Int. J. Engng Ed. Vol. 00, No. 0, pp. 1–10, 2008
Printed in Great Britain.

0949-149X/91 $3.00+0.00

Oral Presentation and Assessment Skills in Engineering Education*

JUAN A. MARÍN-GARCÍA and CRISTOBAL MIRALLES
Polytechnic University of Valencia, 46 071 Valencia, Spain. E-mail: jamarin@omp.upv.es
MARIA-PAZ MARÍN
Department of Spanish and Portuguese, University of Cambridge, Cambridge CD3 9BA, UK

This study examines the possibilities of including students in the assessment process and the conclusions drawn from the experiences of the academic year 2004–05. The fieldwork focuses on examining the reliability of the marks given by students in relation to those given by the lecturer. Similarly, the marks given by the subject lecturer are examined with those assigned by four external markers. In each presentation the assessment of several markers is also studied, whether they are students or lecturers. The results obtained seem to indicate that students can be adequate markers for their peers' oral presentations. Also in their assessments, few significant differences are seen from those suggested by the subject lecturer.

Keywords: oral presentations; agreement among markers; peer assessment; reliability of the assessment

INTRODUCTION

UNIVERSITY EDUCATION CURRENTLY faces a series of different challenges. One of these is the role that the assessment of students' learning can play and the importance of developing certain professional competencies [1–6]. One way of merging these two aspects is to encourage the active participation of students in the assessment process of their own learning or that of their peers [7]. Nevertheless, traditional teaching does not cater for this possibility and assessment falls on the lecturer, whose marks are deemed valid and indisputable [3].

However, it is increasingly usual to acknowledge the importance of educating university students to develop the necessary skills for thinking about the results of their work critically and the process involved in completing it [8, 9]. Not only this, but it is also important that they are able to assess the work of their peers [4]. This skill will be useful to them in their future profession, particularly if the company where they are employed encourages team work [10–12].

Studies on peer assessment usually focus on essays, group projects or oral presentations [1, 13] and, in some cases, poster presentations [14]. Since widely different activities have been studied, it is hardly surprising that results are divergent.

The number of studies carried out up to now has been quite limited and it would be recommendable to offer more experiences related to oral presentation assessments [15] in different disciplines [8, 16]. It also appears necessary to include new ways of studying data that are more practical and enable integrating the conclusions of different studies [8, 17].

To this end, we have studied ways of including students in the assessment process and the conclusions drawn from previous studies. We have only taken into account research related to the assessment of oral presentations. This will be the activity carried out in the fieldwork. We will also summarise the recommendations put forward for improving student participation in assessment. We will conclude the review of literature by highlighting what statistical analysis was used to study the reliability of marks given by students, the results obtained in other studies, and the limitations of these results. Our fieldwork has focused on studying the agreement of student marks by comparing them with those of the subject lecturer. However, we have also compared the subject lecturer's marks to the marks of four other lecturers.

Students as assessors

There are three ways of including students in the process of assessing their academic performance [1, 5, 18].

- Self-assessment: Consists of giving students the opportunity to assess their own academic performance and, in particular, the products and results of their learning.
- Peer Assessment: Consists of a process by which a group of individuals mark their peers.
- Collaborative Assessment, Participative Assessment or Co-assessment: Consists of providing the opportunity for students to assess themselves or their peers, but at the same time enabling the lecturer to intervene and check the final mark submitted.

* Accepted 17 January 2008.

None of these three methods (Self-assessment, Peer Assessment, Co-assessment) is new, since they have all been used in university education for many years and have been the aim of frequent studies since the 1980s [16, 19]. The main areas these studies have focused on concern the use of these procedures as formative assessment or summative assessment; the reliability of the students marks; the fairness of marks given by students; the effects on students when participating in the assessment of their academic performance; what methods may be used and how these processes may be enhanced [1, 16].

In fact, these study aims are closely linked. If these assessment methods are beneficial to students' academic performance, their use as formative assessment is justified. However, it would not be particularly logical to include student marks to summative assessment if the reliability or fairness of these marks was not considerable. Furthermore, students usually place more importance on factors that affect their final mark [13]. Consequently, it is important to identify what methods and under what circumstances we can improve the quality of student marks so that they can, to a certain extent, be included in summative assessment.

We will now make a quick review of the conclusions drawn from studies published on the subject. There appears to be a certain degree of consent about the advantages provided by Self-assessment from the point of view of formative assessment. Many of these advantages are also shared by Peer assessment. Several authors [1, 5, 8, 12] summarise some of these advantages: students are more confident about their skills; it improves the student's perception about the quality of his work; students think more about their performance and their results; it improves students' exam results; it improves the quality and effectiveness of academic performance; students take more responsibility in the learning process; it enhances students' satisfaction. Self-assessment is also very useful in helping students achieve their learning goals [20]. It helps students to be more critical, work in a structured manner, and it encourages them to think more [8, 14]. On the other hand, it is also believed that helping to develop students' skill at assessing their own work and those of their peers is an important factor in the learning process [8, 21]. Indeed it is one of the professional skills for which we should train our university students if we wish to educate thoughtful professionals and encourage learning 'throughout life' [1, 10, 15, 22].

On reviewing previous studies, when students' marks are compared with those of lecturers', it appears that the peer assessments are more similar than those of self-assessment [23]. This is probably because in Peer assessment the marks of several students assessing a peer are usually averaged out before comparing them with those of a lecturer [24]. In any event, the major problems mentioned with regard to agreement between lecturers' and students' marks are applicable to both types of assessment by students. These problems are the following.

- The mark can be negotiated or better marks can be given to friends or oneself than to others [1, 13, 24].
- Students do not usually use the whole range of the scale available and tend to concentrate their marks in the middle range [1, 10].
- Students do not have the same reference level as lecturers [13].
- Students are bad assessors of the activity to be marked and their marks do not tie in with those of lecturers [1, 10, 13].

The reluctance to include students' marks in cumulative assessment due to the problems mentioned above is not merely because of lecturers. Some students do not trust the quality of marks given by their peers or themselves and are not comfortable with the system [8, 10, 25].

Moving on to the following aspect under consideration, we can confirm that the recommendations put forward by different authors for improving the process of Peer assessment, Self-assessment or Co-assessment, are completely coherent. In this connection, the following points are advisable when setting up these systems:

- Assessment is a skill that can be improved by repetition and training [1, 8, 10, 17, 19].
- It is necessary to set aside time so that the student can carry out marking. It is also possible that students may need support or advice at some point during the process [1, 17].
- If we wish to use Self-assessment or Peer assessment as a formative tool, we should ensure that students regard this activity as a tool to make their learning easier [1].
- Use the same criteria for all markers. These should be decided in advance so that students are familiar with them. They should be set out in a simple, straightforward manner. If possible, negotiating with students is recommended [1, 8, 10, 14, 15, 17, 19, 22, 25–27].
- Although the individual marks of a person are not reliable, when the marks of several people are averaged out, reliability is extremely high [13].
- Use Peer assessment—which tends to be more reliable than self-assessment—as part of the mark making up the overall mark submitted by the lecturers [1, 8, 17, 19].

To sum up, it appears that the major area of disagreement in studies lies in the degree of reliability of student marks. If this aspect were improved, it would help to reduce some of the reluctance by lecturers and students to adopt these systems. Consequently, the aim of our work will be to study the degree of reliability obtained in the specific implementation of Peer assessment (the assessment of oral presentations), when taking

into account the recommendations put forward on this subject by scientific literature.

Coherence between students' and lecturers' marks in Oral Presentations

Different measures have been used to establish the degree of coherence between students' and lecturers' marks. The most standard measure is to use the correlation rate between the mean of students' marks with the lecturer's mark (or the mean of lecturers' marks if there are several lecturers) [8, 10, 13, 15, 22, 28–31]. Other measures used are: the percentage of students that give a mark included in a confidence interval compared with the lecturer's mark (usually a standard deviation) [8, 30, 32], compare the variance of students' with lecturers' marks or a t-test for the difference of the means between students' and lecturers' marks [8, 13, 22, 23, 30]. Some studies have used a measure linked to variance analysis (ANOVA) to determine reliability among markers [17].

We can confirm that the degree of coherence between average marks given by lecturers and those of the student group are quite high when oral presentations are assessed. On the one hand, correlations between students' and lecturers' marks are moderate or high. Values lie between 0.44 and 0.79 in studies by several authors [8, 10, 13, 22, 29, 30]. They are above 0.80 in three studies cited by Falchikov and Goldfinch [29] and in two other studies [15, 28, 31]. Furthermore, there are no significant differences between lecturers and students when a t-test is applied [8, 30]. Nevertheless the dispersion of student marks is appreciably less than the dispersion given by lecturers [8, 13, 22, 30]. In short, students tend to concentrate their marks and differentiate less than lecturers.

We must bear in mind however, the warnings by Ward *et al.* [23]. First, all of the studies consider that the marks given by the lecturer are correct and that the difference in marks given by students is due to the fact that they are not as good as lecturers in assessment tasks. However, this difference can also be due to the fact that lecturer marks may not be as valid and reliable as one would normally assume [3, 13, 29, 33]. Suggestions about how to deal with this problem concern improving the reliability of lecturers' marks by using an average of the marks given by various expert lecturers [23, 33].

Secondly, even though there is a series of criteria for assessment, there is no guarantee that all of the individuals taking part in the assessment will interpret them in the same way. The best way to avoid this problem is to offer various criteria that are as simple as possible and to include explicit guides in the assessment form [23, 33, 34]. Another possibility would be to standardize the student marks or those by the lecturers before working out the averages.

Lastly, Ward and collaborators [23] also commented on a possible problem occurring when data are analysed at a group level: in other words, when we compare the average of marks of a group of markers with the marks given by an expert. This kind of analysis does not indicate the degree of agreement of each individual student with the mark given by the lecturer; furthermore it does not give us relevant information when we only have a mark from students (either because they assess themselves or because each piece of work is assessed by a colleague only).

Our aim in this paper is to apply the first two recommendations and we will leave issues of assessment of different levels of analysis for a subsequent study.

Objectives

Bearing in mind the aforementioned research, we think it is worthwhile to offer further data in order to establish to what extent are marks given by students reliable. This way we try to reflect on the possibility of incorporating them to the summative assessment of university subjects.

The questions, which we seek to answer in our research, are the following.

1. Do the marks given by one lecturer agree with those given by a group of lecturers?
2. Does the average of the marks given by several students assessing the same presentation agree with the mark given by the lecturer?
3. How many student markers would be necessary in order to attain a level of reliability similar to that of a lecturer?

Methodology

In order to answer the questions posed by the study, we shall use the following procedure.

- The degree of agreement of the marks given by one lecturer shall be measured by comparing the marks given by the subject lecturer with the average of the assessments given by the other four lecturers participating in this study. Given that it was not possible for more lecturers to attend the final end-of-term presentation, we shall use the data from assessments of video-recorded presentations, both for the subject lecturer as well as for the lecturers collaborating in the study; hence the situation is common to all concerned. With these data we shall calculate the correlation and do a t-test to ascertain if the assessments correspond to the same population. We shall also calculate how many assessments by the subject lecturer differ from the average of the other four lecturers, by less than standard deviation [8, 13, 22, 30].
- We will attempt to prove the agreement of the average of the student marks by way of a similar procedure. But, in this case, the reference value will be the marks given by the subject lecturer on the day of the oral presentations.
- Finally, we will complete the analysis with the estimation of the number of students or lecturers that should be simultaneously assessing an oral presentation so that their reliability is similar to

the assessment of a sole expert lecturer (both in the subject taught and in the assessment of the oral presentations); to this end, we shall follow the procedure put forward by Magin [17] (see Appendix A).

Kane and Lawler [35] suggest three procedures for assessment by colleagues: ranking, nomination and grading. We have chosen the third one; in this situation each student assigns a mark to his or her colleagues in consideration of his or her performance by using an assessment scale. We consider this to be the easiest option to apply (even though the other methods might be more reliable or allow for greater discrimination).

In order to encourage greater student participation, we decided that the criteria to be assessed ought to be selected by the students themselves and are set forth later in this paper. The task of the lecturer was to integrate the visions of the different groups and to come up with a definitive version of the evaluation form, including a series of guides for the assessment of each criterion. The evaluation form contained nine criteria and each one was given a mark between 0 and 3. Hence, the maximum mark for a presentation was 27 points.

The students who were the object of the study are enrolled in one of the two classes of the third year course of the degree in Industrial Organisation Engineering. None of the students had previously participated in peer assessment activities. The subject (Management) is taught over a 15-week term with 2.5 hours of classes weekly. The activity evaluated was carried out in pairs. It consisted of interviewing two company managers, comparing their answers with the theory taught in the course and presenting the result of their work to other colleagues in class. This presentation was video-recorded. The presentation took place on the last day of class and was not compulsory, even though it accounted for 10% of the final mark in the subject—5% was for the presentation (average of the marks given by the colleagues and the lecturer) and 5% was for the degree of agreement with the marks given by each student when it was compared with the marks average for all of the students). There were 23 presentations and 44 students took part in them.

One month before the presentation, the activity was explained in class. Two groups with 15–20 students each were set up. They met to establish the criteria for giving marks in the activity. First of all, there was a brainstorm with a nominal group focusing on aspects which defined a good oral presentation. Afterwards, the criteria were filtered (using the affinity diagram technique) with the premise that the criteria they came up with would have to be easily understood and objectively observed by the students who would act as markers. The brainstorm and the affinity diagram are two techniques contained in the programme of the subject and these were practised in these meetings. The lecturer gathered up the two lists and summed them up after the meeting. Some guides to facilitate the marking were also given.

Three weeks before the oral presentation, the final list was given to the students (see Appendix B). Furthermore, students were given the possibility of watching the video-recordings of class presentations on the WEB. These presentations were independent of the activity that is the object of this paper and consisted of the performance of various students (almost 50% of those attending the class) who had acted as spokespersons for group activities in class. The objective was that during the week, students would assess themselves and two other colleagues using the evaluation form. By doing so, they would obtain feedback on their skills as speakers and familiarise themselves with the use of the evaluation form. In the following class, the students were given the opportunity to talk about the difficulties they found in using the evaluation form. No further clarifications or corrections of the form were necessary.

During the following two weeks, the students conducted their own interviews. On the day of the oral presentation, the students were given one hour of class time to prepare their presentations (with a maximum duration of 3 minutes). During the second hour of class time, the presentations were given and then marked by their colleagues and the lecturer.

With the aim of not taxing the students and so that they would be able to pay attention to the content of the presentations, the students marked only one of every four presentations. The assignation of which presentation they would evaluate was based on the seat they occupies in tables (each table had benches to seat four people), hence we can consider the assignation to be random. The lecturer marked all the presentations using the same evaluation form as the students.

Four months after the presentation, a group of lecturers met with the subject lecturer to evaluate the oral presentations that had been video-recorded. Of the four lecturers, one taught the same subject but at another teaching institution; another belonged to the same department, while the other two belonged to a different department.

We gathered up four sets of data:

1. the marks given by students for presentations delivered by their colleagues (each presentation was assessed by 7–12 students) (AA);
2. the marks given by the subject lecturer for presentations delivered by the students (PP1);
3. the marks given by the subject lecturer for the video-recorded presentations (PPV);
4. the marks given by four lecturers for the video-recorded presentations (PGP).

Analysis and discussion of the results

In Fig. 1 we presented a graph of the marks obtained for each of the presentations given. We have calculated the average in that series of data where there was more than one mark given per

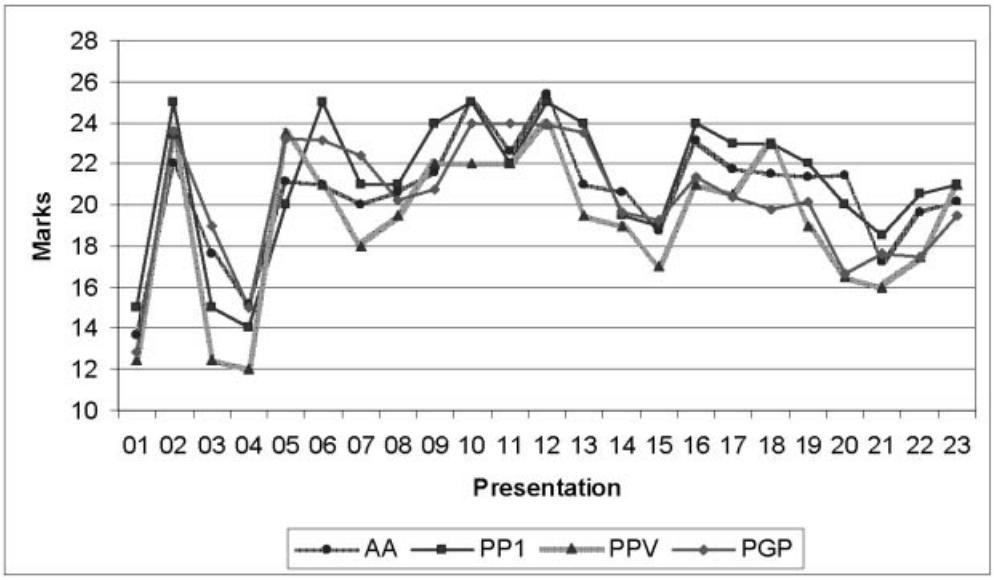

Fig. 1. Marks obtained in the presentations. AA: average of student marks on the day of the presentation. PP1: marks given by the subject lecturer on the day of the presentation. PPV: marks given by subject lecturer a for the video-recorded presentations. PGP: average of marks by four lecturers for the video-recorded presentations.

presentation (AA & PGP). A cursory glance reveals that there is considerable similarity between the different sources of data; however, we will carry out a series of statistical tests to give scientific rigour to this interpretation.

We first calculated the descriptive statistics and we have confirmed, by way of the Kolmogorov–Smirnov test for a sample, that the four series of data (AA, PP1, PPV and PGP) have a normal distribution [36]. We have summarised in Table 1 Pearson's correlations between variables.

The marks given by the subject lecturer for the video-recorded presentations (PPV) show a high correlation with the average marks of the other four lecturers (PGP) (0.808). The differences between the means of both assessments are significant with a t-test ($\alpha = 0.02$). This indicates that, on average, the marks by the subject lecturer are 1.08 points less than those given by the other lecturers. There are a total of four presentations (17%) where the difference between the marks in the PPV and PGP series is greater than 3 points (a standard deviation of the PGP series).

Similarly, the correlation between the average of marks given by students (AA) and that of the subject lecturer, the day of the oral presentations (PP1) is significant and very high (0.875). The differences between the means of both assessments are not significant with a t-test. Overall, there was only one presentation (4%) where the AA assessment differs from PP1 by more than one standard deviation of the PP1 assessment (3.27 points).

Hence, with respect to the first two questions posed in our study, we can conclude that the marks given by a single lecturer match up adequately with those given by a group of four lecturers. The average difference between both marks, though statistically significant, is only 1 point over a total of 27. Furthermore, in 83% of cases the difference between the PPV and PGP marks is

Table 1. Descriptive statistics

	N	Mean	Mean deviation	Min.	Max.	AA	PP1	PPV	PGP
AA: average of student marks on the day of the presentation	242	20.538	2.7155	13.7	25.4	1	0.875(**)	0.856(**)	0.811(**)
PP1: marks given by the subject lecturer on the day of the presentation	23	21.152	3.2768	14.0	25.0	–	1	0.878(**)	0.791(**)
PPV: marks given by subject lecturer a for the video-recorded presentations	23	19.239	3.5543	12.0	24.0	–	–	1	0.808(**)
PGP: average of marks by four lecturers for the video-recorded presentations	87	20.326	3.0168	12.8	24.0	–	–	–	1

** The correlation is significant at the 0.01 level (bilateral).

less than 3 points on a scale of 27 points. We do not have data from similar studies that have concerned themselves with these aspects; hence we cannot evaluate these findings in comparison with other findings.

Similarly, and more significant, the concordance of the averages of marks given by students to their colleagues with the marks given by the subject lecturer is practically complete. Our results coincide with previous studies which found high correlation between the marks given by the students and those given by the lecturers [10, 13, 15, 28–32]. Our data also fall in line with all of the studies that found that the dispersion of marks given by students is lower and that they differentiate less [8, 13, 22, 29, 30]. Although, in our case, the differences in dispersion are quite small. Lastly, the marks given by students on their colleagues' performance are slightly less than those given by the lecturer.

Once we have analysed the questions concerning agreement, we will comment on the results related to the reliability of the marking by students and lecturers. In both cases we will follow a similar procedure.

We will firstly calculate the reliability between markers (r_{nn}) of the group of students who assess their colleagues or the group of lecturers who watch the video-recorded presentations, without including the marks given by the subject lecturer. We will later add to the analyses, the marks given by the subject lecturer in order to establish the level of reliability (r^*_{nn}). Lastly, if the estimated reliability of the subject lecturer (r_{tt}) is higher than the estimated reliability of the students, or that of the whole set of lecturers collaborating in the study (r_{11}), we will calculate how many students or lecturers in other disciplines (DeltaN) would be necessary to assess each presentation so that the average of their marks would have a reliability similar to the subject lecturer's.

The results of these analyses (Table 2) could not be more eloquent. The reliability of the student-given marks ($r_{nn} = 0.90$) is quite high when there is a large number of markers (in our case, more than 10 per presentation). Nevertheless the estimated reliability of the marks given by one student marker, is somewhat low ($r_{11} = 0.47$), but almost the same as the reliability of marks given by just one of the four lecturers assessing the videos.

However, the estimated reliability of marks given by the subject lecturer are appreciably greater ($r_{tt} = 0.66$) and practically constant if we compare it to both sets of data. The reliability of the marks given by students as well as that corresponding to the subject lecturer respectively (r_{nn}, r_{11} *y* r_{tt}) is higher than the values obtained in the study by Magin and Helmore [13]. Nevertheless, the reliability of the marks of the other four lecturers is virtually the same as that found in that study.

We consider that there may be several reasons why the reliability levels are so high. First, the fact that we have included fairly objective criteria, and that each of these criteria had assessment guides incorporated. We acknowledge that we have left subjective aspects aside that are also important in the assessment of public presentations (for example: Is the presentation enjoyable? Does it capture the listener's attention? Is it interesting? . . .) Nonetheless, we were more concerned with improving the reliability of the marks than addressing subjective aspects. Furthermore, we conceived the evaluation form as an educational aid serving as a guide for students doing the presentation (e.g. speak out loud and clear, maintain eye contact with the audience . . .).

Secondly, the students, together with the subject lecturer, have been able to try out the evaluation form and have familiarised themselves with the criteria and manner of the assessment prior to the assessed presentations dealt with in this paper. In contrast, the other four lecturers who collaborated

Table 2. Analysis of the reliability of marks given by students and lecturers

	Marks given by students to colleagues	Marks given by group of lecturers watching video of presentations
Number of markers	43	4
Number of presentations marked	23	23
Number of observations	242	87
Average number of markers per presentation (N)	10.52	3.78
ANOVA's F of the marks given by n markers	10.34	4.20
ANOVA's F* when the marks of the subject lecturer are added to the marks given by n markers	12.24	6.13
Reliability of the marks given by N markers (r_{nn})	0.90	0.76
Estimation of the reliability of a single marker (r_{11})	0.47	0.46
Reliability of the marks when those of the subject lecturer are added to those of n markers (r^*_{nn})	0.92	0.84
Estimation of the reliability of marks by subject lecturer (r_{tt})	0.66	0.66
Estimation of number of students/lecturers who should assess each presentation so that their reliability would be similar to that of the subject lecturer. (DeltaN)	2.14	2.27

in the study did not benefit from this. They were informed of the criteria in the same session in which they had to mark the video-recorded presentations. After listening to a 15-minute explanation by the subject lecturer, they spent a few minutes reading the criteria and clarifying any doubts. Once any queries were sorted out, the videos were screened and assessment scores were given to the oral presentations. This may be one of the reasons why the estimated reliability of one of these lecturers (r_{11}) is appreciably lower than that of the subject lecturer (r_{tt}). We consider that these four lecturers had similar previous experience, and in some cases their experience in the assessment of oral presentations of students was greater than that of the subject lecturer. Therefore this enhancement in reliability can be explained by degree of familiarity with the criteria used. Furthermore, given that r_{11} was similar in the lecturers (as they are people accustomed to marking their students but not familiar with the particular set of criteria that we have used for assessment in this subject) and in the students (people who are not used to marking their colleagues, but trained in the use of assessment criteria), an intuitive deduction would be that the criteria used and the previous training have functioned as a substitute for experience.

Conclusions

First we wish to point out that our conclusions are applicable only to situations similar to the one reported on: oral presentations by students, marked by using an evaluation form designed by the students themselves; they were also trained as to their use. Furthermore, the student assessments contribute to the summative assessment of their colleagues [12] but they also received a mark as markers, which made them take the task seriously. We also attempted to motivate students by reminding them that, in their future careers, it would be quite likely that they would have to assess the performance of their subordinates, perhaps even of their own colleagues. All these features have been drawn from previous research; without them, the results would have been different.

Under these conditions, we could consider that the assessments of the presentations, calculated as an average of the marks given by various students acting as markers is not significantly different from the marks given by the subject lecturer. Hence we could use the marks from Peer assessment in the final assessment of the students without this distorting the results.

Furthermore, the assessments by the subject lecturer presented some differences when comparing them to the average marks proposed by the group of four lecturers who collaborated in the research. Even though these differences are not big, they serve to remind us that regardless of the experience a lecturer may have, we cannot assume that the assessments carried out are always beyond dispute.

This conclusion is backed up by the fact that the estimated reliabilities of a single marker are moderately low in the case of the subject lecturer and somewhat low in the case of the students or the other lecturers participating. Hence, the most advisable approach, in terms of reliability, would be to use the assessments of various markers simultaneously.

The results of our study show fairly clearly that students can become good markers of oral presentations given by their colleagues and that their assessments could be useful in establishing the marks of their colleagues without there being differences in the marks proposed by the subject lecturer.

If we can trust the assessments by students, we lecturers could partially free ourselves from the responsibility and time commitment involved in assessment. At the same time we enhance the autonomy, responsibility and participation of students in the process (aspects which usually generate motivation in students); consequently more time will be freed up in the rather full agendas of lecturers.

If we do not wish to leave all of the responsibility with the students, we have two possibilities: make several lecturers intervene or use collaborative assessment. In the context of this subject, the usual scenario is that more lecturers are not available to participate in assessment (because only one lecturer teaches the subject or because, even though there may be various lecturers, they do not have the time to duplicate assessments). Hence, we could improve reliability if the subject lecturer, together with various students (the minimum should be between two and four), would mark each oral presentation. In this way we could obtain some r^*_{nn} values greater than 0.80 that would generally be considered as acceptable [13]. With this procedure, the lecturer would not save time as the oral presentations would still have to be marked as well as spending time in working out the averages of the assessments of the participating students. Nonetheless the reliability of the assessment would be improved and students would become more involved in the process; hence there would be educational and motivational benefits as a result.

We also wish to point out, as factors for an enhanced reliability in the assessments, a selection of criteria that are as objective as possible. This also includes the creation of assessment guides, familiarity or training in assessment criteria as well as experience in the marking of oral presentations. The more these factors are present, the greater the reliability will be in the assessments carried out. Furthermore, if we adequately manage the two factors that fall within the ambit of the lecturer's role (selection of criteria and training) we can obtain very high reliabilities, even with few markers of each presentation. Hence, there does not seem to be any limitations to the fostering of greater student participation in the assessment of the final mark of their colleagues.

The manner in which assessment criteria on the

evaluation form used in the active participation of students, and the activities aimed at training students as to the use of the evaluation form, may have also positively influenced the reliability of the marks given. However, we are in no position to value its effect at present. Our intention has been to pursue and expand upon the research of later-year students who have not participated in the design of the evaluation form and to see what results are generated. We would also like to analyse what happens when students use the evaluation form under the same conditions as the group of four lecturers. In other words, watching video of the oral presentations after listening to a 15-minute explanation of the criteria without training on the use of the evaluation form.

As a concluding remark we need to bear in mind that including peer assessment and its use in working out the students' final mark will be hindered by traditional university scenarios, particularly if the transfer of responsibilities shifts the balance of power between lecturers and students [1, 4, 10, 12, 37]. Furthermore, when self-assessment or peer assessment is activated, the management of the process becomes more complicated and more time is needed to see it through. Hence, it would be advisable to review, simplify and computerise the procedure to avoid placing an extra burden on teaching staff [22, 38].

Acknowledgements—We would like to thank the Grupo de Innovación en la Evaluación para la Mejora del Aprendizaje Activo (Innovation in Assessment for Enhanced Active Learning Group) of Universidad Politecnica of Valencia for their help financing and preparing this paper, and the R+D+i Linguistic Assistance Office at the Universidad Politecnica of Valencia for their help in translating this paper.

REFERENCES

1. Filip Dochy, Mien Segers, and D. Sluijsmans, The use of self-, peer and co-assessment in higher education: a review, *Studies in Higher Education*, **24**(3), 1999, pp. 331–350.
2. Effie Maclellan, Assessment for learning: the differing perceptions of tutors and students, *Assessment & Evaluation in Higher Education*, **26**(4), 2001, pp. 307–318.
3. Paul Orsmond, Stephen Merry, and Kevin Reiling, The importance of marking criteria in the use of peer assessment, *Assessment & Evaluation in Higher Education*, **21**(3), 1996, pp. 239–250.
4. Armando Rugarcia, *et al.*, The future of engineering education. A vision for a new century, *Chemical Engineering Education*, **34**(1), 2000, pp. 16–25.
5. J. McGourty, Using multisource feedback in the classroom: A computer-based approach, *IEEE Transactions on Education* **43**(2), 2000, pp. 120–124.
6. D. Chadha and G. Nicholls, Teaching transferable skills to undergraduate engineering students: Recognising the value of embedded and bolt-on approaches, *Int. J. Eng. Educ.,* **22**(1), 2006, pp. 116–122.
7. Norrie S. Edward, Mark my words: self and peer assessment as an aid to learning, *European Journal of Engineering Education*, **28**(1), 2003, pp. 103–116.
8. Kam-Por Kwan and Roberta Leung, Tutor versus peer group assessment of student performance in a simulation training exercise, *Assesment & Evaluation in Higher Education*, **21**(3), 1996, pp. 205–214.
9. C. Boiarsky, Teaching engineering students to communicate effectively: a metacognitive approach, *Int. J. Eng. Educ.,* **20**(2), 2004, pp. 251–260.
10. Karen Macpherson, The development of critical thinking skills in undergraduate supervisory management units: efficacy of student peer assessment, *Assessment & Evaluation in Higher Education*, **24**(3), 1999, pp. 273–284.
11. D. M.O'Doherty, Working as part of a balanced team, *Int. J. Eng. Educ.,* **21**(1), 2005, pp. 113–120.
12. Humphreys, P. *et al.*, Developing transferable groupwork skills for engineering students, *Int. J. Eng. Educ.,* **17**(1), 2001, pp. 59–66.
13. Douglas J. Magin and Phil Helmore, Peer and Teacher Assessments of Oral Presentation Skills: how reliable are they?, *Studies in Higher Education*, **26**(3), 2001, pp. 287–298.
14. Paul Orsmond, Stephen Merry, and Kevin Reiling, The use of student derived marking criteria in peer and self-assessment, *Assessment & Evaluation in Higher Education*, **25**(1), 2000, pp. 23–38.
15. J. M. K. MacAlpine, Improving and encouraging peer assessment of student presentations, *Assessment & Evaluation in Higher Education*, **24**(1), 1999, pp. 15–25.
16. Terry Gatfield, Examining student satisfaction with group projects and peer assessment, *Assessment & Evaluation in Higher Education*, **24**(4), 1999, pp. 365–377.
17. Douglas J. Magin, A novel technique for comparing the reliability of multiple peer assessments with that of single teacher assessments of group process work, *Assessment & Evaluation in Higher Education*, **26**(2), 2001, pp. 139–152.
18. P. D. Fisher, D. M. Zeligman, and J. S. Fairweather, Self-assessed student learning outcomes in an engineering service course, *Int. J. Eng. Educ.,* **21**(3), 2005, pp. 446–456.
19. Mike Searby and Tim Ewers, An evaluation of the use of peer assessment in higher educartion: a case study un the school of music, Kingston University, *Assessment & Evaluation in Higher Education*, **22**(4), 1997, pp. 371.
20. Paul Orsmond, Merry Stephen, and Kevin Reiling, A study in self-assessment: tutor and students' perceptions of performance criteria, *Assessment & Evaluation in Higher Education*, **22**(4), 1997, pp. 357–370.
21. M. Reynolds and K. Trehan, Assessment: a critical perspective, *Studies in Higher Education*, **25**(3), 2000, pp. 267–278.
22. Winnie Cheng, and Martin Warren, Peer and teacher assessment of the oral and written tasks of a group project, *Assessment & Evaluation in Higher Education*, **24**(3), 1999, pp. 301–314.

23. M.Ward, L. Gruppen, and G. Regehr, Measuring self-assessment: current state of the art, *Advances in Health Sciences Education*, **7**(1), 2002, pp. 63–80.
24. Douglas J. Magin, Reciprocity as a source of bias in multiple peer assessment of group work, *Studies in Higher Education*, **26**(1), 2001, pp. 53–63.
25. S. E. Earl, Staff and peer assessment -measuring an individual's contribution to group performance, *Assessment & Evaluation in Higher Education*, **11**(1), 1986, pp. 60–69.
26. Keith Sullivan and Cedric Hall, Introducing students to self-assessment, *Assessment & Evaluation in Higher Education*, **22**(3), 1997, pp. 289–305.
27. D. Dancer and P. Kamvounias, Student involvement in assessment: a project designed to assess class participation fairly and reliably, *Assessment and Evaluation in Higher Education: An International Journal*, **30**(4), 2005, pp. 445–454.
28. Ibrahim Al-Fallay, The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment, *System*, **32**(3), 2004, pp. 407–427.
29. Falchikov, N. and J. Goldfinch, Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks, *Review of Educational Research*, **70**(3), 2000, pp. 287–322.
30. M. Freeman, Peer assessment by groups of group work, *Assessment & Evaluation in Higher Education*, **20**(3), 1995, pp. 289–300.
31. Mark Langan, *et al.*, Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria, *Assesment & Evaluation in Higher Education*, **30**(1), 2005, pp. 21–34.
32. N. Falchikov, Peer feedback marking—developing peer assessment, *Innovations in Education and Training International*, **32**(2), 1995, pp. 175–187.
33. Eric Van Duzer and Flora McMartin, Methods to improve the validity and sensitivity of a self/peer assessment instrument, *IEE Transactions on Education*, **43**(2), 2000, pp. 153–157.
34. Geoff. Isaacs, *Assessment for Learning,* University of Queensland. (2001).
35. J. S. Kane, and Edward E. Lawler III, Methods of peer assessment, *Psychological Bulletin*, **85**(3), 1978, pp. 555–586.
36. VVAA. SPSS 10 guia para el análisis de datos. 2002, Sevicio de Informática Universidad de Cádiz, 2006 (internet communication).
37. V. N.Tariq, *et al.*, Developing a new approach to the assessment of project work, *Assessment & Evaluation in Higher Education*, **23**(3), 1998, pp. 221–240.
38. D. Briedis, Developing effective assessment of student professional outcomes, *Int. J. Eng. Educ.,* **18**(2), 2002, pp. 208–216.

APPENDIX A

Calculation of reliability—adapted from Magin (2001a) and Magin and Helmore (2001)

Analysis of one factor results.

The data we need are the average number of markers for each oral presentation (N) and the ratio F (sum of the squares between groups of markers, divided by the sum of the squares within the group of markers). Both data are given in the ANOVA summary tables of any computerised programme that carries out these statistical calculations. The symbol F^* will be used for the calculation of ANOVA values, when the marks of the subject lecturer are added to the marks of the N markers.

The variables in the study have been calculated thus:

- Reliability of the marks in the group of N markers

$$r_{nn} = (F - 1)/F$$

- Estimation of the reliability as individual marker

$$r_{11} = (F - 1)/(F + N - 1)$$

- Reliability of the marks within the group when assessments by the subject lecturer are included in N markers

$$r^*_{nn} = (F^* - 1)/F^*$$

- Estimation of the reliability of the marks of the subject lecturer

$$r_{tt} = (F^* - F)/(F^* - F + 1)$$

- Estimation of the number of markers who would have to assess each oral presentation to ensure that the reliability would be similar to the subject lecturer's.

$$\text{Delta } N = N(F^* - F)/(F - 1)$$

APPENDIX B

Table B.1. Evaluation form for marking of presentations

	Levels/Marks			
Criteria	**0**	**1**	**2**	**3**
Eye contact	Read only notes or transparencies	Looks at one or few people	Occasionally, keeps eye contact with everybody	Frequently, keeps eye contact with everybody
Looks calm	No	Little	—	Yes
Speaking	Not audible	Audible but the speaker speaks too fast/slow	Audible but the speaker speaks a bit fast/slow	Audible and appropriate pace
Readable transparencies	No, they are too difficult to read	A bit difficult to read	—	Yes
Both people take part in actively in the presentation	No	There is a significant difference between the two speakers	—	Yes
				Subtotal A
Transparency	Without colours or pictures	More than one colour	Includes charts or pictures	Colours, charts and pictures
Summarised transparency	No, long paragraphs	Yes, but more than 14 text lines	—	Yes, 14 text lines or less per transparency
End before the three minutes given	No	—	—	Yes
Presentation focused on topics subject-related	The topic is not subject-related or does not fulfil the activity requirements.	—	—	Yes
				Subtotal B

Total marks: Subtotal A + Subtotal B (maximum score: 27)

Juan Marin is a lecturer at the Department of Business Administration in the Polytechnic University of Valencia (Spain). He lectures on management, teamwork and human resources management. With regard to these areas he has also worked as a consultant for some companies in Spain and El Salvador (Central America). His main research fields are participative management, lean manufacturing systems, performance evaluation and active learning in higher education.

Paz Marín is a teacher at the Department of Spanish and Portuguese, in the University of Cambridge, UK. She lectures on Spanish language and Translation into Spanish. Her research interests lie on Didactics of Spanish as a foreign language as well as Specialized Translation.

Cristobal Miralles is a lecturer at the Department of Business Administration in the Polytechnic University of Valencia (Spain). He lectures on operations management. His main research fields are logistics, operations management, adapted jobs for disabled or handicapped workers and quantitative methods.