The automatic generation of ‘Marks for Marking’ within the computerised peer-assessment of essays

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Additional Information:

- This is a conference paper.

Metadata Record: [https://dspace.lboro.ac.uk/2134/1908](https://dspace.lboro.ac.uk/2134/1908)

Publisher: © Loughborough University

Please cite the published version.
THE AUTOMATIC GENERATION OF ‘MARKS FOR MARKING’ WITHIN THE COMPUTERISED PEER-ASSESSMENT OF ESSAYS

Phil Davies
The Automatic Generation of ‘Marks for Marking’ Within the Computerised Peer-Assessment of Essays.

Phil Davies
School of Computing
University of Glamorgan
South Wales

p.davies@glam.ac.uk
01443 – 482247

Abstract
Peer-Assessment of essays is not new to higher education. Over the last few years a few web- and internet-based systems have been developed in order to support and manage the peer-assessment processes. There is significant research that supports the use of peer-assessment, not only as an evaluative tool, but also as an aid to the learning process. However, it should be noted that the peer-marking process is one that requires a significant time resource from a student. Therefore, the gain in following this evaluative process of marking their peers must result in some form of tangible reward that relates to the quality of the marking process shown by the student.

The CAP (Computerised Assessment by Peers) system has been used for a number of years in order to support both the self- and peer-assessment of essays. It incorporates a facility where students may anonymously question their peers with regard to marks and comments produced. In previous uses of the system, a mark has been subjectively awarded by the tutor with regard to ‘rewarding’ the student for the quality of their marking (i.e. their demonstration of evaluative skills). This has proven to be a long and time consuming task for the tutor.

This paper highlights how this allocation of a ‘mark for marking’ can be fully computerised, in order to reward the student both for the quality and consistency of the marks they award, and also for the comments that they supply to the essay owner.

Keywords

Peer-Assessment, Evaluation, Marking.
Background
The benefits of peer-assessment are well documented with respect to student benefits (Boud et al, 1999; Falchikov, 1995). Brown et al (Brown et al, 1997) have noted that students do learn a great deal from working with their peers and therefore it would make sense that they also learn from sharing the assessment process. These benefits are not only limited to students, but can also benefit the staff involved in the assessment process (Pond et al, 1995).

The use of computerised peer-assessment of essays is now becoming a reality with a number of systems having been developed throughout the higher education sector (Davies, 2000; Bostock, 2001; Bhalerao & Ward, 2001). These computerised systems provide a framework for the peer-assessment processes to be built around, and supply management tools that reduce the need for tutor involvement.

Robinson (2002) notes that peer-assessment is of great benefit to peers, but highlights the point that the amount of work involved in providing a peer-assessment environment is significant, with not all of the peer comments being of use to the essay owner.

The CAP (Computerised Assessment by Peers) system has been presented as a means of facilitating the anonymous marking of essays by peers (Davies, 2000). It has also been utilised as a means of supporting self-assessment and reflective self-assessment (Davies, 2002). The system permits students to email in their essays and then to have them marked by their peers. The students initially utilise the system as a means of self-assessing their own work, and this provides a mark that acts as a standard for assessing the critical quality of the student’s marking .... standard of expectation.

The students use the system to peer-comment and peer-mark a number of essays (normally eight). The system includes embedded web browsing, and has been enhanced to permit the students to anonymously discuss the comments with their peers before agreeing to these comments, and hence viewing the allocated marks. This process has anonymity maintained throughout (Davies, 2003).

The benefits of utilising such methods of computerised peer-assessment have been well documented however the heavy burden of marking is transposed to the student from the tutor. The students require a ‘tangible’ reward for performing this process of assessing in a critical and authoritative manner. It has been proposed in the past that this aspect of the assessment process is in fact utilising the higher order skills of evaluation, and should be assessed by the tutor as such.
Due to this, the tutor has been involved in the lengthy process of reading each comment, and assessing the marking student's ability in providing comments and marks. In order to take into account high/low markers, a normalisation process is automatically undertaken. This utilises the 'standard of expectation' mark against the actual mark awarded for an essay, and utilises this difference as a standard that the marker should have adhered to throughout e.g. self-assess their work as 65%, peer grade for this student's work equals 60%, hence they have over-marked by 5%, and should do so throughout the marking process of their peers' work.

This statistical formula/process can be easily programmed. What is not so easy is the judgment of the feedback that they have produced in supporting these marks.

Study
This study involves a cohort of 65 students, in a Distributed Systems module at level three of an under-graduate programme within the School of Computing at the University of Glamorgan. Initially the students were provided with an essay to write in the topic area of 'n-tier computing'. Having submitted their essays in electronic format, they then progressed through a process of self- and then peer-assessment (figure one) making use of the CAP system.

![Figure 1](image)

Having completed this process, the students were then permitted to view the comments of their peers. The essay owners could agree to the comments or request further clarification from the marker. If the essay owner agrees to the comments they may then view the mark awarded (figure two).
If they request further clarification, then the marker is automatically emailed and requested to re-mark the essay (figure three). Once this is completed the owner of the essay is automatically emailed and may view the reviewed comments. This iterative process may be followed a number of times, with anonymity maintained throughout (Davies, 2003).
The final mark awarded for an essay is the median mark of all the peer-markings. In calculating this median mark it should be kept in mind that unfairness could appear due to a student being marked by a group of students who always mark hard (or low). If this were the case then their final mark would be less than a similar student who is marked by a group of student who normally mark easy (or high). To take this into account, the CAP system generates a compensated mark for each marking. This views the markings of a particular student, and compares them with the average peer-mark produced for a piece of work. If they consistently mark high, then the average amount they over-mark is taken from the marking before the compensated average is produced.

e.g.
Student B marks five pieces of work

<table>
<thead>
<tr>
<th>Essay</th>
<th>Median Mark</th>
<th>Student B Marks</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>65%</td>
<td>67%</td>
<td>+2%</td>
</tr>
<tr>
<td>X</td>
<td>48%</td>
<td>54%</td>
<td>+6%</td>
</tr>
<tr>
<td>Y</td>
<td>53%</td>
<td>57%</td>
<td>+4%</td>
</tr>
<tr>
<td>Z</td>
<td>61%</td>
<td>59%</td>
<td>-2%</td>
</tr>
</tbody>
</table>

+2% on average

In the case above Student B has on average over-marked by +2%, therefore his/her markings are all reduced by this amount. When the compensated average is calculated for each essay this is then taken into account. The over- and under-markings tend to cancel themselves out and normally makes little difference to the final mark. However, by automatically performing this process, it has reduced any fears that the students have in accepting high and low markings.

In order for the peer-marker to be rewarded for demonstrating skills in evaluation through marking, the tutor would normally view all comments and marks and subjectively produce an overall mark for the marking process (figure four). This takes into account

a) consistency of marks awarded (taking into account the pre-determined ‘standard of expectation’

b) the detail in the comments

c) the consistency in the comments with regard to the marks awarded

d) the re-action to queries from the essay owner
The rewarding of the marking aspect appears to be a process that can be achieved statistically.

e.g.
Student A self-assesses his/her own work as being worth 65%.
When it is marked by their peers it ends up with a mark of 56%.
Therefore, student A’s **standard of expectation** is +9% (i.e. they have over-marked by 9%).

Therefore, they would be expected to over-mark their peers by a similar amount when performing the peer-marking process.

If having performed the marking the results looked like the table below

<table>
<thead>
<tr>
<th>Av. Peer Mark for Essay</th>
<th>Student A gave</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>45%</td>
<td>57%</td>
<td>+12%</td>
</tr>
<tr>
<td>58%</td>
<td>64%</td>
<td>+6%</td>
</tr>
<tr>
<td>63%</td>
<td>50%</td>
<td>-13%</td>
</tr>
<tr>
<td>71%</td>
<td>66%</td>
<td>-5%</td>
</tr>
<tr>
<td>33%</td>
<td>42%</td>
<td>+9%</td>
</tr>
<tr>
<td>56%</td>
<td>64%</td>
<td>+8%</td>
</tr>
<tr>
<td>62%</td>
<td>67%</td>
<td>+5%</td>
</tr>
<tr>
<td>45%</td>
<td>54%</td>
<td>+9%</td>
</tr>
</tbody>
</table>

Would Student A get a good ‘mark for marking’?

a) has shown a reasonable degree of consistency (not perfect)
b) we are looking for +9% throughout
c) the two negative difference marks are actually out by an absolute difference of 22% (-13%) and 14% (-5%)
An algorithm could be created that could automatically go through the marking results and produce a grade for student A.

However, when the initial viewing of the results takes place, the essay owner would only view the comments, not the marks. The feedback produced is as important, if not more important, if the marker is to attain any benefits from utilising this assessment method. Therefore, merely viewing and judging the marks is not enough. The comments must be quantified in some way as to assess their quality.

If the quantification of these comments could result in a numerical value i.e. the feedback index, being created for each commenting, then by comparing the average feedback index for an essay against the feedback index produced by a marker could result in a method that could be utilised in producing a ‘mark for commenting’ as well as a ‘mark for marking’.

**Creation of the Feedback Index**

In order to create a quantified value for the comments, each essay’s comments must be analysed. As they are free text responses this brings into question the efficiency and accuracy of performing content analysis on these markings. It was decided that each student’s marking would be analysed manually rather than attempting to computerise this process. A list of comments was created that were used by each marker.

Although this was a very time consuming process, it provided the author with a good understanding of the types of comments produced, and also the context that they were found within the markings. A number of the comments would naturally amalgamate and could be covered by a single comment e.g. ‘this essay is really good’

‘that was really cool’

‘wow that was choc’

Having performed this task it was surprising that amongst all of the markings there were only 96 comments used having previously amalgamated common types of comments.

On analysing the comments produced it was noted that they led themselves naturally to be grouped into ten common categories:

- Readability
- Aimed at Correct Level
- Personal Conclusions
- Referencing
- Research and Use of the Web
- Content and Explanations
- Examples and Case Studies
- Overall Report Quality
- Introduction and Definitions
- Report Presentation and Structure
Within these categories there were both positive and negative comments. A database of category/comments was then created based upon this feedback content analysis. The next stage of the process was to go through all of the feedback again, making use of a marking tool (figure five). The outcome of this process was to numerically quantify the feedback produced for each peer essay marking.

Figure 5

By utilizing these comments, a quantification of each feedback marking was produced i.e. a feedback index based upon positive and negative comments (figure six). The average feedback index was then correlated against the final peer mark produced for each essay. If positive correlation exists between these figures then

a) it can be assumed that the student feedback maps directly to the marks awarded

b) this index could be used as a method of grading the essays, rather than depending upon the raw marks

c) by utilizing this average feedback index, the quality of an individual marker’s feedback could be judged, and hence provide an immediate mark for the marking process.
The feedback index range produced for the essays was -6.625 to +5.125. The average feedback index grade being -0.60, with a standard deviation of 2.6.

A mapping of final peer marks to feedback indexes was performed:

<table>
<thead>
<tr>
<th>Feedback Index to Peer Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>34</td>
</tr>
<tr>
<td>34</td>
</tr>
<tr>
<td>39</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>34</td>
</tr>
</tbody>
</table>
The results of this analysis produced a significant positive correlation of 0.825 between the mark awarded for an essay and the feedback index generated. If an essay’s feedback index can be used as a yardstick of the quality of an essay, then the feedback index for a student’s marking of an essay can be measured against this yardstick to judge the quality of a student’s commenting. Therefore, we can produce an algorithm that will

a) check a student’s marking against the final mark awarded (possibly taking into account that this student may naturally be a high marker by looking at the standard of expectation)

b) check the feedback index generated by a student against the average feedback index for an essay.

In this way the mark for marking can be generated automatically. However, the content analysis process is very time consuming, and probably beyond the scope of creating an application to provide the feedback index automatically. Having analysed the common comments within the ten categories, it was clear that they were fairly generic and could be utilised for many different essays. The database of comments was already present and the CAP system could easily be amended to provide these comments to the marker via a pull-down menu system (figure seven).

Figure 7
By utilising this menu driven system, the students when marking could make use of pre-set comments for their marking. Each of these comments has a value associated with it, and this numerical value is stored on the marking database for each essay’s marking. This brings into question whether by restricting the marker to fixed comments the benefits to a marker are being reduced? In its initial use the CAP marking permitted both free text input and fixed comment selection from the pull-down menu of comments. The analysis of this data is currently being scrutinised.

**Future Work**
It should be noted that students marking work
a) only tend to use a subset of these comments.
b) from their feedback have a different regard to the weighting of each of the comments with respect to their commenting on the quality of an essay.

With these factors in mind the CAP system is being further developed for each student to be able to input their own comments, both negative and positive, prior to commencing their marking process. Also prior to marking they will be expected to attribute a personal weighting to each comment used. By using this method a more precise ‘weighted feedback index’ can be created for each student marking. Also the need for free text input will be removed as each student is able to input their own comments into the database, and this will remove the need for content analysis of the comments post-marking.

**Conclusions**
The development of the CAP system has provided a means of managing the process of peer-assessment. The addition of the anonymous discussion between marker and essay owner has removed the initial problems of a student requiring further feedback or clarification and not being able to contact the marker. Also the student having to agree to the comment rather than just looking at a mark has improved the students commenting as they know that they must be precise and clear to avoid lengthy discussions with the essay owner. In order to reward a student for performing the marking process in a diligent and constructive manner, then a mark must be awarded. This tutor based assessment has proved to be time consuming and subjective. By developing the ‘feedback index’ generation, this tutor involvement can now be removed, and the awarding of a ‘mark for marking’ can now be fully automated both for quality of marks, and perhaps more importantly for the quality of the comments produced.
References


