# Monitoring the university admissions process in Spain

Anna Cuxart i Jardí and Nicholas T. Longford
Department of Economics and Business Studies, Universitat Pompeu Fabra,
Barcelona, Spain, and Department of Medical Statistics, De Montfort
University, Leicester, England

## Acknowledgements

The paper was written while Anna Cuxart was visiting De Montfort University, Leicester. She was supported by research grants from *Comissionat per a Universitats i Recerca*, the Autonomic Government of Catalonia, and *Concurso Nacional de Proyectos de Investigación Educativa*, Spanish Ministry of Education.

The paper was presented at the 10th European Meeting of the Psychometric Society, Santiago de Compostela, July 15-18, 1997 and European Conference on Educational Research, ECER 97, Francfurt, 24-27 September, 1997.

## ABSTRACT

The examinations taken by high-school graduates in Spain and the role of the examination results in the university admissions process are described. The following issues arising in the assessment of the process are discussed: reliability of grading, comparability of the grades and scores (equating), maintenance of standards, and compilation and use of the composite scores. Studies to assess rater reliability and other imperfections of the grading process, and their integration in the operational grading are proposed. Various schemes for score adjustment are reviewed and feasibility of their implementation discussed. The advantages of pretesting of items and of empirical checks of experts' judgements are pointed out. The paper concludes with an outline of a planned reorganisation of the higher education in Spain, and with a call for a comprehensive programme of empirical research concurrent with the operation of the examination and scoring system.

Key words: *admissions process, examinee's choice, pretesting, rater reliability, score adjustment*
*Journal of Economic Literature classification:* C99, I29

## High-school examinations in Spain

Students in the last year of the high school in Spain sit a set of examinations called *Curso de Orientación Universitaria* (COU). The examinations are taken in seven or eight subjects. Three subjects are compulsory (Philosophy, Spanish, and a foreign language) and four are selected by the students. The optional subjects are in four categories (Science and Technology; Biomedical Sciences; Social Sciences; and Humanities and Languages), and each student has to select them from one category. Within each category there are two compulsory subjects (e.g., Mathematics and Physics in Science and Technology), and a number of options, from which the student has to select another two subjects. In addition to the three universally compulsory subjects, students in Catalonia have a compulsory examination in Catalan. The examinations in the various subjects are prepared and graded by the students' high-school staff. The Spanish Ministry of Education issues general guidelines which promote, though do not enforce, standardisation and at the same time allow for regional differences. The examinations are graded with regard to student's performance in the current academic year. For each examination, the continuous grading scale 0–10 is used. A student is classified as having passed COU if he/she achieves grade 5 or higher on every examination. For historical reasons, the grades are classified into four categories: 'pass' for scores 5.00–5.99 (coded as 5.5), 'good' for scores 6.00–6.99 (coded as 6.5), 'very good' for scores 7.00–8.49 (coded as 7.5), and 'excellent' for scores 8.50–10.00 (coded as 9). This system has been in place since 1974.

The COU examinations take place in May, with substitute dates in September. At present the minimum school-leaving age in Spain is 14, although students leaving school at this age receive vocational training for at least two further years. Most high-school students who reach the fourth year (typical age 18) sit the COU examinations and about half of them pass at the first attempt. Students who fail one or several examinations in May can resit them the following September.

High-school students receive grades at the end of each of the first three academic years for every subject they study. These grades, classified into the same four categories as COU (5.5, 6.5, 7.5, and 9), are combined with COU into an overall mean. In Spanish, this composite score is referred to as the *Expediente*. Here we refer to it as the *high-school record*.

Another set of examinations taken by the students who have passed COU and want to enrol at a public university in Spain is *Pruebas de Aptitud para el Acceso a la Universidad* (PAAU). (Until recently, there were only a few private universities, most of them theological seminaries. Now there are several new privately funded business schools and similar institutions.) These examinations have a much stronger element of standardisation enforced by examination boards responsible for a geographical area (*distrito universitario*). For instance, Catalonia (population of about six million, approximately 15 per cent of Spain's population) is administered by one such board. In Catalonia, the examinations are prepared, administered, and scored by *Coordinació del COU i les PAAU*, an institution funded by the universities in Catalonia. *Coordinació* also maintain an extensive database of PAAU and COU examination grades and scores. The PAAU examinations have eight components, the seven subjects taken in COU (plus Catalan in Catalonia), and an essay on a prescribed topic. The examinations have a variety of formats, varying both from subject to subject and from board to board. For instance, an examination may consist of two sections of four open-ended problems, and students are given credit for solutions to the four items in the section they select.

About 97 percent of the students successful in COU sit the PAAU examinations. The success rate averaged over the subjects is about 92 percent, but there are substantial differences, especially among the four categories of the subjects. Since students have to pass every examination, the percentage of successful students is much smaller. Universities reserve a small number of places for students who have completed vocational training. The applicants for these places do not have to take PAAU, and decisions about their applications are based solely on their record during vocational training.

The items (problems) for a subject are prepared by an academic appointed by *Coordinació* and the students' papers (responses) are graded by raters appointed, trained, and instructed by the author of the examination and the members of *Coordinació*. Prior to the examinations, the problems are subjected to a peer review.

The examinations are organised within so-called *tribunals* consisting of a few high schools with contiguous catchment areas. A typical tribunal comprises schools with a total of about 200 examinees. At present, the raters are engaged only in their local tribunals. The students may ask for a revision of their grade, in which case the author, as a chief examiner, is involved in the arbitration. An informal process of feedback from the schools is in place; for instance, schools may propose new item types or various improvements in the presentation of the items.

The examinations in PAAU take one to one-and-half hours, and the eight or nine examinations are taken over two or three days. The students' papers are collected within each tribunal separately for each subject, and are distributed to the raters. Papers in subjects taken by very few students may be collected over several tribunals. The Ministry of Education issues only very general instructions to the raters, and these are often supplemented by more detailed instructions from the examination board.

Traditionally, each paper is graded only once. From the corresponding data (grades) it is not possible to assess whether the observed differences of the mean grades across the tribunals are due to different levels of ability or different (severity) scales applied by the raters. Another concern not attended to by this design is rater reliability. These issues are addressed in Section 'Rater reliability'.

The PAAU grades are combined as weighted totals of the components. The high-school record (containing COU as a component) is combined similarly. Finally, the criterion for admission to universities is based on the mean of the high-school record and the mean PAAU score.

Each university school (*facultad* in Spanish, such as a School of Computing Studies) has a given number of available first-year places. The students are allocated to schools in such a manner that each school will be assigned students with the highest composite scores, and each student will be enrolled in the school highest on his/her list of preferences, for which the attained score qualifies.

The examinations are in a continual state of change, so as to respond to changing demands of the university system and to changing high-school curricula, to improve the validity of the composite scores, and to promote equity in the whole university admissions process. In this paper, we discuss the main educational measurement issues associated with the examinations: reliability of grading and score adjustment, combining scores, basing decisions on scores which are subject to rater's judgement (measurement error), making inferences about board-to-board differences, and others. Our focus is on research concurrent with administration of the examinations and on continual improvement of the entire university admissions process.

So as to place the current research agenda in a historic perspective, we give a brief of past research of PAAU and COU.

## Past research

A large number of appointed raters are involved in grading students' papers in PAAU. The Ministry of Education imposes an upper limit on the amount of work that can be assigned to any one rater. In addition, there is a strict time limit in which all the examinations have to be graded. These constraints make it impossible for all the examinations to be graded more than once.

One of the earliest documented studies of PAAU scores is reported by Sans (1989). It is a descriptive study of students who took PAAU in a few adjacent tribunals in the city of Barcelona (the capital of Catalonia) in 1987. Different test forms were used each day, but all students from a school took the examinations on the same day. Sans (1989) carried out an analysis of covariance to explore differences among the tribunals and test forms, after adjusting for students' high-school records. Significant differences were found both among the forms and among the tribunals. However, since no replication was implemented, the extent to which the observed differences due to differential difficulties of the test forms and due to differences in the schools' mean academic abilities cannot be established. In addition, the raters may have altered their conduct from one day to another (presumably the same group of raters was engaged on every examination day), which would be another confounding factor.

Muñoz-Repiso *et al.* (1991) reviewed several studies and carried out an informal meta-analysis of the results in PAAU examinations in a few adjacent tribunals in Madrid, the capital of Spain, in years 1987–89. In the categories Science and Technology and Biomedical Sciences the mean PAAU scores tended to be lower than their high-school record counterparts. In the other two categories, Social Sciences and Humanities and Languages, the PAAU scores were also lower, but by a smaller margin. They attributed this finding to the fact that grading of the former two categories is more precise and more consistent. They pointed out the necessity for standardisation of the grading process.

Apart from their main objectives, these two studies highlight the need for assessing the reliability of the grading process. This could be done only if some form of replication in grading of the papers were implemented. Escudero and Bueno (1994) designed and analysed a study with replication of grading. All the examinations taken in one tribunal, involving 348 students, were photocopied and re-graded by a different set of raters. Comparisons were made mainly at an aggregate level, using analysis of variance. However, from our perspective, the most important outcome of the study is that, if the scores from the second (experimental) grading were used, decisions about admission to university would be reversed for about 33 students (9.5 percent).

Satorra and Udina (1993) carried out a small-scale quality-control study involving 20 PAAU examination papers in Mathematics. The papers were photocopied and sent for re-grading to several raters. Some raters failed to respond, but each paper was re-graded at least twice. Application of a measurement-error model led to the conclusion that about ten percent of total variation in the scores was due to discrepancies among the raters.

Further research on the issue of rater reliability, and possible schemes for adjustment of scores is in progress (Martí, 1995). The generalisability theory (Cronbach, 1972) is adopted as a framework, with the adaptation to the setting of examinations and raters due to Longford (1994 and 1995, Ch. 2).

Memoria de Actividades (1994) proposed several changes in the grading process with an emphasis on improving the rater reliability of both PAAU and COU scores, and for monitoring of the school-level differences between COU and PAAU scores. As a direct response, Cuxart, Graffelman, and Martí (1995) explored the association of the COU and PAAU composite scores for students from a random sample of schools in Catalonia. Using multilevel analysis (Goldstein, 1995; Longford, 1993), significant between-school variation was found, indicating that the scales (standards) of COU grades vary from school to school. In contrast, the within-school means of the scores vary much less, suggesting that some form of self-norming takes place in each school. The design of the study did not allow to distinguish between tribunal- and school-level variation (the majority of tribunals were represented in the study by at most one school).

In summary, the main concerns associated with the use of the examination scores have been reliability of the grading process and the between-school variation in the standard of the COU grades. In the following sections we discuss a more comprehensive agenda for research concurrent with (future) administrations of the tests. Additional issues are: the effect of the choice of sections in the examinations, definition of composite scores (weighing of the component grades), score adjustment due to imperfect reliability, and pretesting of items.

A related general area relevant to structured and regularly administered examinations is that of equating. It is relevant in several contexts: taking account of different difficulties of the items in alternative sections, making unimportant the choice of alternative examinations (e. g., of the two optional Science and Technology subjects), maintaining the same standard of the PAAU scores across the years and across examination boards. Such concerns are not specific to Spain; see Fitz-Gibbon and Vincent (1994) for a study comparing public examination standards across subjects in England and Wales.

## Rater reliability

The study by Cuxart (1996) explores the feasibility of an on-going system of monitoring the quality of grading. Examination papers for two subjects, Mathematics (187 papers) and Philosophy (363 papers), from students taking PAAU in June 1995 in two tribunals in Catalonia, were photocopied and re-graded. The examination papers were collated in batches of up to 20 papers and distributed at random to a set of raters recruited from the neighbouring tribunals (ten raters for Mathematics and 20 raters for Philosophy). Elementary summaries, such as cross-tabulations of operational and experimental grades, provide a clear evidence of between-rater differences. Also, there is much more agreement in the grading of Mathematics papers than of Philosophy papers. For instance, the pairs of grades differ by one point or less for 72 percent in Mathematics but only for 51 percent in Philosophy (the grading scale is 0–10). They differ by more than three points for 48 Philosophy papers (13 percent), but only for four Mathematics papers (2 percent). A statistically more profound way of summarising the quality of grading is in terms of the variances due to ability, severity, and inconsistency (Longford, 1994). In an ideal setting, the ability variance dominates the severity and inconsistency variances; poor quality grading corresponds to large variances due to severity and inconsistency. Severity refers to different standards of grading by the raters. The term 'inconsistency' stands for the amalgam of irreconcilable differences among the raters, their temporal variation, and other non-systematic influences. In analyses of several datasets from testing programmes administered by the Educational Testing Service in the U.S.A. (Longford, 1994, 1995, and 1996), inconsistency variance was several times greater than severity variance.

The observations based on the comparisons of the pairs of grades are confirmed by the more formal approach: the inconsistency variances are 13 and 34 percent of the total variance for Mathematics and Philosophy, respectively. Severity variance is negligible for Mathematics, but in Philosophy it is 6 percent of the total variance. If in the composite scores (mean of the PAAU grades and the high-school record) the contribution of Philosophy and Mathematics grades from the operational scoring were replaced by their conterparts from the experimental (second) scoring, decisions about admission to university would be reversed for eleven students (3 percent). Of course, if the grades for the other seven or eight subjects (Mathematics is not compulsory) were also replaced, the number of reversals would likely be much greater.

Although the outcomes of the study do not contradict intuition, generalisation to other subjects or examination boards, to future (or past) administrations, or to different examination formats, is not warranted. The study demonstrated the value of monitoring the quality of grading, as well as the feasibility of an on-going scheme of partial duplicate grading which would be used solely for quality control and for understanding the inconsistencies occurring in the grading process. If the large between-subject differences in the quality of grading were confirmed, efforts to improve grading could be concentrated on the subjects with the poorest quality of grading. If sufficient resources (experts and funds) were available, duplicate grading in these subjects could be introduced operationally.

An unavoidable limitation of the study is that the raters knew that their grading was not part of the operational grading. They were grading photocopies of the papers; the original papers were graded at the same time at a different location. In any case, the original papers would have been annotated by the 'operational' raters, and could not be re-used for grading.

The study implemented an interpenetrating allocation design. At present, the raters for the examinations in a tribunal are recruited locally, from the same tribunal. Concerns arise about the causes of differences among the tribunals. One explanation is that the mean proficiency of the students varies among the tribunals, but these differences may be confounded with different mean severities of the raters from the tribunals. The confounding could be resolved by assigning examination papers for grading by raters from outside the tribunal. This raises substantial logistic problems, but various compromise solutions may be adequate. For instance, a small number of neighbouring tribunals may form a group which would share the grading of the examinations. These groups may overlap and be reconfigured from year to year. Also, the grouping arrangements may differ from subject to subject. If such a scheme were implemented operationally, comparisons across tribunals would be more meaningful. The maintenance of the standards of grading could be further enhanced by re-grading small samples of examination papers outside this group.

## Prospects of score adjustment

If the raters' conduct has a systematic element, such as some raters being consistently stricter than others, validity of the grades may be improved by compensating for such differences. This issue has been addressed by Longford (1994) who described several schemes for score adjustment. In fact, the adjustment schemes fulfill a more general function; they also protect against large errors by 'shrinking' the raw (original) grades toward the mean. The motivation for this is that in the case of very poor quality grading each student would get the same score, rather than being exposed to the total arbitrariness of the grading process. The adjustment schemes based on shrinkage estimators do not eradicate the sources of grading error, but reduce their impact.

Implementation of an adjustment scheme requires specialised software and its integration in the process of reporting the scores. Such a scheme could be improved if the performances of the raters were recorded over the years. For instance, if the raters' severities were maintained over the years, their performances in the past could inform about the adjustments to be applied in future. The approach is not without transparent flaws: for instance, if a student writes a perfect paper but the paper is graded by a lenient rater who (rightly) awards the highest grade, score adjustment would reduce this grade. In brief, no student graded by a lenient rater could get the highest score. In general, such a flaw is more than compensated by the overall improvement of the scores, especially since the principal concern is about the scores much further away from the extremes of the scale.

Although the weights of the component grades in the composite score are set *a priori*, they ignore the uncertainty associated with each component. Longford (1997) showed that such composite scores can be estimated more efficiently when the measurement- or grading-error variances are known or are estimated. The gains of this approach over the trivial method are substantial when the hypothetical 'true' grades are highly correlated and the grading-error variances differ a great deal. To motivate the approach, suppose the grades for two subjects, A and B, are to be combined, with equal weights. If A is graded with much more precision than B, and the true grades for A and B tend to be very similar (a student who is good in A is very likely to be good also in B), a combination with a larger weight for the grade A estimates the composite score (A+B) more efficiently.

Even though the grades for each subject are on the same scale, the distributions of the grades differ from subject to subject. For instance, extreme grades (0–1 and 9–10) are much more frequent in Mathematics (14 percent of the papers in the study) than in Philosophy (8.5 percent). Arguably, such differences should be taken into account when setting the weights.

Most schemes for combining scores pay equal attention to the entire range of scores. For PAAU scores and their combinations, precision at the extremes (for students with very low or very high ability) is less important than for students who are on the borderline for admission to the desired university school. If the combinations of the scores were defined with a focus on reducing the uncertainty about the borderline cases, the purpose of the scores could be taken into account more effectively.

## Score equating issues

It is desirable that any given grade, in PAAU or COU, have the same meaning, irrespective of the subject, examination board, or the year of the examination. There is no truly external reference for the score scale; if, for instance, all the scores were multiplied by 10, or transformed in a similar manner (linearly), their meaning and validity would not be affected. However, the associations among the scores would be altered if the transformations applied were specific to the subjects and/or boards (years). Such transformations can, in principle, make scores on different subjects, in different boards or years, to have a uniform meaning, that is, refer to a universal standard. This may not be straightforward for arrange for technical reasons such as problems with estimating the appropriate, not necessarily linear, transformations and because the standard has to apply for very disparate subjects. Methods and procedures for transforming grades or scores so that they would be close to such equivalence are generally referred to as *equating*. For background, see Holland and Rubin (1982).

Since high schools have a considerable autonomy in the construction of the COU examinations, a variety of formats, item types, and grading schemes are used. Therefore, a general discussion of equating issues in COU is not feasible, and so we focus on equating in PAAU. However, we conclude this section by discussing the association of PAAU and COU scores, which has a relevance to studying regional differences.

Maintaining standards from year to year is a well-established goal of many national educational assessment systems. In Spain, an important contributor to this goal is the uniform interpretability of the PAAU scores across the years. At present, examinations are set by experts who use their professional judgement about the difficulties of the examination items. They consult examinations from previous years and the proposed items are reviewed by a committee, but there is no empirical check on the quality of the expert judgement. Simple approaches could be applied to compare the distribution of scores in each subject from one year to another. A potentially serious flaw of such approaches, for optional subjects in particular, is that examinations in each subject may be taken by different subpopulations of students. For instance, over the years Biology may become more popular among high-ability students; then, assuming unchanged difficulty of the examination, the mean scores may increase over the years. However, substantial drifts in examination-taking patterns within the span of a few years are unlikely.

Since PAAU comprises many subjects, it can be argued that the small yearly deviations in the difficulties of the examinations may even themselves out. On the other hand, an overall (grade-inflation) trend common to all examinations may be present. In addition, if the standards of COU are slipping, more low-ability students are admitted to take PAAU, and so lower mean scores would not necessarily be an evidence of more difficult PAAU examinations. Some insights can be gained from the comparison of numbers of students taking PAAU with the size of the population cohort.

Since the composite PAAU score is of central importance, the issue of weighing of the constituent subjects cannot be ignored. For instance, if particular subjects were identified as being easy, informed students would tend to select them, thus subverting the validity of the scoring system. On the one hand, disparate subjects, such as Mathematics and Spanish, are inherently incomparable; on the other hand, for related subjects, such as the options in one of the four categories (Science and Technology; Biomedical Sciences; Social Sciences; and Humanities and Languages), a common scale is meaningful.

This issue of equating across subjects is technically resolvable, owing to the compulsory subjects within the categories. For instance, in the category Science and Technology, each student sits examinations in Mathematics and Physics. The grades for optional subjects can be equated by conditioning on the grades in these compulsory subjects. The solution is not so straightforward in other categories within which the subjects are much less comparable. For example, in category Social Sciences the compulsory subjects are History and Mathematics (different from the examination of the same name in Science and Technology).

A similar problem of equating arises for students who select subjects from different categories. Then the universally compulsory subjects (Spanish, Philosophy, a foreign language, and an essay) can be used for comparison (as *anchors* in the terminology of equating). The flaw of this approach is that the compulsory subjects may have uneven affinity to the four categories of optional subjects. Preliminary studies indicate, for instance, that the highest-ability students tend to be overrepresented in the category Science and Technology (Cuxart, unpublished). Fitz-Gibbon and Vincent (1994) arrived at a similar conclusion about the high-school examinees in England and Wales.

## Choice

In order to cater for different preferences and strengths of the students, in most subjects, examinees have a choice of two parallel sections (say, two sections comprising four problems each in Mathematics). The sections are graded according to the same set of rules, and no adjustment is made for differing difficulties of the sections. However, such an adjustment would be very difficult to make because the sets of students who select the sections are not necessarily comparable. For instance, higher-ability students may choose section A more often than lower-ability students, in which case section A would appear to be easier. The choice, as a feature of the examination, is important because it reduces the impact of the context of a small number of items selected for the examination.

Choice is not a uniformly desirable feature. In some subjects, comprehensive mastery of the curriculum is required. For instance, if a student copes well with Arithmetic, but is not proficient in Algebra, he/she should not pass the examination in Mathematics. On the other hand, in History, some form of specialisation or emphasis on part of the general curriculum should be promoted (e. g., history of Spain vs. World history, or 20th century history vs. medieval history). However, even in Mathematics, the performance of an examinee may depend on the context of the problem, which is incidental to the ability in the subject. Therefore, choice improves the chances that no student will come across several problems with unfamiliar context. Obviously, the strategy engaged in selecting the section is important. Examinees who ignore or misjudge this element of the examination may not present themselves in the best light.

When each section contains only a small number of items, it is difficult to arrange that the two sections be balanced with respect to several item specifications. For instance, the formats of the questions should be similar, important elements of the curriculum should be represented evenly in the two sections, the range of scores that can be awarded should be the same, and the difficulties of the items should be similar. The latter condition is particularly problematic when the difficulty of each item is assessed only subjectively, by an expert. Also, the sections are unlikely to have similar difficulties purely by chance when they contain small numbers of items. Yet another issue is that papers in the two sections may be graded with unequal precisions.

In ideal circumstances, students would be compensated for having selected the more difficult section of the examination. This presumes that the difficulties of the sections would be established, that is, the scores for the two sections of an examination would be *equated.* An equating procedure would be relatively straightforward if the examination had a common section; the score on this section could be used as the anchor. Even without a common section, we can use as anchors the scores in the compulsory subjects in the category or in the universally compulsory subjects, but these anchors are not as effective.

### Pretesting

We have seen that difficulty of the items plays a central role in defining equitable scores. Reliance on non-empirical assessment of difficulty by experts is problematic. The quality of such a judgement could be gauged by comparing experts' subjective judgements made prior to the examination date with empirical results from the examinations. A more complex scheme would involve organising the following system of continual item pretesting: items would be written one or two years before their operational use, and they would be pretested as part of the operational examination by a different examination board. In this way, a few years after introducing this system, each examination would contain several items which have been pretested and only a few (or only one) item which is being pretested. It would be necessary to keep confidential the location where items for future examinations are being pretested. Different boards can be used for different subjects, and even individual items can be pretested by different boards. Of course, this raises a host of logistic issues, but the benefits would be higher-quality items and tailor-made composition of the examinations. Similar schemes for interpenetrating design of rater assignment, possibly even for item construction and review, but generally for every aspect of the examination, would greatly enhance the equity (standardisation) and validity of the PAAU scores.

### Geographical variation and association of COU and PAAU

Since the PAAU scores are (intended to be) standardised the association of COU and PAAU scores can be interpreted as differences in the standards across geographical units (boards, tribunals, and schools). There is some rationale for standardising the COU scores as well, so that their secondary uses would be meaningful. This is difficult to enforce throughout the country, but systems could be put in place to encourage outlying units to adjust the process of grading COU examinations.

Cuxart, Graffelman, and Martí (1995) fitted regression models with random effects to assess the strength of association among the two sets of examination scores, and to explore its between-school differences. The study used data from a random sample of schools in Catalonia, and its important outcome was identification of schools in which the association of COU and PAAU differed a great deal from the average. In a future operation, the identified schools could be informed, so that they could investigate whether a change in the grading standards is the cause of outlying. In principle, the COU scores can be adjusted for differing standards, but this is possible only after disentangling the sources of regional differences. However, such level of uniformity may not be desired, and any adjustment scheme is likely to be controversial. The COU scores are unduly affected by rounding. In each subject, students' final scores fall into four categories. The corresponding rounded scores are then averaged. Less information would be discarded if the original scores were averaged and, perhaps, these averages were rounded at the end.

The proportion of students who pass the various examination hurdles is the most important aspect of geographical differences. In recent years, the percentage of students passing COU has been increasing, but this is difficult to attribute to the competing causes, such as lower grading standards, improved strategies in the selection of subjects, or an improvement in exam taking or in the education in general.

It is important that the examination scores be used only for intended purposes. In particular, prominent reporting of various aggregate scores may subvert certain elements of validity of the scoring system, and thus undermine their undisputed status as measures of suitability for university admission.

## Conclusions

The educational system in Spain is undergoing a reorganisation. In the new system, the minimum school-leaving age will be raised from 14 to 16 years. The streaming to vocational and academic education, which has taken place at the age of 14 will now be postponed till the age of 16. The motivation for the reform is to extend formal education of the lower-ability students, to encourage them to conclude their education with a formal examination, and to enhance the diversity of career choice. Although cohorts of 18-year-olds will be less numerous in the coming years, a higher proportion of students are expected to complete secondary education.

The COU (or its successor) and PAAU will be taken at the age of 18, as at present. In the new system there will be less restriction in the choice of the examination subjects, thus catering for the diverse interests of the students and employers. The format and content of the PAAU examinations will be updated. A working party organised by the Inter-university Council of Catalonia recently submitted a report (Ferrer, 1996), which compares the higher-education systems in several EC countries and the U.S.A., and discusses the advantages of implementing various features of these systems in a prospective reform in Spain.

At present, the sole purpose of PAAU is for admission to universities, although, informally, it is also used as a standard for high schools. In the future, this secondary role may be formalised. Goldstein and Spiegelhalter (1995) focus on technical issues of institutional comparisons, but they point out that the uncertainties associated with such comparisons and the uncritical interpretation of the ranks produced by formal analyses greatly undermine their value.

Although we have presented a long list of suggestions for improving the examinations and their scoring, implementing them requires further research and careful adaptation to the specific circumstances. Any changes have to be introduced gradually, so that their impact can be monitored without being confounded with other innovations. Best intentions can be eroded by incomplete understanding or uneven interpretation of the new instructions by the personnel involved, as well as by insufficient quality control of the entire process. Our proposals may seem largely orthogonal to the changes in the educational system, but there is an added impetus for their consideration as the existing system is subjected to closer scrutiny.

Our general emphasis is on research concurrent with the operation and on empirical evidence about the properties of the examination items and scores. Since the system of examinations is dynamic, involving changes in the population of students, in the curricula, in the demands of the economy and society, and in the expertise of the raters and other personnel, any statistical conclusions have a strictly temporary nature, and so their generalisability is doubtful. Also, apart from the major reform, the system is continually undergoing minor changes which have subtle and difficult-to-assess effects.

Some elements of our proposal are gradually being implemented. For instance, in some geographical areas essays in a few subjects are graded by two distinct raters, and the discrepancies among the raters are analysed. In addition to resources, a key to the success of the effort is consultation with all the parties involved (raters and exam coordinators in particular), so the aims of the effort are well understood and supported. In selecting the priorities in implementing other components of a comprehensive monitoring system, this aspect has to be given a prominent consideration.

## References

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

Cuxart i Jardí, A. (1996) 'Los sistemas de corrección de las pruebas de Selectividad en España', unpublished draft report.

Cuxart i Jardí, A., Graffelman, J. and Martí, M. (1995) 'La nota PAAU y su relación con la nota COU: un modelo de regresión con coeficientes aleatorios para el estudio del efecto centro en la nota PAAU', in *Actas de la 5 Conferencia Española de Biometría.* (Valencia, Spain).

Escudero Escorza, T. and Bueno Garcia, C. (1994) 'Investigaciones y experiencias: Examen de Selectividad. El estudio del tribunal paralelo', *Revista de Educación*, 304. Madrid, Spain: Ministerio de Educación y Ciencia.

Ferrer i Juliá, F. (1996) *Els sistemes d'accés a la Universitat des d'una perspectiva internacional.* Barcelona, Spain: Generalitat de Catalunya, Consell Interuniversitari de Catalunya.

Fitz-Gibbon, C. T. and Vincent, L. (1994) 'Candidates' performance in public examinations in Mathematics and Science', a report commissioned by SCAA from the Curriculum, Evaluation and Management Centre, University of Newcastle upon Tyne. London: School Curriculum and Assessment Authority.

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

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

Holland, P. W. and Rubin, D. B. (1982) *Test Equating.* New York: Academic Press.

Longford, N. T. (1993) *Random Coefficient Models.* Oxford: Oxford University Press.

Longford, N. T. (1994) 'Reliability of essay rating and score adjustment', *Journal of Educational and Behavioral Statistics*, 19, 171–200.

Longford, N. T. (1995) *Models for Uncertainty in Educational Testing.* New York: Springer-Verlag.

Longford, N. T. (1996) 'Adjustment for reader rating behavior in the Test of Written English', TOEFL Research Report No. 55. Princeton, NJ: Educational Testing Service.

Longford, N. T. (1997) 'Shrinkage estimation of linear combinations of true scores', *Psychometrika*, 62, 237–244.

Martí i Recober, M. (1995) 'Los sistemas de corrección de las pruebas de Selectividad en España. Análisis y propuestas', Grant from Concurso Nacional de Proyectos de Investigación Educativa. Ministerio de Educación. Madrid: Spain, CIDE.

'Memoria de Actividades del Consejo de Universidades, Junio 1991–Julio 1993'.

Muñoz-Repiso Izaguirre, M., *et al.* (1991) 'Las calificaciones en las Pruebas de Aptitud para el Acceso a la Universidad', *Colección Investigación*, 61. Madrid, Spain: CIDE.

Sans, A. (1989) 'Fiabilidad y consistencia del proceso de Selectividad', in *Actas de las jornadas 'La investigacion educativa sobre la Universidad'*, 201–208. Madrid, Spain: CIDE.

Satorra, A. and Udina, F. (1994) Personal communication.