Student-Graded Oral Presentations

http://dx.doi.org/10.3991/ijep.v5i4.4841

O. M. Ågren
Umeå University, Umeå, Sweden

***Abstract*—We describe a way to use peer-graded oral presentations as a way of reducing the load on the teacher, and show that almost identical results as can be achieved as with teacher graded presentations. Moreover, we have found that very little in the form of explicit criteria are needed.**

***Index Terms*—Didactics, Peer assessment, Teacher offloading.**

I. INTRODUCTION

Some of the most imperative skills for a new engineer seem to be those involving oral presentations [1]. This implies that the prospective engineers need to practice these skills during their education. Moreover, these are skills that are easier to master with practical experience.

There has fortunately been an increase in student involvement in courses lately, mostly in the form of self or peer assessment. The main reason for this is that the students will be more active and will thus gain more from their studies (see e.g. [2, 3]). Both self and peer assessment can be used to assess writing [4, 5, 6, 7, 8] as well as presentations [9].

It has been shown that well defined assessment criteria are helpful in getting good (or at least consistent) assessments, but no conclusive evidence have been shown with regards to the influence of age brackets, educational levels or sub-assessments of various criteria [9]. We will, partly because of this, be moderate in our discussions.

Another important question to look at is that of the trustworthiness of the results from these assessments, i.e. the reliability and validity of the results. While most of the papers on the quality of peer assessments focus on either reliability (mainly between peer assessments) and validity (between peer assessments and teacher assessment) as can be seen in the meta-analysis of [9], we will look at both in our assessment of our dataset. This dataset is also bigger than any of those found there, meaning that we can apply advanced statistical methods to it.

Main ideas of this paper:

1. Students in advanced courses are able to grade fairly without being given explicit grading criteria.
2. The students and teachers will, on average, give the same grades to each group.
3. Given enough students, summative assessments can be given by their peers, using the teachers (or teaching assistants) as fail safes.

A. Background

The course in Computer Architecture at the Computing Science department at Umeå University was a C level course (second highest level) in the pre-Bologna system that was used in Sweden. The course could be used as either a last course in a Bachelors degree or as an advanced course in a Masters degree. The course had three mandatory assignments and a written exam, and used a system where 20% of the final marks came from the assignments. These assignments were performed in pairs, but could be done individually if the student so chose. The 20% was given in lumps of 5% for each assignment if they were handed in on time (with deductions for being late) and had a passing grade before the exam. The final 5% came from an oral presentation, which was originally graded by the teacher.

The first and second assignments were to write assembly language for a number of virtual machines [10, 11, 12] and to implement one of the virtual machines in any computer language, respectively.

The third assignment was to write a short technical report on something within the computer architecture field, such as a processor, a bus or any type of storage media. This assignment was heavily edited and collected in a proceeding in order to model a workshop as closely as possible. This increased the likelihood that the students actually turned in their assignments in due time; everyone wanted to be in the proceedings. This also meant that the final assignment could yield up to 10% of the final marks.

During the first few years there were a number of students that contested the gradings, all the way up to heated arguments. We wanted to see if that could be alleviated by letting the students perform peer-grading [13], and the results from those experiments are presented here.

The same assignments were used a few more years after this, but the teacher that took over the course did unfortunately not keep any records. The peer-reviewed oral presentations were after this moved to another course with format changed in such a way that later data cannot directly be compared to those shown here.

II. METHODS

The students could make any type of presentation that they could think of. Moreover, they could use any means available for the actual presentation, including overhead slides, the whiteboard, a tape recorder, etc. The one rule that had to be followed was that the presentation must fit in the allotted time slot, between eight and ten minutes (depending on year).

Each presentation was graded by one teaching assistant (called teacher in all tables and figures) and at least all the students that presented in the same hour. The presentations were open for anyone to attend and grade, including other teachers and students. I personally sat in on one of the presentation tracks, as backup and extra support for that teaching assistant.

The grades were given individually by each grader, one grade for each presentation group[1]. There were six possible grades to give to a presentation, ranging from zero to five. The rationale behind this was twofold:

- The grades given by the presentation should match what it would be worth on the final exam.
- There should be no single average score to choose, thereby forcing the students to make a choice.

We had, moreover, added the extra rule that there had to be differences in the grades between presentations, e.g. a grading paper with all fives would be ignored in the process.

The only guidelines given to the students were the following:

"Grade each group according to how well you thought they managed to get to the core of the subject, how prepared they were, the disposition that was used and how the presentation was done. Do not grade them according to how nervous they were."

The grades were collected in a spreadsheet. The average, median and mode results of each presentation were calculated, and was used directly as a given grade if they agreed with each other. If they disagreed, the grade given by the teaching assistant was used as a decisive vote to show what grade to give to that presentation.

III. RESULTS

There was a total of 2310 votes given to the 112 presentations done in 2003–2006, disregarding no-shows that automatically got a zero. All averages and counts of this dataset can be seen in Table I. Fig. 1 contains the average given grade as well as the 99% confidence intervals of each type using normal distribution for the students' grades and t-distribution for the teacher's grades. There is very little difference between each year and most of the differences between years are not statistically significant.

The difference between the staff and the student grades have been checked as well, yielding a very interesting pattern over all gradings. It is normally distributed with μ=−0.047316, skew of 0.021278, kurtosis of −0.13622 and $\sigma \approx s$=1.0391 over all four years. Looking at each year (rather than each presentation track or in total) yields a slightly different picture but there are still very small differences, as can be seen in Table II.

A. Reliability Estimates

It is possible to make estimates of the reliability for the numbers using analysis of the variance in the dataset. The test statistic F (defined as variance between groups divided by variance within groups) given by one-way analysis of variance (ANOVA) can be used to calculate the reliability of the averaged mark (r_{nn}) and the estimated reliability of the individual raters (r_{11}), as given in (1) [14, 13]. The results of these calculations can be seen in Table III.

$$r_{nn} = \frac{F-1}{F} \quad \text{and} \quad r_{11} = \frac{F-1}{F+N-1} \tag{1}$$

[1] Each presentation group corresponded to a subject and usually consisted of one or two students.

TABLE I.
DATA PER YEAR, GIVEN AS AVERAGE (BOTH FOR TEACHING ASSISTANTS AND STUDENTS), AVERAGE MEDIAN, AVERAGE MODE, AVERAGE GIVEN GRADE AND COUNTS (N) FOR BOTH TEACHING ASSISTANTS AND STUDENTS.

Year		2003	2004	2005	2006
Teacher	***Average Grade***	3.35	3.57	3.65	3.94
	n	43	21	31	17
Student	***Average Grade***	3.47	3.66	3.61	3.59
	Std. Dev.	0.93	0.94	0.93	0.93
	Average Median	3.48	3.60	3.66	3.68
	Average Mode	3.56	3.57	3.65	3.65
	n	1104	316	481	298
Given	***Average Grade***	3.49	3.62	3.65	3.71

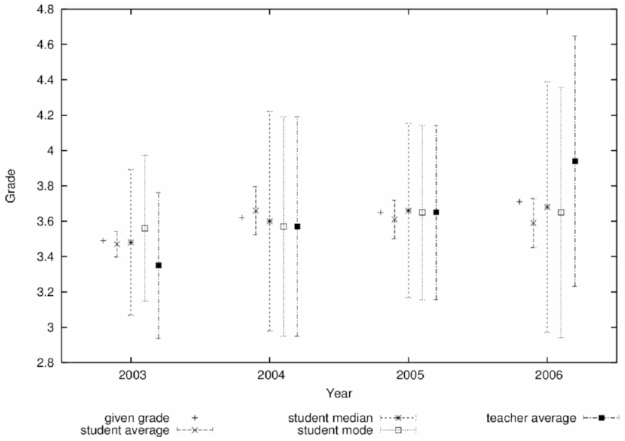

Figure 1. The average given grade as well as the 99% confidence intervals for the grades given by students (average, average median and average mode) and the teaching assistants per year.

TABLE II.
DISTRIBUTION OF THE DIFFERENCE BETWEEN TEACHING ASSISTANT AND STUDENT GRADING

Year	2003	2004	2005	2006	Total
-3	8	2	3	0	13
-2	92	24	19	9	144
-1	301	96	132	45	574
0	412	111	177	112	812
1	232	65	121	107	525
2	54	14	26	24	118
3	5	4	2	1	12
Skew	0.044	0.256	0.038	-0.212	0.021
Kurtosis	-0.133	0.099	-0.094	-0.115	-0.136

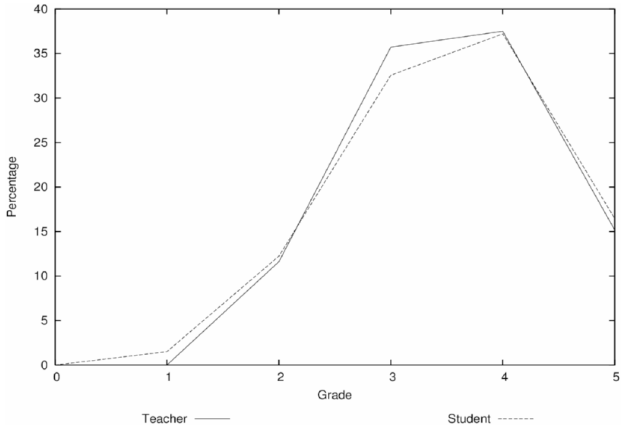

Figure 2. The almost identical distribution of the grades by the teaching assistants and the students.

TABLE III.
STUDENT ASSESSMENTS: RELIABILITY COEFFICIENTS PER YEAR

Year	Reliability of averaged mark (r_{nn})	Mean number of peer raters (k)	Single rater reliability (r_{11})	Number of groups assessed (N)
2003	0.937	25.67	0.259	43
2004	0.858	15.05	0.224	21
2005	0.866	15.52	0.172	31
2006	0.904	17.53	0.355	17

IV. DISCUSSION

Looking at the results in Table 1 reveals some rather interesting tidbits of information; While the student average median and average mode grades was non-decreasing, the student average was actually decreasing 2004–2006. We attribute the student average to slightly more critical students, as well as a decrease of students from one program. The change in the other two are, however, not significant, but might indicate that more than six levels could have been used to get more information.

One of the teaching assistants from 2003 and 2004 was probably a bit too critical and the teaching assistant from 2006 was instead overly positive, according to the data. It is unfortunately very hard to guard against things like this, but it did not make that much difference in the end because of the large number of students that were more critical.

A very interesting question is "What should have been done differently?" The most obvious thing to change would be to increase the number of grading levels and incur the same increase in the number of points given by the assignments. Doubling the points from the oral presentation would not be entirely out of order, since it was a very important and large part of the course. It was also one of the most frequent suggestions found in the course evaluation.

As a closing remark, I would say that the average grades over these four years are very balanced between teaching assistants and students. The closeness of grading can be seen in Figure 2. In fact, the teaching assistants gave out on average 3.56 points per group and the students had an average of 3.55 points, meaning that the results are closer in grouping than any of the studies found in [9]. The students did an excellent job of grading each other, and using it in a course will not incur any extra costs except possibly for collecting the data and performing the calculations.

V. ACKNOWLEDGEMENTS

The author would like to thank Peter Jacobsson, from whom the Computer Architecture course was inherited. Moreover, the author is forever in debt to all of the teaching assistants and students who participated in the experiment over the years.

VI. REFERENCES

[1] Morley, L. Producing new workers: Quality, equality and employability in higher education. *Quality in Higher Education*, 7(2):131–138, 2001. http://dx.doi.org/10.1080/13538320120060024

[2] Piaget, J. (1971). *Science of Education and the Psychology of the Child*. Penguin Books. Written 1969 and translated by Derek Coltman.

[3] Cross, K. P. (1992). *Adults as Learners: Increasing Participation and Facilitating Learning*. Jossey-Bass.

[4] Marcoulides, G. A. and Simkin, M. G. (1995). The consistency of peer review in student writing projects. *Journal of Education for Business*, 70(4):220–223. http://dx.doi.org/10.1080/08832323.1995.10117753

[5] Orsmond, P. and Merry, S. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, 21(3):239–250. http://dx.doi.org/10.1080/0260293960210304

[6] Topping, K. J., Smith, E. F., Swanson, I., and Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education*, 25(2):140–169. http://dx.doi.org/10.1080/713611428

[7] Orsmond, P., Merry, S., and Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, 25(1):23–38. http://dx.doi.org/10.1080/02602930050025006

[8] Saito, H. and Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research*, 8(1):31–54. http://dx.doi.org/10.1191/1362168804lr133oa

[9] Falchikov, N. and Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3):287–322. http://dx.doi.org/10.3102/00346543070003287

[10] Ågren, O. (1999). Teaching Computer Concepts Using Virtual Machines. *SIGCSE Bulletin*, 31(2). http://dx.doi.org/10.1145/571535.571578

[11] Ågren, O. (2000). The DARK-Series of Virtual Machines. Technical report, Umeå University, Umeå, Sweden. UMINF 00.15, ISSN 0348-0542.

[12] Ågren, O. (2000). Virtual Machines as an Aid in Teaching Computer Concepts. *IEEE TCCA Newsletter*.

[13] Magin, D. and Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education*, 26(3):287–298. http://dx.doi.org/10.1080/03075070120076264

[14] Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, 16(4):407–424. http://dx.doi.org/10.1007/BF02288803

AUTHOR

O. M. Ågren is with the Department of Applied Physics and Electronics, Umeå University, SE-901 87 Umeå, Sweden (e-mail: ola.agren@umu.se). The data presented in this paper was collected while he was finishing his PhD thesis at the Computing Science Department at the same university.

This work was supported in part by the Computing Science Department at Umeå University. Submitted 03 July 2015. Published as resubmitted by the author 10 October 2015.