University admission marks in Catalonia: some highlights from the empirical research[1]

Anna Cuxart[2], *Rosa Grau*
Departament d'Economia i Empresa, Universitat Pompeu Fabra
Manuel Martí-Recober
Departament d'Estadística i I.O.,Universitat Politècnica de Catalunya

Abstract

The results of the examinations taken by graduated high school students who want to enrol at a Catalan university are here studied. To do so, the authors address several issues related to the equity of the system: reliability of grading, difficulty and discrimination power of the exams. The general emphasis is put upon the concurrent research and empirical evidence about the properties of the examination items and scores. After a discussion about the limitations of the exams' format and appropriateness of the instruments used in the study, the article concludes with some suggestions to improve such examinations.

Keywords: Admissions process, reliability of grading, index of difficulty, discrimination power

Journal of Economic Literature Classification: C89, C99, I29

1 Introduction

Till 1997 students in the last year of high school in Spain took a set of examinations called the *Curs d'Orientació Universitària* (COU). These examinations were prepared and marked by the staff of the respective high school. Students who have passed the COU and want to enrol at a public university had to take another set of examinations: *Proves d'Accès a la Universitat* (PAU). These examinations have a much stronger element of standardization than COU exams. In Catalonia (population of about six million, approximately 15 per cent of the population of Spain) the PAU examinations are prepared, administered and scored by the *Coordinació de les PAU*, an institution created by the seven public Catalan universities.

The *Coordinació*, which also maintains an extensive database of PAU and COU marks and scores, started a project off in 1995 to monitor the quality of the university admission process in Catalonia. The research[3] allowed a better knowledge of the process, spotting those aspects which would require higher control or improvement. Among those aspects we point out the confirmed variability of COU evaluation standards among secondary school centres, the discrepancy of markers' scores to the PAU[4] questions with open answers, and the level of difficulty of the exams.

[1] Research supported by Concurso Nacional de Proyectos de Investigación Educativa, Spanish Ministry of Education, and Coordinació del COU i les PAU, Interuniversity Council of Catalonia.
[2] anna.cuxart@upf.edu, rosa.grau@upf.edu, manuel.marti-recober@upc.es
[3] For more details, see Cuxart and Longford (1998), Cuxart (2000) and Grau, Cuxart and Martí-Recober (2002).
[4] In the frame of the referenced PAU improvement project, A. Cuxart performed a double-marking experiment with a sample of Mathematics I and Philosophy exams, which involved all markers of these subjects from 18 PAU examination courts in June/1995. The experiment was performed simultaneously with the exams, to highly guarantee the usual conditions of the official marking. The study allowed measuring the precision (or reliability) of the marking and revealed that some of the marking variability causes were related to the building and format of the

The aim of the PAU exam is to orderly assign the students to the different degrees. So, following an equity principle, the exam should be able to separate them sufficiently and adequately.

We think that in the building of the PAU exams several aspects should be taken into account, in addition to the appropriateness of the questions. They are the following: the score given by the markers should lead to a minimum variability, the level of difficulty should be adequate, and the students' answers should allow sufficient discrimination. As far as these principles are applied to the building of the exams, the analysis of results stability would be meaningful and the results of one year would be able to be compared with others'.

The aim of this work is the research and experimentation of analysis instruments that would allow knowing the level of difficulty of the PAU exams and their ability to "separate" students adequately.

In section 2 the data are presented and the uncertainty associated to exam review is analysed. Section 3 is devoted to the study of question difficulty and discriminative ability. In section four we discuss aspects related to the exploitation of the information generated by the exams. Finally, we conclude with some pedagogical considerations in section 5.

2 Data and reliability

The results analysed in the following sections come from four samples of approximately 100 exams each from the subjects of Philosophy, Mathematics I, Catalan Literature and Biology from the June/1997 PAU exams (COU). Those samples were chosen with two purposes: 1) to measure the quality of marking by comparing it with a previous study[5], and 2) to initiate an analysis on difficulty and discriminatory ability of questions. Each exam was marked by the official marker and two other supporting ones. So, for each exam three marks were available, which we would label correspondingly *official mark*, *replay1* and *replay2*.

In Table 1 (see also Chart 1) a summary of the official marks assigned to the four subjects (between 0 and 10 in each exam) is presented. The number of exams, average mark, standard deviation (SD) and pass percentage (marks greater or equal to 5) are shown. The exams of the four subjects were composed of questions with open answers.

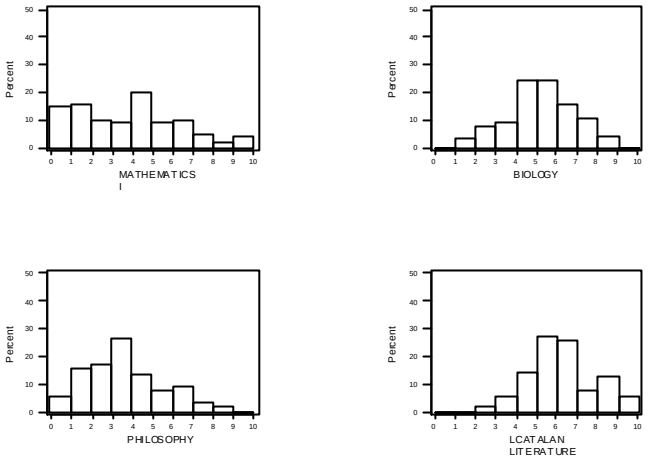
Table 1. Summary of global results. Sample of exams, PAU 97.

	Philosophy	Mathematics I	Biology	Catalan Literature
n	100	100	100	70
average	3.40	3.60	5.20	5.90
SD	1.86	2.49	1.61	1.68
% pass	22	30	59	79

exams. This study confirmed the need to perform systematically empirical studies and the possibility to perform experiments related to the writing of exams

[5] See Grau, Cuxart and Martí-Recober (2002) for a report of the compared study of marking reliability 1995/1997.

Chart 1. Histograms of the score distribution for each subject. Official marker.



Marking reliability

The mark given by each marker to each question intends to be a measure of the knowledge the student has on the subject. This measure is taken from the student's answers and, naturally, it is not error free. So, in any analysis about difficulty and discriminatory ability of questions the uncertainty inherent to each mark and the need to choose a good estimator of the *true mark* must be taken into account, meaning by *true mark* the mark the student deserves for his exam.

A common instrument in analysing concordance between marking is Pearson's correlation coefficient. Table 2 shows these figures by subjects and pairs of markers. The correlation coefficients show that marking concordance is very good in Mathematics I and Biology. However, we must not forget that the sample correlation coefficient is a good estimator of the marking precision or reliability whenever the *severity*[6] degree between markers is equivalent[7]. Longford (1994) introduced a decomposition model of the observed marks' variation that allows the calculation of marking quality indicators in complex situations. In this model the mark given by each maker to each exam is decomposed as the sum of the *true mark* that would correspond to the exam, the marker's *severity* and *inconsistency*[8]. This model allows reliability estimation by means of a unique coefficient calculated as the proportion of the total variation that corresponds to the true mark. For more details about this model and its applications, see Longford (1995) and Cuxart (2000).

[6] By marker *severity* we mean the difference between two non-observable figures: "the marker's mean" (that we would know if she/he corrected all the exams) and "the global mean" (that could be calculated if all exams were marked by all markers).

[7] When markers have different severity levels the value of the sample correlation depends on the design (Longford, 1995), that is, on the specific assignment of exams to markers.

[8] The inconsistency or "non systematic error" is a mixture of flaws that are present in the marking process. The specific inconsistency of each exam and marker would be "the deviation of the given mark from the mark the specific marker would give to the exam in average".

Table 2. Marker reliability estimations

	Philosophy	Mathematics I	Biology	Catalan Literature
Pearson's coeffic. of correlation				
oficial, replay1	0.60	0.95	0.84	0.69
oficial, replay2	0.52	0.91	0.87	0.63
replay1, replay2	0.60	0.91	0.87	0.56
Reliability coeffic. (modelization)	*0.42*	*0.92*	*0.85*	*0.57*

In each subject the student was allowed to choose between two exam options. As Table 3 shows both options were not in general equivalent regarding the concordance between markers. In Philosophy option B the value of the correlation coefficient between the official marker and replay2 (of 0.29) is to be pointed out. This value does not allow to discard the correlation null hypothesis equal to zero. The correlation between marks is much better in Mathematics I and Biology than in Philosophy and Catalan Literature.

Table 3. Marker reliability estimations. Results per exam option

	Philosophy		Mathematics I		Biology		Catalan Literature	
exam option	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
number of exams	78	20	46	54	72	28	62	8
oficial, replay1	0.61	0.59	0.95	0.95	0.87	0.80	0.63	--
official, replay2	0.61	0.29	0.91	0.91	0.92	0.72	0.65	--
replay1, replay2	0.66	0.43	0.84	0.94	0.87	0.88	0.56	--

Marking reliability of questions

The variability produced in an exam marking is the sum of the small discrepancies in each question. The correlation coefficient has some limitations in studying the reliability in the marking of each question because the theoretical marking rank in some questions is very narrow. If this were, its interest would be found through the detection of behaviour patterns and extreme deviations.

As Table 4 shows there is a strong agreement in marking some questions and strong discrepancies in marking others, in all subjects. In Philosophy the correlation coefficient values are low, being 0.76 the maximum value and reaching negative values in question 1B. On the contrary, in Mathematics I most values are higher than 0.8.

The exam formats[9] are different for the four subjects. Biology questions have a very short answer. On the contrary, the questions in Philosophy and Catalan Literature are open and not bounded. This fact may probably and partially explain the higher discrepancies found between marks assigned by markers in these two subjects.

[9] Questions' marks depend on the subject: Philosophy (1.5, 1, 3.5, 2.5, 1.5), Mathematics I (2, 2, 3, 3), Biology (2, 2, 2, 2, 2) and Catalan Literature(5, 5).

Table 4. Agreement in marking questions. Correlations between official marker/replay1, official marker/replay2 and replay1/replay2.

	Question 1	Question 2	Question 3	Question 4	Question 5
Philosophy A	0.38	0.28	0.53	0.47	0.29
	0.42	0.47	0.50	0.37	0.52
	0.30	0.50	0.60	0.46	0.18
Philosophy B	**-0.11**	0.68	0.18	0.42	0.61
	-0.39	**0.07**	0.32	0.43	0.76
	0.49	0.17	0.10	0.75	0.51
Mathematics I A		0.93	0.83	0.97	
	*	0.92	0.65	0.95	
		0.85	0.59	0.90	
Mathematics I B	0.78	0.90	0.94	0.92	
	0.50	0.90	0.94	0.89	
	0.62	0.92	0.99	0.99	
Biology A	0.81	0.96	0.92	0.65	0.75
	0.73	0.96	0.93	0.70	0.71
	0.82	0.97	0.85	0.73	0.76
Catalan Literature A	0.63	0.39			
	0.60	0.50			
	0.51	0.38			
Catalan Literature B**					

* No one of the students who chose option A in Mathematics answered question 1 correctly.
** Option B in Catalan Literature was chosen by only 8 students

3 About the difficulty and discriminating power of exams and questions

The arithmetic mean and standard deviation are indicators of exam difficulty and discriminating power, respectively. The difference in the exams' level of difficulty of the different subjects is self-evident in Table 1. We must not forget that students which attend the PAU examination have previously passed all subjects in the centre where they come from. Table 1 and Chart 1 show that Mathematics I and Philosophy exams were difficult, giving Mathematics I a better discrimination.

In order to know the behaviour of the different questions that compose the exams several indexes used in psychometry[10] were tested.

Estimation of the true mark starting from the median

Since we had three marks for each exam and question, we chose to estimate the *true mark* starting from the median of the three marks. From this point on we have used the median as an estimation of the *true mark* that would correspond to the student exam in all estimations about difficulty and discriminating power.

[10] See Del Rincón, D, Arnal, J, Latorre, A and Sans, A. *Técnicas de Investigación en Ciencias Sociales*, 1995.

The difficulty of the questions

The difficulty of a question was estimated through the percentage of right answers, taking into account that a student had hit a question when her/his *true mark* were greater than half of the points assigned to the question. In Table 5 the hit percentages on several questions in the four subjects are shown. A classification of the level of difficulty in five levels (Del Rincon *et al*., 1995) was taken as a reference to guide the analysis: very "easy" question, hit by more than 75% students, easy (between 55 and 74%), average difficulty (between 45 and 54%), difficult (between 25 and 44%) and very difficult (less than 25%). Taking into account these categories one may see that question 1 in Mathematics I (in both options) was very difficult while question 5A in Biology could be labelled as very easy.

Table 5. Percentage of students which hit each question.

	Question 1	Question 2	Question 3	Question 4	Question 5
Philosophy A	64	42	33	32	33
Philosophy B	30	30	55	25	45
Mathematics I A	0	20	9	50	--
Mathematics I B	6	44	39	26	--
Biology A	21	72	42	17	85
Biology B	32	46	64	75	62
Catalan Liter. A	55	40	--	--	--

Experts say that a balanced exam should be composed of 10% of very easy questions, 20% of easy questions, 40% of average difficulty questions, 20% of difficult questions and 10% of very difficult questions. On account of the short number of questions in the exams under analysis and the broad range of marks assigned to those questions, it is almost impossible to get a similar empirical distribution. With these constraints on account, the percentage of the total mark of the corresponding exam at each difficulty level was estimated, once the difficulty of the question was known. Table 6 shows the observed distribution for each subject and exam option. We could draw our attention to the 85% of the whole mark (8.5 points) of the Philosophy option A exam which went to questions that turned out to be difficult and the 15% (1.5 points) which went to easy questions, being this distribution different from the one corresponding to option B of the same subject.

Table 6. Observed level of difficulty as a percentage of the whole mark.

	very easy	easy	average	difficult	very difficult
Philosophy A	0	15	0	85	0
Philosophy B	0	35	15	50	0
Mathematics I A	0	0	30	0	70
Mathematics I B	0	0	20	60	20
Biology A	20	20	20	0	40
Biology B	0	0	20	60	20
Catalan Liter. A	0	0	50	50	0

Exam and question discrimination

Variance and its square root, the standard deviation, measure the discriminating ability of an exam. A zero variance means that all students have got the same mark and there is no chance to

put them in order, so discrimination is null. A higher variance is associated to a higher separation among students. One of the known variance's limitations is that it is very susceptible to change with extreme values. Hence, it is always recommended to enclose graphical representations with the statistical summaries, to weigh up the information that they carry in a better way.

The discrimination indexes

In psychometry it is common to use a discrimination index defined in the following way: Students (exams) are ordered in three groups of the same size (A: highest marks, B: medium and C: lowest) according to a previous mark or reference. For each question the difference between the hit percentage in group A and the hit percentage in group C is calculated. This index measures the ability the questions have to separate the best students from those which show les performance regarding the reference. A question that was hit by all students in group A and none in group C would have a discrimination index of 100%, and, in this sense, would be considered optimum. Questions with no hit and those which are answered by all students have a discrimination index of zero. In the study we are presenting, the reference chosen was the PAU *true mark* of he subject.

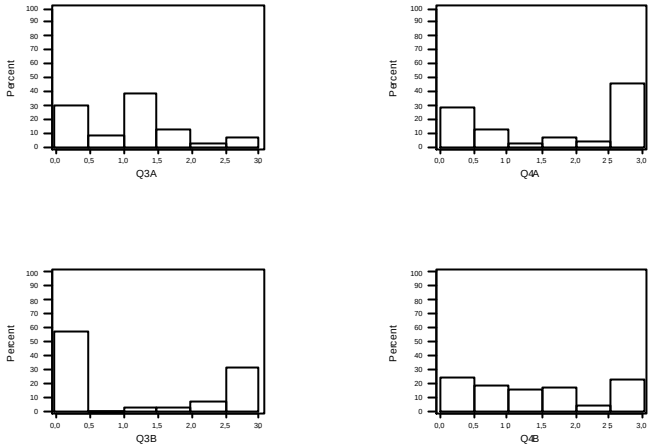
Experts think that discrimination index values under 20% show a poor discrimination ability. Table 7 shows the values obtained per subjects and questions, being evident that, in general, the best discriminating questions are those which are considered to have an average difficulty (see Table 5).

A methodologyc objection to the use of the index defined above is founded in the fact that the same question, whose discriminating ability is being measured, is part of the used reference. Another important objection about this index stems from the groups' definition, because it is possible for two students with the same *true mark* not to be classified in the same group. Those problems led us to consider alternative estimators for measuring the discriminating ability of each question. Taking into account the reduced and diverse maximum mark assigned to the questions, we chose as an alternative estimator the quotient between the observed standard deviation (SD) and half the mark allocated to the question, which means we divide SD by 1.5 for the three-point questions and by 1 for the two-point questions. This coefficient measures the variation in connection to the magnitude of the question. Therefore it is a relative variation coefficient.

Table 7. Discrimination index per subjects and questions. Values in percentage.

	Question 1	Question 2	Question 3	Question 4	Question 5
Philosophy A	26.9	38.5	80.7	57.7	61.5
Philosophy B	57.1	28.6	71.4	28.6	57.1
Mathematics I A	0.0	40.0	20.0	100	------
Mathematics I B	6.0	72.0	89.0	72.0	------
Biology A	42.0	50.0	87.5	29.2	20.8
Biology B	66.5	55.6	88.9	77.8	22.3

Chart 2. Histograms of the distributions of the *true mark* of Mathematics I. Questions of three marks.



In Chart 2 and Table 8 that follow we show the application of the different discriminating level estimators for the Mathematics I exam, as an example. Chart 2 and Table 8 show the differences in questions and the role of the estimators.

No student hit question 1 in option A, therefore this question was not useful to separate or discriminate students. Similarly, question 1B was hit by 6% of students only, so it is not a useful question for discriminating. On the contrary, question 2 in both options establish separation among students, neither being very difficult or optimum. Question 3 in option B and question 4 in option A should be emphasized as optimum in the studied context because they are the ones which best discriminate, with discrimination indexes of 89% and 100%. In addition, both question practically do not cause variability in the marking, being their correlation coefficients higher than 0.9. The hit percentage is 39% in question 3B and 50% in question 4A.

Table 8. Summary of the questions in Mathematics I (*true mark*).

	Questions of 2 marks				Questions of 3 marks			
	1A	**1B**	**2A**	**2B**	**3A**	**3B**	**4A**	**4B**
Observed mean	0	0.59	0.52	1	0.87	1.12	1.64	1.19
Standard deviation (SD)	0	0.51	0.69	0.69	0.79	1.34	1.35	1.04
Discrimination index (%)	0	6	40	72	20	89	100	72
Coef. of variation (%)	0	51	69	69	53	89	90	69

4 Discussion: About the instruments and their application

About the exam format and its limitations.

The variety of exam formats (number of questions and marks) is a limitation when comparing indicator values among subjects. In some subjects the short number of questions limits the ability to estimate student knowledge. Exams with a large number of questions allow to embrace not only a large part of the program but a broader range of difficulty levels too.

Some of the questions have lead to huge discrepancies between markers. Teachers should discern among those questions which arouse agreement in marking and those which arouse discrepancies. These last ones may be of importance as learning instruments but their inclusion in an external objective test is, in fact, more debatable.

About the indexes used

The discrimination index, being very useful in other fields (psychometry, tests with many questions ...) has great flaws in its application to the PAU marks and student record marks, as it has been shown. The need to define three groups of students, for instance, leads to separate students with the same mark in different groups, introducing an arbitrary factor. An alternative to this index is the relative variation coefficient introduced in this study, which shows two advantages. The first one is that the coefficient gives us a measure of discrimination without the need to have a reference previously, the second one is that it removes the ordainment and further categorization of students. Another alternative measure could be the percentage of hits conditioned to the different mark intervals used as a reference.

The difficulty index is a good estimator of a question's difficulty level. In the building of an exam we should not forget that the best discriminating questions are those of medium difficulty.

About the observed samples

Students chose an exam option. To compare results from the two options adequately it would be necessary to perform an experiment to vanish the choice effect, by assigning the options to students randomly or by asking each student to answer both options.

The samples, being valid to study marking reliability, could be biased to study difficulty when they do not represent student population (students were from 6 secondary centres located in the same area in the city of Barcelona).

5 Main conclusions and pedagogical considerations

We point out the main conclusions and pedagogical considerations that emerge from the present study:

- We think that the study of question quality analysing the variability generated in their marking and their separating power is just a previous step to the building of an experienced question bank.

- A fact to be pointed out is that teachers responsible for the PAU exam preparation actively participated in the discussion created by this study, and they took its results into account for the building of the exams.

- The exams, as objective tests, are useful not only as part of student evaluation but also to contribute to sound information to analyse and improve student learning and centre performance. To which extent are secondary teachers used to considering exams as learning instruments? Have teachers the habit of analysing exam results to find out about students'

difficulties and progress? In this sense we think that the indicators used in this study can be a useful tool in the internal evaluation processes of secondary school centres.

?? Extreme situations, as questions hit by almost every student or questions not answered by any student, should not be the general rule of an exam. Such questions provoke a deficient separation between students because they equate marks of students with different levels of knowledge. Against the opinion of many evaluating teachers, extremely difficult questions do not allow to capture adequately the different student preparation levels, so they simply do not discriminate enough.

?? At present, the explicit purpose of the PAU is for admission to universities; however, it is well-known that the existence of the PAU also have contributed to the homogeneity of the high schools' teaching. As the educational system in Spain is undergoing a reorganization and some politician have considered the elimination of the PAU, it is important do not subestimate this indirect effect of the PAU.

References

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

Cuxart, A. (2000): "Modelos estadísticos y evaluación: tres estudios en educación", en *Revista de Educación*, núm.323, pp. 369-394.

Del Rincón, D., Arnal J., Latorre A., Sans A. (1995): *Tecnicas de Investigación en Ciencias Sociociales*. Madrid. DYKINSON.

Grau, R., Cuxart, A y Martí-Recober, M. (2002): "La calidad en el proceso de corrección de las pruebas de acceso a la universidad: variabilidad y factores", en *Revista de Investigación Educativa*, Vol. 20, nº 1, págs. 209-223.

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