WEIGHTING FOR COMPUTERIZED PEER-ASSESSMENT TO BE ACCEPTED

Phil Davies

Weighting for Computerized Peer-Assessment to be Accepted

Phil Davies
School of Computing
University of Glamorgan
pdavies@glam.ac.uk

Abstract

This paper details the process undertaken in developing a peer-assessed, essay-based assignment making use of the CAP (**C**omputerised **A**ssessment by **P**eers) system. The walkthrough provides information detailing the needs of a peer-assessed assignment, taking into account the requirement to automatically provide a reward to the student for performing the marking and commenting processes in a qualitative manner. In order to quantify these marking attributes, a menu-driven marking system has been developed, and this menu system is populated with student derived comments that include a weighting factor, so that the 'importance' of said comments can also be included within the overall assessment process.

Background

The bespoke development of computerized peer-assessment systems for essay marking has increased over the past few years (Bhalerao & Ward, 2001; Davies, 2000; Lin et al, 2001; Parsons, 2003). The benefits of peer-assessment have been reported upon in various research papers over a number of years (Falchikov, 1995). There are many issues that affect the ability to manage peer-assessment, and the introduction of these computerized systems has introduced methods to remove some of these management constraints.

However, merely utilizing these ICT systems as management solutions is minimizing the prospective benefits such systems are capable of providing e.g. communication, anonymity, etc. In a past CAA conference (Davies, 2001), the CAP (Computerized Assessment by Peers) tool has been presented as a means of integrating ICT with the pedagogical benefits of peer-assessment. One of the main negatives associated with peer-assessment is the need to provide a significant 'up front' training period to ensure that students are 'able' to perform the marking and commenting processes to an acceptable degree of quality.

By introducing methods of compensating high and low markers automatically via the CAP system has reduced this need, and produced a level of consistency of marking. However, research has identified considerable benefits that students achieve in receiving quality and structured feedback. This paper identifies a method of structuring the computerized peer-marking process via a drop-down menu driven enhancement. The students are able to utilize this system to easily provide pre-set comments. However, it is important that subjectivity of the marking process of peer-assessment is maintained and the students are able to use their own comments. Therefore prior to the assessment process taking place, the students create their own comments within the pre-defined categories of assessment. A problem that has occurred in the past use of the CAP system has been the difficulty dyslexic students have had in providing comments for their peers. The use of this menu driven system has enhanced their ability and confidence in providing such feedback.

It should be noted that in questioning tutors, certain feedback comments that they provide can have a higher degree of importance to some rather than others. Therefore it can be assumed that the same degree of importance of comments is present with students in their peer commenting. To take this into account the students also have to provide a level of importance associated with each peer-comment.

It has been proposed that the comments provided via peer-marking are as important as the marks. Therefore it is essential that the comments provided have a significant correlation to these marks (Davies, 2004). The use of the menu driven/weighted comments provides a means of easily quantifying these comments via the production of a weighted feedback index.

This paper evaluates the quality of this index with respect to the compensated marks produced within the marking process. It questions whether there is a need for marks to be used within the peer-marking process and whether a grade can be attributed to an essay solely via this quantified weighted feedback index from the comments. Also an analysis of the ability of various groups of students is performed to address peer-marking ability and also the categories of comments that have greatest importance to these groups of markers.

By measuring the quality of a marker’s marking and commenting, an automated solution is described that will present a ‘mark for marking’ based upon consistency factors.

Assignment Walkthrough

This study was undertaken with a group of 46 students within a module in the area of Distributed Systems at the University of Glamorgan, in the academic year 2004-5. The students were set an assignment that required them to develop an essay that explained the features associated with Grid Computing, and also were expected to develop proposals where they felt this new strategy could be developed within the commercial sector over the next five years.

A strict word limit of 3,500 words was set within the assignment specification, and also the students were provided with a template of headings to be used within the essay. This provided a standard framework in order to create consistency of appearance of the essays. Also within each topic section within this framework, a separate reference section was expected. The students were instructed to use the web as the main source of information. By utilizing the web, the CAP system permits a method of viewing these web pages via an embedded web browser. This aids the marking of students with regard to viewing how a student has developed their essays (process) and not just the final text (product).

The student having completed their essay then submitted it via the digital drop box facility of the Blackboard VLE used within the University of Glamorgan. Before the students were allowed to enter the marking stage they had to run a registration application in order to create a password for themselves. This also provides a method for students to provide an email address for themselves which could then be used if the communication aspect of the CAP system were to be used (not in this study). When each student registers, a letter code is automatically allocated to them and stored on the database. It is this code that is used to name the provided essays. These are stored, under letter code, in a read-only folder on a network drive within the School of Computing file server.

In past uses of the CAP system, some basic categories have been developed from student markings that were generic to most if not all essays in the area of distributed systems. These were:

Readability, Aimed at Correct Level, Personal Conclusions, Referencing, Research and Use of Sources, Content and Explanations, Examples and Case Studies, Overall Report Quality, Introduction and Definitions, and Presentation.

It was decided to use these categories within the marking of this essay. However, each student would possibly have different comments that they would like to provide within these categories, both positive and negative. Therefore, the next stage of the assessment process required the students to add their own pre-defined comments to aid them to perform the marking process. Each of the comments that they provided was also assigned a weighting in the range of 1-5, with 5 being the highest importance. In this way the subjectivity of the marking process was improved.

Having developed this weighted comments database, the students then sent this to the tutor via digital drop box, renaming it with their student enrolment number.

In the past uses of the peer-assessment process, the initial student self-assessment of their own work has been reported as being of great benefit to the marker (Davies, 2002). It has provided both a means of setting a standard of expectation and also getting the student to be reflective towards their own assignment.

The menu driven marking application was now used to mark/comment their own essays. The server application that sits on the tutor's PC was set for self-assessment, hence when a request comes in from the Client CAP application for an essay it is the student's own essay code that was returned. The Client application then fetched the required essay from the network server's folder.

The student marked their own essay making use of their menu driven database of comments. On completion, their comments (weighted) and marks were saved on the marking database held on the tutor's machine.

The students were then expected to mark at least six of their peers' essays within a two week period. It is important that the students receive a mark for performing the marking process that equates directly to the quality and consistency of the marks they award. However, the comments that they provide are equally important and must also equate directly to the quality of the marked essay. In the assessment detailed in this short paper, the essay itself was worth 70% and the 'mark for marking' was worth 30% of the assignment mark.

In past uses of the CAP system, it is the tutor who has provided the 'mark for marking' by going through each marking etc. This is obviously a long and time consuming task. The CAP system has been augmented to provide a means of automatically providing a mark that reflects the consistency/quality shown by the marker both for comments and marks provided. In order to provide this mark the server aspect of the CAP system has a series of stages that it must go through on completion of the student marking in order to automatically provide this 'mark for marking'.

These stages are:

a) Provide a raw, median based average peer-mark for each essay
b) Look at each marker, and ascertain whether they have in general over- or under-marked the essays. This is done by comparing each mark they have awarded with the raw peer mark awarded for an essay, and then creating an average over- or under- marking grade.
c) This over- or under- marking grade is now used to amend the marking provided for an essay by the marker in question. By modifying these marks it is possible to create a compensated peer-average mark for each essay that takes into account high and low markers.

d) It should be noted that the comments provided by each marker are easily quantifiable by taking into account the use of the menu system i.e. the number of positive comments against the number of negative comments (within each category) (Davies, 2006). The server application is now able to develop a feedback index that matches the comments awarded for a particular essay. By taking into account all of the markings for a particular essay, an average feedback index is created.

e) It is now possible to judge the consistency of the comments provided by a marker against the average feedback index generated.

f) However the comments used by each marker have been weighted in importance, therefore a weighted average feedback index can be generated, and this can be measured against the marker for a better assessment of the marker's commenting.

In order to be able to judge and create a 'mark for marking/commenting', there is a need to able to objectively assess the marker's consistency of marking. Looking at the provision of a mark, a difference factor has been created with regard to a marker (+D or –D) i.e. has the student on average over or under marked. Looking at the marking produced for an essay, the marker could have given it X and the compensated average peer mark awarded could be Y. Therefore it is possible to note for this marking an individual difference of F = Y – X. However, this marker should have provided a difference of +/– D. Therefore, by calculating the absolute difference between F and D for this marker produces a consistency value for this single marking. This procedure is then repeated for each essay marked and an average consistency factor for marking can be created. The lower this value, then the more consistent the marker has been. This process may then be repeated to produce the feedback and weighted feedback consistency factors By mapping these consistency factors for marking and commenting to some pre-defined grading figures (Davies, 2004), it is possible for the CAP system to automatically provide a 'mark for marking/commenting'.

Results

Before attempting to automatically create a 'mark for marking/commenting', it is important to check that using the method of having menu driven comments has not resulted in a loss of correlation between the marks achieved and the comments. It is a key principle of using this method of peer-assessment that the quantification of the comments maps to the quality of an essay. Table 1 shows the mapping of the compensated peer-mark awarded to the average feedback indexes. This table shows a positive correlation of 0.880. The average standard deviation within the feedback indexes to the marks being 3.35.

Table One

-5	-4	-3	-2	-1	-0	+0	1	2	3	4	5	6	7	8	9
29	44	41	49	46	53	64	49	53	60	62	69	68	69		82
	38	48	47	51	45	54	58	53	62	62	64	65	73		
					49	51	50	60	57	57	67	66			
							51	58	53	50		59			
										57		63			
										59		65			
										64					
0	4.2	5.0	1.4	3.5	4.0	6.8	4.8	3.6	3.9	4.7	2.5	3.1	2.8		0
29	41	45	48	49	49	56	52	56	58	59	67	64	71		82

However, this study has made use of weighted comments in order to provide a more specific mapping of comments to map to the subjectivity of the marker. Table 2 shows the mapping of the compensated peer-marks awarded to the weighted average feedback indexes for the essays. This results in a ‘slightly’ better positive correlation between the marks awarded and the weighted comments of 0.883. The average difference of the average standard deviations within the weighted feedback indexes to the compensated peer average marks has reduced to 3.05.

Table Two

-20	-16	-12	-8	-4	0	4	8	12	16	20	24	28	32	36
29	44	48	49	51	58	64	60	64	69	67	73		82	
	38	48	47	49	53	57	60	62	69	66	69			
			41	46	51	54	57	59	66	65	69			
				45	51	53	57	52	62	64				
					50	49	53	50	62	64				
							53		57	63				
										59				
										57				
0	4.2	0	4.16	2.75	3.21	5.59	3.14	6.15	4.71	3.44	2.31		0	
29	41	48	46	48	53	55	57	57	63	63	70		82	

Therefore the results shown in the previous tables (1&2) both resulted in a very positive correlation between the marks and comments (both weighted and un-weighted) for the peer-marking of the essays.

It is interesting to assess whether certain groups of students, based upon their ability within this assignment area, tended to be over- or under-markers and/or -commenters. Using their essay grades achieved as an assessment of their knowledge, Table 3 shows the average differences of each group of students.

Table Three

Range	Frequency of Students	Mark Difference	Feedback Difference	Weighted Feedback Difference
80>	1	0.57	0.58	-3.1
70>	1	-4.8	-1.12	-4.8
65-69	8	-0.93	-0.03	-0.34
60-64	10	-2.00	-0.36	-1.90
55-59	8	0.72	0.08	-0.04
50-54	10	-2.69	-1.2	-3.1
45-49	8	2.14	1.42	5.13
40-44	2	4.17	0.99	5.01
35-39	1	-4.67	-1.84	-1,24
25-29	1	4.6	0.27	-0.87

Table 3 shows that just by a student looking at the basic comments provided for their essay, the true emphasis of the comments may not be fully achieved e.g. 80> gave comments that were on average positive yet by including the weightings were in fact negative.

Table 4 below shows the actual number of comments set up in the database (averaged per number in each grade category). It should be noted that some groups of students tend to have more negative that positive. By using the weightings then it appears to produce a better ratio of positive to negative comments.

Table Four

	# Pos	# Neg	Ratio	Weight Pos	Weight Neg	Ratio
80>	47	48	0.98	144	154	0.94
70-74	42	54	0.78	155	177	0.88
65-69	39.88	55.63	0.72	123.88	148.63	0.83
60-64	41.7	54.2	0.77	136.10	154.80	0.88
55-59	37	49.88	0.74	129.88	143.38	0.91
50-54	39.6	53.7	0.74	142.7	158	0.91
45-49	33.63	46.38	0.73	122.38	150.88	0.81
40-44	32	34	0.94	125.5	118.0	1.06
35-39	40	57	0.70	96	57	1.68
25-29	39	54	0.72	116	156	0.74
			0.77			0.91

Table Five

Category of comments	# Pos	# Neg	Ratio	Total	Weight Pos	Weight Neg	Ratio	Total
Readability	4.58	3.72	1.23	8.3	13.76	10.42	1.32	24.18
Aimed at Correct Level	1.5	4.62	0.32	6.12	5.12	12.48	0.41	17.6
Personal Conclusion	6.86	6.12	1.12	13.0	22.22	18.86	1.18	41.08
Referencing	3.92	3.88	1.01	7.8	13.62	12.58	1.08	26.2
Research & Use of Sources	3.08	6.24	0.49	9.32	11.5	17.5	0.66	29.0
Content & Explanation	4.72	6.74	0.70	11.5	16.26	19.74	0.82	36
Examples & Case Study	3.82	4.62	0.83	8.44	12.78	13.4	0.95	26.18
Report Quality	4.64	4.62	1.00	9.26	16.58	15.06	1.10	31.64
Intro & Definitions	2.34	4.62	0.51	6.96	8.04	13.86	0.58	21.9
Presentation	3.12	6.32	0.49	9.44	11.3	15.12	0.75	26.42
				90.1				280.2

Table 5 above looks at the comments databases and works out the average number of comments etc by each assessment category.

Table 6 presents the various average differences for each category of student for both their marking and commenting, with the overall maximum range of mark consistency being 11.41 to 1.4, feedback consistency being 9.54 to 0.57 and weighted feedback consistency being 27.75 to 6.0.

It should be kept in mind that to attain a GOOD grade for marking, then the student’s consistency factors should be low.

Table Six

Range	Frequency	Mark Difference	Mark Consistency	Mark Consistency Ranges	Feedback Difference	Feedback Consistency	Feedback Consistency Ranges	Weighted Feedback Difference	Weighted Feedback Consistency	Weighted Feedback Ranges
80>	1	0.57	**8.32**	8.32	0.58	**3.62**	3.62	-3.1	**11.93**	11.93
70>	1	-4.8	**10.48**	10.48	-1.12	**4.18**	4.18	-4.8	**8.0**	8.0
65-69	8	-0.93	**7.22**	11.41-4.77	-0.03	**3.05**	5.06-1.98	-0.34	**11.51**	15.60-6.32
60-64	10	-2.00	**5.90**	10.20-1.63	-0.36	**2.89**	4.36-0.57	-1.90	**12.57**	15.89-9.13
55-59	8	0.72	**5.48**	9.74-2.97	0.08	**2.88**	4.44-1.83	-0.04	**16.12**	27.75-7.63
50-54	10	-2.69	**7.20**	10.64-1.4	-1.2	**4.55**	9.54-2.04	-3.1	**11.37**	16.24-6.0
45-49	8	2.14	**5.61**	6.79-3.49	1.42	**2.43**	3.37-1.26	5.13	**15.56**	24.2-8.41
40-44	2	4.17	**4.56**	5.42-3.68	0.99	**1.81**	2.38-1.25	5.01	**16.35**	17.69-15.01
35-39	1	-4.67	**3.73**	3.73	-1.84	**2.7**	2.7	-1,24	**7.27**	7.27
25-29	1	4.6	**5.78**	5.78	0.27	**2.12**	2.12	-0.87	**15.6**	15.6

There are various ways that a mark/grade may be awarded for the student marking performance.

1) A linear scale could be used to award marks from 0 – 100. However this would not normally map to marks awarded within higher education.
2) The actual essay grades awarded would be a better reflection of the quality of this group of students, therefore a mapping of the mark for marking to the frequency distribution of marks awarded for these essays could be used.
3) A particular expectation of a set of normalised results may be expected for a particular cohort within a module, this could also be used.

Whichever method is used, it can be easily included within the automatic creation of a mark for marking within the server aspect of the CAP system.

Table 7 below shows the frequency distributions for the consistency factors, and also the awarded essay grades.

Table Seven

	0	1	2	3	4	5	6	7	8	9	10	11
Mark Consistency	0	2	2	5	4	10	8	5	4	2	4	1
	0	1	2	3	4	5	6	7	8	9	10	11
Feedback Consistency	1	6	21	8	5	4	1	0	0	1	0	0
	6	8	10	12	14	16	18	20	22	24	26	28
Weight Feedback Consistency	6	7	6	8	13	2	1	0	1	2	1	0
	80+	75	70	65	60	55	50	45	40	35	30	25
Essay Compensated Peer-Mark	1	0	1	8	10	8	10	8	2	1	0	1

Conclusions

The results of the mapping of the compensated peer-marks to the average feedback indexes are very positive. Although the weighted development of the average feedback index only produces a slight improvement to an already very positive correlation, it addresses a concern that the subjectivity of the comments derived from the menu driven system were not totally subjective.

The main concern of this method of automatically developing a mark for marking & commenting is the mapping of the consistency factors to an absolute grade. The method used of referential mapping although possibly not being totally statistically acceptable is easy to explain to students. It should be kept in mind how difficult it currently is to explain to a student why they have been awarded 69% and their colleague has 71% within a *traditional* assessment.

The important outcome of an assessment is that a student knows what mark they have attained and why. By maintaining **relative** simplicity in the production of both the compensated peer derived mark for the essay, and also the consistency derived mark for marking and commenting, the students are able to assess their own strengths and weaknesses for future work not only in this subject area, but also in becoming more reflective of their own work.

It should be noted, the tutor has had no input into the generation of the marks for this particular assignment. There were no objections raised by the cohort of students concerning this matter.

References

Bhalerao, A. and Ward, A. (2001) ‘Towards electronically assisted peer assessment: a case study’, *ALT-J*, 9, 1, 26-37.

Davies, P. (2000) Computerized Peer Assessment, *Innovations in Education and Training International*, 37, 4, 346-355.

Davies, P. (2001) Computer Aided Assessment MUST be more than multiple-choice tests for it to be academically credible?, *Danson, M. & Eabry, C. (eds) Proceedings of the 5th International CAA Conference, July 2001*, 143-150.

Davies, P. (2002) Using Student Reflective Self-Assessment for Awarding Degree Classifications, *Innovations in Education and Teaching International*, 39, 4, 307-319.

Davies, P. (2004) ‘Don’t Write, Just Mark; The Validity of Assessing Student Ability via their Computerized Peer-Marking of an Essay rather than their Creation of an Essay’, *ALT-J*, 12, 3, 263-279.

Davies, P. (2006) Peer-Assessment: Judging the quality of student work by the comments not the marks?, *Innovations in Education and Teaching International*, 43, 1, < *in press*>.

Falchikov, N. (1995) Improving Feedback To and From Students, Assessment for Learning in Higher Education, *ed. P Knight, London Kogan Page*.

Lin, S.S.J., Liu, E.Z.F. and Yuan, S.M. (2001) Web-based peer assessment: feedback for students with various thinking-styles, *Journal of Computer Assisted Learning*, 17, 430-432.

Parsons, R. (2003) Self, Peer and Tutor Assessment of Text Online: Design, Delivery and Analysis, *Christie, J. (ed) Proceedings of 7th International CAA Conference, July 2003*, 315-326.