios de corrección y por el otro, que el entrenamiento que han recibido los alumnos es muy diverso.

- El hecho de que las pruebas del primer ejercicio aparezcan poco correlacionadas entre sí, siendo éste un resultado común a otros estudios de ámbito estatal, se podría

[23] Con anterioridad, T. Net (1996) había estudiado los resultados de COU y PAU de los estudiantes de la muestra con una atención especial a la matriz de covarianza de las notas PAU por materias. Net realizó dos análisis ACP tomando como variables activas las puntuaciones de COU en primer lugar y las de PAU posteriormente. En el ACP de las notas COU los porcentajes de inercia que recogen el primer y el segundo eje son, respectivamente, 37.6% y 12.6%. En el ACP de las notas PAU dichos porcentajes son 24.3% y 11.1%. En ambos casos aparece un primer eje de *rendimiento académico* o nivel del estudiante y un segundo eje que separa las asignaturas de ciencias de las de letras. El análisis de la nota PAU permite distinguir los resultados de los alumnos según el género, si son o no repetidores de COU, la opción de COU y el tipo de centro, sugiriendo al mismo tiempo la necesidad de estudiar por separado las cuatro opciones de COU.

[24] Parece ser que la prueba de Comentario de Texto en Catalunya presenta diferencias respecto del resto de distritos universitarios. Según un estudio (Muñoz-Repiso *et al.*, 1997) realizado por el equipo que dirige Mercedes Muñoz-Repiso del CIDE, a partir de los resultados de las pruebas PAU de junio de 1995, de 12.117 estudiantes procedentes de 130 centros adscritos a la UAM, en las cuatro opciones, la correlación entre Comentario de Texto y el resto de pruebas del primer ejercicio toma valores alrededor de 0.3. Este mismo estudio, al comparar los resultados de las diferentes universidades del Estado revela que en Catalunya se dan los porcentajes de aprobados en Comentario de Texto más bajos de todo el Estado. En la mayoría de universidades del resto del Estado esta prueba es la que obtiene el porcentaje más alto de aprobados.

interpretar como que dichas pruebas estén valorando aspectos diferentes de la preparación[25] del estudiante.

- Del análisis del grado de correlación entre el primer eje factorial y los resultados de las pruebas (primer y segundo ejercicio) cabe destacar que mientras en las opciones A y B el primer eje está más correlacionado con el segundo ejercicio que con el primero, en la opción C la correlación es prácticamente la misma con los dos ejercicios, y en la D, justamente al revés.

- Las materias que más influyen en la ordenación de los alumnos en las PAU son Matemáticas I en las opciones A y B y Matemáticas II y Historia del mundo contemporáneo en la opción C. Cabe destacar que al mismo tiempo y a lo largo de los tres años estudiados estas tres materias se encuentran entre las de nota media global más baja.

- En cuanto a los grupos que presentan diferencias al realizar los análisis ACP del vector de nota PAU, cabe destacar los siguientes:
  - El ser repetidor es el *efecto* más importante. Se podría interpretar que el nivel exigido en COU es alto y requiere una preparación adquirida con anterioridad.
  - El tipo de centro (público o privado) también presenta diferencias. El análisis de la *nota COU* y la *nota PAU* (primera parte de la tesis) no señalaba diferencias significativas entre estos dos tipos de centros. En cambio, al estudiar el vector de notas PAU sí se revelan diferencias entre centros públicos y privados. Un elemento que diferencia los centros privados de los públicos es que en estos últimos el porcentaje de repetidores es mucho mayor, siendo, como hemos dicho, los alumnos repetidores los que obtienen los peores resultados en la mayoría de pruebas.
  - El género. En especial en la opción A los resultados de las mujeres son inferiores a los de sus compañeros. ¿El tipo de examen las perjudica?

### *Modelización: descomposición de la variación total*

A continuación se resume el modelo especificado y los resultados obtenidos en el estudio conjunto de las materias del primer ejercicio.

Tabla 3
Variación total y variación entre centros para cada materia del primer ejercicio. Muestra de 26 centros, junio de 1993

| | Lengua Catalana | Lengua Castellana | Filosofía | Lengua extranjera | Comentario de Texto |
|---|---|---|---|---|---|
| Varianza total | 3.846 | 2.744 | 3.113 | 3.782 | 2.482 |
| Varianza entre centros | 0.935 | 0.574 | 0.473 | 0.391 | 0.592 |
| Coef. Corr. *intra-centros* $r$ | 0.24 | 0.21 | 0.10 | 0.15 | 0.23 |

[25] Los datos estadísticos sugieren en este caso preguntas pero no ofrecen respuestas. Se plantea la necesidad de definir qué se entiende por *madurez*, cómo evaluarla y si las actuales pruebas de acceso están diseñadas para tal evaluación. ¿La *madurez* tiene una única dimensión? ¿Qué dice al respecto la psicología y, en particular, la psicometría?

En primer klugar, la aplicación de un modelo de efectos aleatorios sin covariantes a cada una de las pruebas del primer ejercicio por separado permitió descomponer la variación total en variación entre centros (entre medias de los centros) y variación dentro de los centros (entre sus estudiantes) y la consiguiente estimación del valor del coeficiente de correlación *intra-centros* $\rho$ en cada prueba. Los resultados de dicha estimación se encuentran en la Tabla 3. Los valores del coeficiente[26] $r$ , sugerían la conveniencia de aplicar modelos multinivel para el estudio de las posibles diferencias entre centros. Al mismo tiempo, puesto que el interés también se centraba en el estudio de la asociación entre materias, se planteó la necesidad de formular un modelo multivariante. En consecuencia, la elección recayó en los modelos multinivel multivariantes (Goldstein, 1995).

Para el estudio de la variación conjunta de las cinco pruebas comunes que constituyen el primer ejercicio de las PAU se especificó un modelo multivariante multinivel que descompone la matriz de varianzas y covarianzas residual en variación entre centros y variación dentro de los centros y que permite, al mismo tiempo, estimar para cada materia el *efecto* $u_{kg}$ debido al centro y el *efecto* $e_{kig}$ específico de cada alumno en la nota observada (el subíndice *i* se refiere al estudiante y el subíndice *g* al centro de secundaria en que ha cursado el COU). Los parámetros correspondientes a las varianzas y covarianzas de estos *efectos* aleatorios forman la matriz de variación entre centros y la matriz de variación entre estudiantes. Dichas matrices son un elemento clave en el estudio de la variación conjunta (Longford, 1994). El modelo en cuestión es:

$$\begin{array}{l} x_{1ig} = \mu_1 + \sum_r \beta_r y_{rig} + u_{1g} + e_{1ig} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{5ig} = \mu_5 + \sum_r \beta_r y_{rig} + u_{5g} + e_{5ig} \end{array} \qquad (3)$$

Según se desprende de los análisis ACP de las puntuaciones PAU, el factor género así como el factor repetidor, el tipo de centro y la opción de COU son elementos significativamente diferenciadores. De ahí que en el modelo (3) se incorporen para cada respuesta una serie de variables explicativas $y_{rig}$, binarias en este caso, que recogen las características citadas. En el modelo (3) se distingue una parte de efectos fijos, integrada por las $\mu_k$ que representan las medias de las *categorías base* (alumna no repetidora de centro privado que cursa la opción D) en cada materia y por los parámetros $\beta_r$ que representan los efectos principales debidos a las diferentes categorías (alumno versus alumna, público versus privado, repetidor versus no repetidor, opción A/B/C versus D) y una parte de efectos aleatorios integrada por los $u_{kg}$ y los $e_{kig}$. En ausencia de covariantes, las $\mu_k$ representan el parámetro media de cada materia.

[26] Con valores por encima de 0.10 en todas las materias y superior a 0.20 en tres de ellas, según se recoge en la Tabla 3. El coeficiente de correlación *intra-grupos*, que ya apareció en el primer estudio de este informe, se suele representar con la letra del alfabeto griego ρ. Es una medida de la homogeneidad de los grupos en relación a toda la población. En situaciones de dos niveles (estudiantes y centros, por ejemplo) el coeficiente de correlación *intra-centros* se interpreta como la proporción de variación total observada que corresponde a variación entre centros. En presencia de correlación *intra*, la hipótesis de independencia entre observaciones, tradicional de los modelos de regresión, no se satisface. Se puede estimar el coeficiente de correlación *intra-grupos* a través de un análisis de la varianza ADEVA clásico o vía descomposición de la variación total por un modelo de efectos aleatorios con ausencia de covariantes (Searle, 1991).

Tabla 4

Estimaciones resultantes de la aplicación del modelo (3) de descomposición de la variación conjunta para las pruebas que integran el primer ejercicio de las PAU. En la diagonal las varianzas, debajo las correlaciones. Muestra de 26 centros, junio de 1993. Todos las estimaciones que aparecen en la tabla son significativas para un nivel de significación de 0.05.

| | Catalán | Castellano | Filosofía | Lengua ext. | Comentario |
|---|---|---|---|---|---|
| *Efectos principales* | | | | | |
| Media categoría base | 5.07 | 6.12 | 4.85 | 5.43 | 5.78 |
| Repetidor de COU | -1.05 | -0.82 | -0.67 | -0.87 | -0.31 |
| *Variación entre centros* | | | | | |
| Catalán | 0.69 | | | | |
| Castellano | - | 0.44 | | | |
| Filosofía | - | - | 0.45 | | |
| Lengua ext. | - | - | - | 0.33 | |
| Comentario | - | - | - | - | 0.51 |
| *Variación intra centros* | | | | | |
| Catalán | 2.74 | | | | |
| Castellano | 0.32 | 2.05 | | | |
| Filosofía | 0.23 | 0.27 | 2.57 | | |
| Lengua ext. | 0.25 | 0.32 | 0.22 | 3.24 | |
| Comentario | 0.23 | 0.23 | 0.19 | 0.25 | 1.87 |

Se ha estimado el modelo (3), desde la versión más simple que admite solamente variación entre estudiantes hasta la versión más compleja, que admite, también, variación entre centros e incorpora variables explicativas. La Tabla 4 recoge los resultados de la estimación multivariante multinivel[27]. De la información que se deriva de las Tablas 3 y 4, cabe destacar:

- Las pruebas del primer ejercicio que en la exploración inicial presentan más variabilidad de resultados son Lengua catalana (varianza total igual a 3.85) y Lengua extranjera (3.78). Las medias de los centros varían significativamente en todas las pruebas del primer ejercicio. Las diferencias entre centros más acusadas se dan en Lengua catalana (varianza entre centros igual a 0.94) y el Comentario de Texto (0.59). Dentro de los centros, las materias que presentan más diversidad de puntuaciones son la Lengua extranjera[28] (varianza entre estudiantes dentro de los centros igual a 3.24) y la Lengua catalana (2.74).

- En el modelo de variación que no contempla variación entre centros, las covariantes opción de COU, género, tipo de centro y ser o no repetidor de COU tienen un efecto significativo en las pruebas del primer ejercicio. Al admitir la existencia de variación

[27] Las estimaciones de la Tabla 4 corresponden al modelo que ofrece un mejor ajuste a los datos de la muestra. En los modelos de nivel múltiple, la significación de los parámetros fijos se suele analizar –siempre que se disponga de una muestra suficientemente grande de datos, como en el caso que nos ocupa- a partir del criterio común (distribución del estadístico aproximadamente normal) de dividir la estimación por el error estándar. Si el cociente es superior a 2, se considera el parámetro como significativo. En el caso de coeficientes aleatorios de los que queremos estimar su varianza o covarianza, no es aconsejable hacer uso exclusivo del criterio anterior (Goldstein, 1995). Es mejor tener en cuenta, al mismo tiempo, la información que suministra el test de la *razón de verosimilitud* que compara el ajuste de los datos observados a los modelos estadísticos que resultan de incluir o no los parámetros en cuestión. Este ha sido el método seguido en la selección del modelo de variación para nuestros datos.

[28] Para más de un 95% de los estudiantes se trata de Lengua inglesa.

entre centros (como en la versión más completa del modelo (3)) desaparecen los efectos de todas estas covariantes a excepción del ser o no repetidor de COU. Este hecho abona la necesidad de considerar el modelo multinivel: los efectos estimados en el modelo de un solo nivel se debían al comportamiento singular[29] de algunos centros y no respondían a un comportamiento más general de los alumnos. El ser repetidor tiene un efecto negativo en los resultados de las cinco pruebas (los repetidores obtienen notas, alrededor de un punto en cada una de las Lenguas y 0.7 puntos en Filosofía, por debajo de sus compañeros).

- No se han estimado correlaciones significativas entre las medias de los centros por materias. El hecho de que un centro se sitúe por encima de la media en una asignatura no lleva asociado que ocurra lo mismo en otra asignatura. La estimación vía estimadores encogidos (Longford, 1994) de los *efectos* debidos a los centros en cada materia confirma este comportamiento no uniforme de los centros: no encontramos centros con resultados por encima de la media global en todas las materias, o por debajo de la media en todas. Sí destacan, en cambio, algunos centros que para alguna de las materias presentan resultados muy alejados del comportamiento general. Quizás sería más adecuado hablar de *efecto profesor* que de *efecto centro*...

### *Discusión*

Las correlaciones entre las diferentes pruebas que integran las PAU son muy bajas, incluso si las calculamos para los estudiantes de una misma opción. De ahí que los análisis de componentes principales deban ser considerados tan sólo como elementos de ayuda en la reflexión. Las diferencias entre centros no explican suficientemente las bajas correlaciones observadas, puesto que al descomponer la variación total tampoco se obtienen correlaciones entre materias más altas entre los alumnos, dentro de los centros.

La pregunta sigue abierta ¿por qué las correlaciones son tan bajas? Entre las respuestas posibles se encontrarían las siguientes:

- Las materias del primer ejercicio que, en un principio, deberían evaluar la *madurez* del alumno, se ajustan más al programa de COU de las correspondientes asignaturas que a un criterio de evaluación que tenga que ver con el concepto de madurez.
- La corrección de preguntas de respuesta abierta conlleva subjetividad, imprecisión. Un error de medida importante[30] en cada evaluación tiene el efecto de *atenuar* (Fuller, 1987) los coeficientes que miden la relación[31] entre variables.

[29] El análisis detallado de los *efectos centro* en cada materia puso en relieve, por ejemplo, que un bajo rendimiento femenino (en promedio) en Lengua catalana se debía a la existencia de un centro de secundaria de 40 chicas que habían obtenido los peores resultados en esta asignatura. Al admitir diferencias entre centros, desapareció este efecto global (y ficticio) siendo capitalizado por el correspondiente *efecto* debido al referido centro en dicha asignatura.

[30] En la Tabla 2 se estimaba la proporción de varianza de la nota observada que correspondía a error de medida (error aportado por la imprecisión en el proceso de corrección). En el caso de Filosofía, esta proporción, tanto en 1995 como en 1997, rondaba el 40%. En la tesis, partiendo de la hipótesis que el grado de fiabilidad de las cinco pruebas comunes sea similar al de Filosofía (60%, aproximadamente) se realiza un cálculo estimativo de las correlaciones dentro de los centros resultando valores notablemente más altos y también más cercanos a los valores observados entre las respectivas asignaturas de COU.

[31] La imprecición en la corrección incrementa la varianza estimada entre estudiantes dentro de los centros. Si se consigue reducir el error de medida, es de esperar que, no solamente se observe una mayor

En el caso de las pruebas específicas de cada opción[32], se añaden las siguientes reflexiones:

- Si los formatos de examen de dos asignaturas propias de la opción son muy dispares, las pruebas pueden estar evaluando no solamente conocimientos distintos sino también diferentes habilidades de los estudiantes.
- Los actuales exámenes no cubren de manera exhaustiva la programación (Muñoz-Repiso *et al.*, 1997). 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.

Respecto a la consideración formulada al inicio de este estudio sobre si valía la pena examinar de dos materias si su correlación era muy alta, podemos decir que no tiene demasiado sentido su planteamiento, puesto que no se han observado correlaciones de dicha magnitud. Según los análisis realizados, se podría inferir que las pruebas están midiendo aspectos diferentes de la preparación del alumno y, por tanto, un examen no debería sustituir a otro automáticamente. Pero, también es cierto que el alto grado de imprecisión en la corrección introduce incertidumbre en el proceso y alerta sobre la formulación de tales conclusiones. Aunque es evidente que las pruebas de corrección objetiva tienen también sus limitaciones, parece aconsejable estudiar la posibilidad de introducir este tipo de pruebas –al menos como parte común del examen de cada materia- para un adecuado seguimiento del proceso.

En el futuro sería aconsejable trabajar con una muestra mayor de centros. Estudios recientes basados en la simulación recomiendan que en el caso de que la *correlación intracentros* sea superior a 0.10 el número de unidades del segundo nivel (centros) sea como mínimo de 30 unidades para asegurar una estimación eficiente de los parámetros relativos a las variables de este nivel y de las interacciones entre los dos niveles. La estimación de los parámetros relativos al primer nivel (características de los estudiantes, en los estudios que nos conciernen) demandan un número total de unidades suficiente, preferentemente el mismo número de estudiantes por cada centro. Una muestra de 30 estudiantes por centro sería un buen punto de partida. El criterio general (Kreft and De Leeuw, 1998) es el siguiente: para asegurar una potencia alta de análisis es preferible un diseño muestral de muchos centros con pocos alumnos cada centro que la situación inversa (pocos centros con muchos alumnos cada centro), sobretodo si la variación entre centros es alta en relación a la variación dentro de los centros.

## Conclusiones

Se detallan, a continuación, algunas de las principales conclusiones que se desprenden del conjunto de los tres estudios. Puesto que la investigación empírica se ha basado en muestras de centros de Catalunya, es obligado matizar que todas las conclusiones que se apoyan en dichos datos deben limitarse a este distrito.

---

asociación entre las notas de las materias sino que también las diferencias entre centros sean más acusadas que las observadas hasta el momento.

[32] No se ha incluido en este informe el análisis de las pruebas del segundo ejercicio de las PAU por opciones. Destaca uno de los resultados del análisis relativo a los alumnos de la opción A: mientras que Matemáticas I es la prueba que presenta más variabilitat globalmente y entre estudiantes dentro de los centros, es en Física donde se dan los resultados más variables de un centro a otro.

- La *nota PAU* revela que la preparación de los alumnos no es tan similar como la *nota COU* nos llevaría a presuponer: mientras la primera varía significativamente entre centros (alrededor de un 20% de la variación total de la *nota PAU* corresponde a variación entre centros, según datos de los tres años analizados), no ocurre lo mismo con la *nota COU*. La diferenciación entre centros que revela la *nota PAU* no solamente se mantiene sino que se incrementa[33] al hacer la regresión de la *nota PAU* respecto de la *nota COU*, confirmando que las diferencias entre centros, en cuanto a los resultados en las PAU, no se deben únicamente a la composición de sus alumnos. Nuestra conclusión es que los centros se rigen por diferentes estándares en la preparación y evaluación de sus alumnos. Los profesores y los centros estarían clasificando y ordenando a sus alumnos sin tener en cuenta un referente común externo introduciendo cada profesor su propio sesgo. Este hecho avala la precaución con que se debería considerar tanto la *nota COU* como la *nota Expediente* y, al mismo tiempo, cuestiona la afirmación recogida en la *Memoria del Consejo de Universidades* de 1993:

  > "Y hay datos concluyentes[34] de que el expediente es mejor predictor del rendimiento en los estudiantes universitarios que las pruebas realizadas (Escudero, 1981, 1986). Por ello no estaría en absoluto justificada la minusvaloración de aquél en el cálculo de la calificación global de acceso".

Tampoco estaría justificado, sin embargo, que el mismo *Ministerio de Educación* se basara en estos argumentos para defender un mayor peso[35] de la *nota Expediente* en el cómputo de la *nota de acceso*, sobretodo porque ha pasado mucho tiempo desde el trabajo de Escudero (más de 20 años) y de Touron (más de quince años) y en este período las universidades han introducido cambios importantes en los primeros cursos

---

[33] El coeficiente de correlación *intra-centros* $r$ calculado a partir de los residuos individuales del modelo de regresión ordinario (modelo A en la Tabla 1) es superior a 0.3, confirmando que dentro de los centros se da mayor similitud que en general y rebatiendo la hipótesis de independencia entre residuos que presupone dicho modelo.

[34] Al hablar de *datos concluyentes* interpreto que se refieren a los resultados de un estudio longitudinal de seguimiento -realizado por un equipo de investigadores, que dirigía Tomás Escudero- de 417 estudiantes, que el curso de 1975-76 superaban las pruebas de acceso y accedían a las universidades de Navarra y Zaragoza. El mismo autor, además de precisar las limitaciones del estudio, nos decía:" En definitiva, la prueba de madurez académica -seguida de cerca por el expediente secundario- aparece como el mejor predictor del rendimiento universitario cuando se trata de la muestra total". Y es que, en mi modesta opinión, el hecho más relevante del estudio de T. Escudero (1987) donde se resumen los dos documentos citados en la *Memoria de ...* es que pone en evidencia las diferencias existentes entre estudios universitarios: para los estudiantes de Medicina (127 de 417, un 30% de la muestra) el rendimiento en la universidad aparece más asociado a la *nota de Expediente* (correlación 0.40) que a la *nota PAU* (correlación 0.05); para los estudiantes de Ingeniería Superior (35 de 417, un 8%) las correlaciones son, en cambio, 0.44 y 0.56, respectivamente.

Otro trabajo empírico al que con frecuencia los analistas se refieren es el de J. Touron (1987). Touron estudia el rendimiento de 165 estudiantes de primer curso de Medicina de la Universidad de Navarra (junio de 1984) y encuentra que la nota media de las cuatro asignaturas de ciencias de secundaria predice mejor el éxito en primero de Medicina que la *nota PAU.* El autor no se refiere al poder predictivo de las cuatro pruebas específicas de las PAU, quizás porque no disponía de dicha información desagregada.

[35] En el momento de redactar este informe ya se ha publicado el Real Decreto 1640/1999 de 22 de octubre (BOE de 27-10-99) por el que se regula la prueba de acceso a estudios universitarios que en su artículo 14, establece

> "Para ser declarado apto por una vía de acceso deberá obtenerse, al menos, cuatro puntos en la calificación global para esa vía. [...] La calificación definitiva para el acceso a estudios universitarios se calculará ponderando un 40 por 100 la calificación global de la prueba y un 60 por 100 la nota media del expediente académico del alumno en Bachillerato".

(como las normas de permanencia y las fases selectivas, por ejemplo). Además son muchos las Facultades y Escuelas Universitarias que han realizado estudios -la mayoría para uso interno, sin llegar a ser publicados- que analizan la relación entre el éxito en los primeros cursos de universidad y las puntuaciones en las diferentes pruebas PAU. Por ejemplo, un estudio de la Facultad de Informática de la UPC sobre 165 estudiantes que ingresaron en 1993, y de los cuales se disponía de todas sus notas en COU y las PAU, reveló que la materia mejor relacionada con el éxito en primer curso era la prueba de Matemáticas en las PAU, por delante de la puntuación de esta materia en COU y de la *nota Expediente* de secundaria.

- Si quisiéramos resumir en pocas palabras el papel evaluador de la *nota COU* y de la *nota PAU*, diríamos que mientras la *nota COU* es el resultado de una evaluación en principio muy completa (el profesor tiene muchas oportunidades para valorar el conocimiento y madurez del alumno) pero afectada de sesgo (los criterios de evaluación no son los mismos de un centro a otro), la *nota PAU* es el resultado de la aplicación de un instrumento estándar (el mismo para todos los alumnos) pero afectado de error de medida (debido al tipo de examen de preguntas de respuesta abierta). Para comparar los efectos de ambas imperfecciones en el cómputo de la nota de acceso hemos de tener en cuenta que dicha nota es la media de la *nota PAU y la nota Expediente*. En la actualidad son muchos los estudiantes que solicitan una revisión de la corrección de alguna de las pruebas PAU. Según se ha podido estimar a partir de los datos, la revisión (en caso de ser oportuna) podría comportar una rectificación al alza de la nota de acceso de 0.23 puntos, si la asignatura revisada es Matemáticas y de 0.32 si es Filosofía. Sin embargo, no se cuestiona en absoluto la *nota Expediente* cuando cabría preguntarse ¿Por qué no se admite la rectificación de la nota de acceso de todos los alumnos de un centro si existen indicios suficientes para pensar que dicho centro ha evaluado con mucho más rigor que el resto?

- El hecho de que el acceso a algunos estudios universitarios en que la oferta es inferior a la demanda no esté restringido a determinadas opciones y la evidencia de la no aplicación de criterios uniformes de puntuación entre las diversas materias de las PAU, provoca una situación desigual. Dicha situación de desequilibrio puede facilitar la aparición de estrategias de acceso no deseables. Al mismo tiempo, cuestiona el sentido de la actual fórmula de acceso, concretamente la necesidad de examinar de determinadas materias y en su caso la ponderación asignada. No parece adecuado examinar de materias que no se consideren importantes para los estudios universitarios que solicita el estudiante si, al mismo tiempo, no se puede garantizar la aplicación de criterios de evaluación uniformes.

- Surge la pregunta de si las pruebas del primer ejercicio están evaluando adecuadamente la madurez del alumno. Parece ser que se reducen a pruebas sobre los contenidos de las materias de COU comunes.

- En la *nota de acceso* a la Universidad (y en consecuencia en la ordenación de estudiantes en el momento de competir por las plazas disponibles) de un estudiante de Ciencias tiene el mismo peso la nota de Matemáticas que la de Filosofía cuando, según hemos podido comprobar, se trata de asignaturas con diferente capacidad discriminadora y diferente calidad en la corrección. ¿Debería ser así?

- El análisis por materias pone en evidencia la heterogeneidad de centros de secundaria en cuanto a los resultados en las PAU y la conveniencia de informar a los centros sobre sus resultados en comparación con la población de centros.

## Monitorización de las pruebas PAU

En el artículo "Monitoring the university admissions process in Spain" (Cuxart and Longford, 1998) se encuentran muchas de las reflexiones que han ido surgiendo a medida que avanzaba la investigación, así como propuestas, alguna de las cuales ya se han podido experimentar. Cabría distinguir tres áreas específicas de reflexión y posible actuación en el futuro:

1. Mejora de los exámenes en cuanto a su elaboración, homogeneización de la corrección y posible intervención para corregir discrepancias o desajustes. Considerar en cada materia la posibilidad de substituir el examen actual, o una parte del mismo, por una prueba de preguntas de respuesta cerrada.
2. Formación de coordinadores y correctores. Posibilidad de separar la labor de vigilancia de la labor de corrección.
3. Creación de un sistema de información útil para la Administración educativa y para los centros de secundaria (profesores y alumnos).

## Futuras líneas de investigación

Son varias las líneas de investigación que surgen motivadas por la necesidad de profundizar en temas que tan sólo han podido ser apuntados:

a) La validación de las pruebas PAU requiere un estudio detallado de los enunciados y contenidos de los exámenes. ¿Son pertinentes las preguntas? ¿Certifican la secundaria? ¿Preparan para la universidad?
b) El conocimiento empírico de la dificultad y el poder discriminador de las preguntas[36] permitiría explicar mejor las diferencias observadas entre convocatorias y facilitaría la confección de nuevos esámenes.
c) El seguimiento de los alumnos en la universidad se plantea como un elemento de estudio imprescindible[37] toda vez que, según parece, en las futuras PAU los alumnos deberán examinarse de un número menor de asignaturas y con una relación más estrecha con los estudios universitarios.
d) El estudio (cualitativo en su mayor parte) de los mecanismos de elección de los alumnos que les lleva a preferir una materia optativa a otra, y para cada materia, una opción de examen frente a otra, permitiría conocer el porqué de las preferencias de los alumnos y proporcionaría un mayor soporte en la elaboración de las pruebas.
e) Tanto en las pruebas PAU-COU como en las pruebas PAU-LOGSE[38], las mujeres han obtenido peores resultados que sus compañeros a igualdad de condiciones (para una misma *nota Expediente*). En cambio, los resultados en secundaria de las mujeres están muy por encima de los de sus compañeros (el porcentaje de aprobadas

[36] La *Oficina de Coordinació del COU i les PAU de Catalunya* ha iniciado una investigación al respecto.
[37] Son varias las universidades que han iniciado estudios de seguimiento de sus alumnos.
[38] Según se deduce del estudio de las primeras promociones del bachillerato LOGSE.

supera en varios puntos al de aprobados). Se plantea un interrogante sobre cuáles son las dificultades específicas de cada género, los mecanismos de aprendizaje, la evaluación de resultados, los incentivos (la valoración por parte de los profesores y familiares,...)

f) Otro tema importante y que hace referencia al comportamiento de los correctores en las preguntas de respuesta abierta, es el estudio de las causas de la inconsistencia en la corrección de cada materia: ¿Qué factores provocan la modificación de criterio de un corrector? ¿Qué elementos le producen inseguridad en la puntuación? El conocimiento de los mismos influiría en la redacción de los exámenes y pautas de corrección así como en la definición de las condiciones adecuadas para una buena corrección (tiempo, lugar, entorno, …).

## Referencias bibliográficas


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

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

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. (1987) Buscando una mejor selección de universitarios. *Revista de Educación*, 283: 249-283.

Escudero, T. y Bueno García, C. (1994). Examen de Selectividad. El estudio del tribunal paralelo. *Revista de Educación,* 304: 281-297.

Fuller, W. (1987). *Measurement Error Models*. Wiley, New York.

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

Kreft, G. (1987). *Models and methods for the measurement of school effects.* Tesis doctoral. Universidad de Amsterdam.

Kreft, G. and De Leeuw (1998). *Introducing multilevel modelling* Sage publications, London.

Longford, N.T. (1994a). Random Coefficient Models. *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* Arminger and Sobel editors. Plenum Press, New York.

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

Martí Recober, M. *et al*. *Los sistemas de corrección de las pruebas de Selectividad en España. Análisis y propuestas*. Concurso nacional de Proyectos de Investigación Educativa (1995-98). Ministerio de Educación y Ciencia, CIDE.

Martí, M., Ferrer, F. y Cuxart, A (1997). El desarrollo de la LOGSE: las nuevas Pruebas de Acceso a la Universidad. *Revista de Educación*, 314:89-114.

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

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., Murillo, F., Arrimadas, I., Navarro, R., Díaz-Caneja, P., Martín, A., Gavari, E., Molinonuevo, J., Gómez, A. y Fernández, E. (1997) *El sistema de acceso a la Universidad en España: tres estudios para aclarar el debate*. Madrid: CIDE.

Net, T. (1996). *Análisis multivariant de la informació continguda a l'expedient dels alumnes que accedeixen a les PAAU*. Projecte final de carrera, Diplomatura d'estadística. UPC, Barcelona.

Plewis, I. (1997). *Statistics in Education*. Arnold. London.

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

Searle, S.R., Casella, G. And McCulloch, Ch.E. (1991). *Variance Components.* Wiley Interscience, New York.

Touron, J. (1987) High school ranks and admission tests as predictors of first year medical students' performance. *Higher Education,* pages 257-266.