1st International Conference on Higher Education Advances, HEAd'15
Universitat Politècnica de València, València, 2015
DOI: http://dx.doi.org/10.4995/HEAd15.2015.441

Optimising Peer Marking with Explicit Training: from Superficial to Deep Learning

S.B. Caldwell, T.D. Gedeon

Research School of Computer Science
Australian National University

Abstract: We describe our use of formative assessment tasks measuring superficial learning as explicit training for peer assessment of a major summative assessment task (report writing), which requires deep learning. COMP1710 at the Australian National University is a first year Web Development and Design course done by over 100 students each year, by many Computing students in their first semester of their first year or at any time prior to graduation; the course also attracts some 25% of its cohort from other academic areas of the University. We found that formative assessment trained peer markers performing a surface learning task can produce peer marks consistent with our expert summative task marker. Weaker students only demonstrating superficial learning were able to reliably assess the reports of the better students capable of the deeper learning required to produce the reports. This significantly increases the usefulness of peer marking, and could have use in large online courses such as MOOCs.

Keywords: peer marking, formative assessment, summative assessment, desired mark

Introduction

As the landscape of education continues to be transformed via evolving pedagogies and technologies, finding ways to practically implement these as teaching enhancements has become a priority in our first year web development and design course. As a result, we are examining ways to combine formative and summative assessment to create multi-layered learning outcomes for students while creating efficiencies in marking for this well-attended course.

The contributions of the work reported in this paper include *explicit training* of peer assessors, use of superficial/formative assessment tasks for peer assessor training, peer assessment by comparative marking, and use of weaker students to reliably assess stronger students' work in a deep learning summative task.

Theoretical Background

Formative assessment is designed to facilitate learning and typically involves qualitative feedback rather than scores. Summative assessment is a snapshot of the learning at a particular time, and is usually the mechanism by which final results and grades are reported. However, it can be difficult to balance these two types of assessment. William (2000) indicates that "few teachers are able or willing to operate parallel assessment systems," and suggests formative assessments could only provide an 'envelope' of overall scores. Formative assessment with marks is only primarily formative but is not useful as a replacement for summative assessment (MacLean and McKeown, 2012), with formative assessment not predictive of final grade, but predictive of pass/fail.

There is a large body of work extant on peer assessment, thus we introduce this area only briefly, in particular mentioning work relevant to our study. The paper by Hamer *et al.* (2009) reports on the difference between student (peer assessment) marks and expert marks in a large programming course. They found "good correlations that *improve with student ability* and experience" (our emphasis added, see our results section). Kulkarni *et al.* (2015) used a “fortune cookie” approach to provide qualitative and personalised feedback. This was found to have no effect on the amount of feedback returned, however, they noted that multiple assignments assessed via peer assessments provided incremental improvements on the quality of assessments produced over time. Reilly *et al.* (2009) found that combining just 2 peer marks produced high reliability, which informed our grouping approach to peer evaluation.

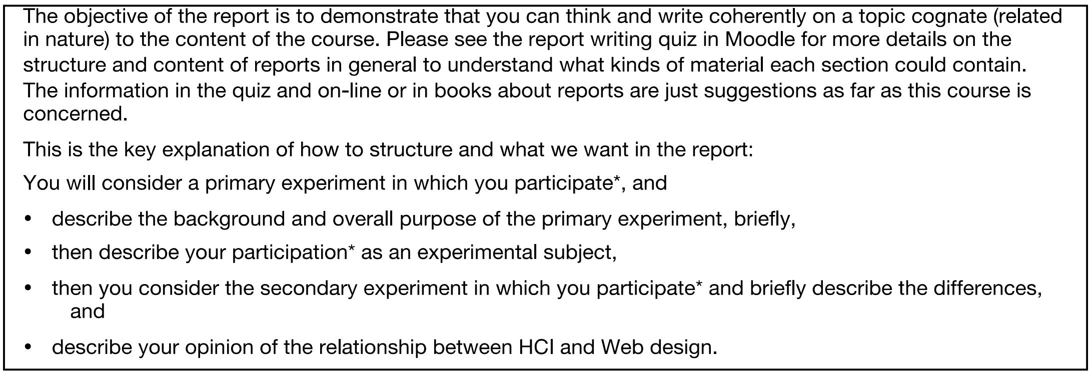
Methods – Assessment in COMP1710

COMP1710 Web Development and Design is a course in the Research School of Computer Science at the Australian National University (ANU) delivered annually to over 100 students. Many Computing students take it in their first semester of their first year, or at any time prior to graduation, as there is not a strong prerequisites tail. The course also attracts some 25% of the cohort from other academic areas of the University. The authors are chief tutor and course co-ordinator, respectively.

Key to our approach is to separate the surface learning / competency parts of our course assessment from deeper learning assessment. We consider this to be essentially the same problem as separating the formative and summative marking we all do. Briefly, our solution to the surface / deep learning or formative / summative evaluation quandary is to explicitly separate the marks so that surface and formative marks can only be collected to achieve a Pass in the course, and higher grades require *qualitatively different kinds of marks* which add conventionally to the pass marks for students to achieve higher grades. Thus, by ‘*qualitatively different*’ we mean that the students have to perform qualitatively differently to earn such marks. While reporting on the overall outcomes will be done elsewhere, as far as this paper is concerned we note that the training in report assessment and the marks for doing that assessment are from the surface/formative category, while the report itself is from the deep/summative category.

Formative – quiz on report writing

The formative quiz on report writing discusses an online technical report with commentary, a more abstract discussion of the usual components of technical reports, the use of images and charts, mistakes to avoid, and two final short essay questions. Additional formative training was provided with the essay questions 9 (report structure) and 10 (experiment participation reflection), which required students to assign a mark out of 6, then justify their mark with a short explanatory paragraph. Both questions 9 and 10 were marked twice. This required three marking sessions, as the second marking



The objective of the report is to demonstrate that you can think and write coherently on a topic cognate (related in nature) to the content of the course. Please see the report writing quiz in Moodle for more details on the structure and content of reports in general to understand what kinds of material each section could contain. The information in the quiz and on-line or in books about reports are just suggestions as far as this course is concerned.

This is the key explanation of how to structure and what we want in the report:

You will consider a primary experiment in which you participate*, and

- describe the background and overall purpose of the primary experiment, briefly,
- then describe your participation* as an experimental subject,
- then you consider the secondary experiment in which you participate* and briefly describe the differences, and
- describe your opinion of the relationship between HCI and Web design.

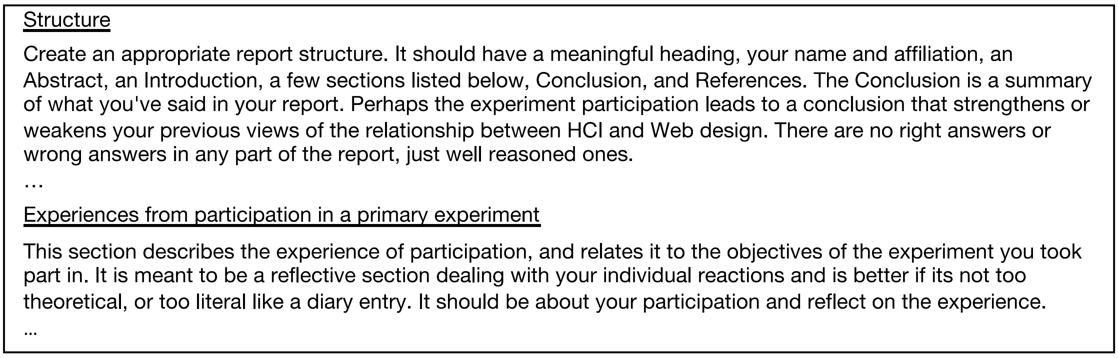
Figure 1. Report Specifications – Introduction

of question 9 was done at the same time as the first marking of question 10. The marking was all done by the course co-ordinator, and took substantial effort. The major reward was an unsolicited comment by email from the (independent) marker for the reports: “The quality of the reports is unrecognisably better than when I first marked these reports. Many came very close to completing what was required of them.”

Summative – report specifications

The report itself was primarily worth ‘deep marks,’ which would demonstrate understanding beyond a Pass level. Students wrote about their experiences when participating in real experiments as described in Figure 1.

Figure 2 demonstrates the link between the formative quiz and summative report tasks; question 9 of the quiz asks students to mark a report structure, and question 10 asks students to mark the reflection on experiment participation in sample reports, preparatory to students writing their own reports. The sample reports provided are similar in terms of overall topic, but with different experiments, thus ensuring that the samples are useful as examples of work but can not be directly copied.



Structure

Create an appropriate report structure. It should have a meaningful heading, your name and affiliation, an Abstract, an Introduction, a few sections listed below, Conclusion, and References. The Conclusion is a summary of what you've said in your report. Perhaps the experiment participation leads to a conclusion that strengthens or weakens your previous views of the relationship between HCI and Web design. There are no right answers or wrong answers in any part of the report, just well reasoned ones.

…

Experiences from participation in a primary experiment

This section describes the experience of participation, and relates it to the objectives of the experiment you took part in. It is meant to be a reflective section dealing with your individual reactions and is better if its not too theoretical, or too literal like a diary entry. It should be about your participation and reflect on the experience.

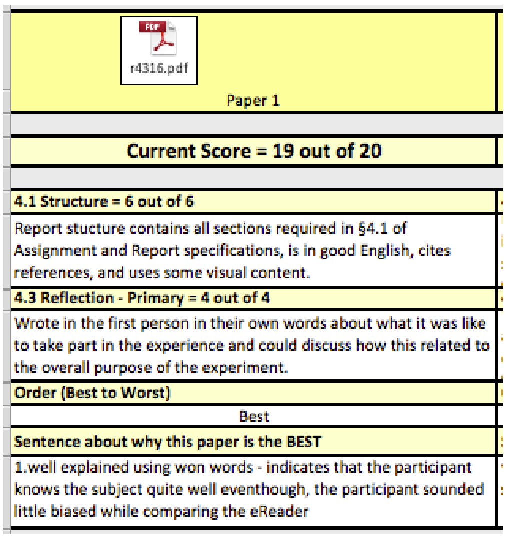
…

Figure 2. Target report sections related to Questions 9 and 10 of formative assessment

Peer assessment – data and analyses

Students participating in the peer evaluation were generally those who had not attained enough marks to achieve a Pass in the course, while students being evaluated were generally those who performed well in the course. Report peer evaluators had a mean total course score of 52.5 SD 13.4, while report writers had a mean total course score of 74.6 SD 9.6 (the mean total course score for all students was 61.2 SD 24.1).

The students doing peer evaluation are sent Excel spreadsheets with five anonymised embedded reports and both qualitative and quantitative evaluation tasks (Figure 3). Below each report is a set of sections with dropdown lists of alternative descriptions. Thus, for “4.1 Structure”, the student chose the description shown as being the one most correct for that part of the evaluation, which in this case was worth 6 out of 6. The student must then choose (again via dropdown lists) for each paper: best / good / middle / bad / worst and for these



Paper 1

Current Score = 19 out of 20

4.1 Structure = 6 out of 6

Report stucture contains all sections required in §4.1 of Assignment and Report specifications, is in good English, cites references, and uses some visual content.

4.3 Reflection - Primary = 4 out of 4

Wrote in the first person in their own words about what it was like to take part in the experience and could discuss how this related to the overall purpose of the experiment.

Order (Best to Worst)

Best

Sentence about why this paper is the BEST

1.well explained using won words - indicates that the participant knows the subject quite well eventhough, the participant sounded little biased while comparing the eReader

Figure 3. Peer marking spreadsheet

reports write a sentence to explain why. The overall score at the top of “19 out of 20” is composed of these scores automatically.

Table 1. Sample results from 1 peer marker

	1st	2nd	3rd	4th	5th	sum
	r6994	r1483	r9549	r3348	r4195	
S rank	1	5	4	2	3	
S mark	17.5	11.5	7.5	16.5	12.5	
D rank	1	4	5	2	3	
D mark	15	7.5	6	14.25	12	
SqErr rank	*0*	*1*	*1*	*0*	*0*	*2*
SqErr mark	*6.25*	*16*	*2.25*	*5.06*	*0.25*	***29.81***
Total	17.5	11.5	7.5	16.5	12.5	
Structure	6	3	1.5	4.5	1.5	
Background	3	3	1	4	4	
Reflection	4	2	2	4	4	
Reflec-Diffs	1.5	1.5	1	1	1	
HCI-Design	3	2	2	3	2	
Ranking	Best	Worst	Bad	Good	Middle	

The benchmark for our comparisons is the marks given by our expert marker. He is a senior colleague with significant relevant experience. He provides two guest lectures and does the marking of all of the reports, with no other engagement with the course. This is as close as seems possible to fully independent marking.

An example of our analysis is shown in Table 1. The bottom 7 rows show the numeric results from the Peer marking spreadsheet (Figure 3). The last row is the student-chosen rankings (converted to numeric form and shown as *S rank*). Notice that the student-chosen *Worst* report is not necessarily the one with the lowest mark. In the table, the *Total* line is copied to the *S mark* line. The two lines below (*D rank/mark*) should be read as “Desired mark” etc., being the mark from our expert marker. We then calculated the squared difference of the ranks, with the sum of these values shown in bold, right. The sum of squares eliminates negative values and penalises large differences, and is commonly used to compare information retrieval rankings. The values of ***2***, ***29.81*** can now be compared to the equivalent *SqErr rank / mark* values for all other students as two estimates of their reliability where that is a measure of similarity to the *Desired* marks and ranks.

These calculations allow us to derive three possible peer marking results for each report. Firstly, the average mark for each report (Ave_S), based on the students’ marks. Second, we can pick the most reliable marks for any report by just picking the mark given by the most reliable student (by_rank) based on the similarity of their ranking of the 5 papers to the *Desired* ranking using the sum of the *SqErr* rank values as the reliability measure. Third, the same as the second but using the sum of the *SqErr* mark values (by_mark) as the reliability measure.

Table 2. Sample results by report

ReportID	Ave S	by rank	by mark	**D mark**
…				
r3348	17	16.5	16.5	**14.3**
r3406	13.2	17.5	17.5	**13.5**
r3626	10.5	10.5	10.5	**12**
r3790	18	19	15.5	**15**
r3841	16.5	16.5	16.5	**13.5**
r4195	12.6	15	12.5	**12**
…				

So for example the average mark for the first report shown in Table 2, *r3348,* is 17, but the *by rank* and *by mark* values are both 16.5, being the mark given by the student shown in Table 1. That student also gave the *by mark* value for the last report in the table (*r4195*). This is possible because another student ranked their 5 reports in the same order as our expert marker, but gave more different numerical marks, hence was more reliable on one measure but not the other.

Results and Discussion

We received 155 peer marks in total for 53 reports, yielding a mean of 2.9 'marking events' per report (SD 1.1). We received a single mark for 6 reports, and six marks for just 1 report. For this dataset, we can compare *Ave_S*, *by_rank*, and *by_mark* for similarity to the estimator *Desired* values in a number of ways. We choose the simplest here, again using sum of squared differences, producing 3 numbers:

- D-to-Ave_S = 655.3
- D-to-by_rank = 806.7
- D-to-by_mark = 650.9

The magnitudes of the numbers are not meaningful, and we just want to compare differences in magnitude. The results suggest that using selection by match of marks to the *Desired* mark produces results no different from simple averaging of marks, and that the use of the ordering of student marks as a selection means is not useful. This was contrary to our intuition.

Examining the statistical significance of peer evaluation results using a two-tailed t-test with *Desired* mark as estimator shows that all three results are highly statistically significant, using the $p < 0.05$ measure:

- p(D-to-Ave_S) = 0.0009
- p(D-to-by_rank) = 0.0006
- p(D-to-by_mark) = 0.006

Unfortunately, this means that none of these three results could be used to approximate the *Desired* marks. Instead of turning to more complex statistical measures such as the Pearson correlation coefficient, we performed a simple check of the averages of the *Student* and *Desired* marks and we discover that they differ by 2 marks out of 20: *Student* marks have a mean of 13.48 with standard deviation 3.3, while *Desired* marks have a mean of 11.53 with a standard deviation of 3.0.

To cope with the difference in mean, we applied the simplest measure, and one most often applied in our experience by examiners' meetings: subtraction. We subtracted two marks from each student mark and then we recalculated our measures:

- D-to-Ave_S = 455.1
- D-to-by_rank = 543.7
- D-to-by_mark = 511.9

and

- p(D-to-Ave_S) = 0.46
- p(D-to-by_rank) = 0.36
- p(D-to-by_mark) = 0.30

These results show that the match to the *Desired* marks is better as the sum of squared error values are smaller, and the differences between the benefit of one over another in approximating the *Desired* marks is less. The *p* values are interesting, as none of the columns is now statistically significantly different from the *Desired* column, which is what we want here. The average error for the best result (*D-to-Ave_S*) is 2.9 marks. In Figure 4 we can see that most of the differences between the peer review marks and the *Desired* marks ('errors'), are below 4 marks, with 5 outliers with errors of: 4.8, 5.8, 7.0, 7.3, and 9.0 marks. The other two

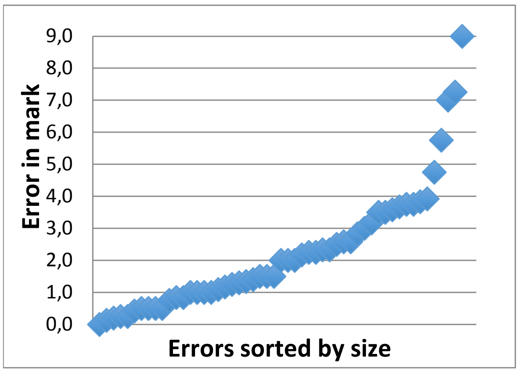


Figure 4. Distribution of Peer Marking 'errors'

distributions with sum of squares errors of 543.7 and 511.9 are similar with a few more outliers.

Conclusions

We have described an experiment involving 31 peer marking events comparing 5 reports at a time, with 53 reports marked in total. Similar to previous work in the literature (Kulkarni, 2015, Hamer, 2009), our work shows that we could reproduce the expert marking for this sample of 53 reports from formatively trained student peer evaluators, as we can produce (a number of) lists of marks which are not statistically significantly separable from the 'true' list of marks provided by our expert marker.

We have made 3 significant contributions. The first contribution is that the use of a surface learning task done as a formative task can perform the role of *explicit training* in the assessment task, and produces high quality results on the first peer assessment task, unlike previous work in the literature (Kulkarni, 2015) which has focused on repeated assessments (which was not possible in our course as only one report is written). The second contribution is the introduction of *comparative assessment* where a number of submissions are evaluated in parallel (five in our case). Finally, the third and perhaps most significant contribution is that we have achieved our results using the weakest students in our cohort marking the rest of the students including the best students, with a high level of accuracy. This use of a surface learning task to reliably predict the results of a deep learning task for better students is impressive. The implication of this is that our students in their surface task were able to correctly recognise the outputs of deep learning tasks from other students.

Acknowledgements

We sought and received approval from the ANU Human Research Ethics committee to undertake this work. We would also like to acknowledge the ANU Vice-Chancellor's Teaching Enhancement Grant that made this study possible.

References

Hamer, J., Purchase, H. C., Denny, P., & Luxton-Reilly, A. (2009, August). Quality of peer assessment in CS1. In Proceedings of the fifth international workshop on Computing education research workshop (pp. 27-36). ACM.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., ... & Klemmer, S. R. (2015). Peer and self assessment in massive online classes. In *Design Thinking Research* (pp. 131-168). Springer International Publishing.

Maclean, G. and McKeown, P. (2013): Comparing online quizzes and take-home assignments as formative assessments in a 100-level economics course, New Zealand Economic Papers, 47(3): 245-256.

Reilly, K., Finnerty, P. L., & Terveen, L. (2009, May). Two peers are better than one: aggregating peer reviews for computing assignments is surprisingly accurate. In Proc. ACM 2009 Int. Conf. on Supporting group work (pp. 115-124). ACM.

William, D. (2000): Integrating formative and summative functions of assessment. In Working group, 10(November).