A Method of Automatic Grade Calibration in Peer Assessment

John Hamer **Kenneth T.K. Ma** **Hugh H.F. Kwong**

Department of Computer Science
University of Auckland
Private Bag 92019, Auckland New Zealand
J.Hamer@cs.auckland.ac.nz

Abstract

Once the exclusive preserve of small graduate courses, peer assessment is being rediscovered as an effective and efficient learning tool in large undergraduate classes, a transition made possible through the use of electronic assignment submissions and web-based support software.

Asking large numbers of undergraduates to grade each others work raises a number of obvious concerns. How will mark reliability and validity be maintained? Can plagiarism be detected or prevented? What effect will "rogue" reviewers have on the integrity of the process? Will effective learning actually occur?

In this paper we address the issue of grade reliability, and present a novel technique for identifying and minimising the impact of "rogues." Simulations suggest the method is effective under a wide range of conditions.

1 Introduction

Peer assessment is attracting increasing attention from educators looking for new ways of improving learning outcomes in undergraduate courses. Many of the tasks associated with peer assessment are associated with Bloom's (1956) "higher" learning outcomes of *analysis* and *evaluation*. More specifically, literature surveys by Ballantyne, Hughes & Mylonas (2002) and Topping (1998) suggests that peer assessment can:

- help to consolidate, reinforce and deepen understanding, by engaging students in cognitively demanding tasks: reviewing, summarising, clarifying, giving feedback, diagnosing misconceptions, identifying missing knowledge, and considering deviations from the ideal;
- highlight the importance of presenting work in a clear and logical fashion;
- expose students to a variety of styles, techniques, ideas and abilities, in a spectrum of quality from mistakes to exemplars;
- provide feedback swiftly and in quantity. Feedback is associated with more effective learning in a variety of settings. Even if the quality of feedback is lower than from professional staff, its immediacy, frequency and volume may compensate;
- promote social and professional skills;
- improve understanding and self-confidence; and
- encourage reflection on course objectives and the purpose of the assessment task.

Historically, peer assessment has been largely confined to small graduate courses or in tutoring contexts. However, the potential benefits for large undergraduate classes are considerable. In addition to the suggested learning benefits, time saving is also often given as a pragmatic reason in favour of peer assessment (Ballantyne et al. 2002).

Realising the benefits of peer assessment in large, undergraduate classes remains a challenging problem. Issues of particular concern include:

- mechanisms for distributing assignments and collecting reviews;
- maintaining validity and reliability in the grading;
- motivating students to complete the reviews;
- minimising the influence of "rogue" reviewers;
- ensuring anonymity of reviewer and/or the student being reviewed;
- detecting and preventing plagiarism;
- dealing with grading disputes.

The problems of distribution and anonymity have largely been solved through the use of electronic submission and web-based reviewing software (Gehringer 2001, Chapman 2001).

Grading validity can be achieved using *scoring rubrics* (Moskal, Miller & King 2002). A scoring rubric is a descriptive scoring scheme that guides the reviewer in assessing various aspects of the work. Scoring rubrics are typically employed when judgement is required, and are used in a broad range of subjects and activities.

An example of a rubric question is the following:

Followed the Assignment's Directions

Inadequate The paper has no apparent relation to the directions of the assignment.

Needs Improvement Some of the paper follows the directions.

Adequate Most of the paper follows the directions.

Excellent The paper follows the directions precisely. (i.e. the sections are labeled, directions for finding the article are clear, all required information, etc.)

Rubrics can be used on their own, or in combination with free-format comments that allow the reviewer to provide more elaborate feedback.

Plagiarism has become an issue of major concern in recent years, and, while it is in no way confined to peer assessment, detection methods that rely on individual markers clearly cannot be used in this context. We analyse the opportunities for detecting plagiarism with peer assessment in Section 4.

The primary focus of this paper is the problem of rogue reviewers. Every undergraduate class is likely to have a proportion of disruptive or incapable students who will inject random or arbitrary grades into the peer assessment.

Several different responses to this issue are possible. Often, peer assessment is used for formative feedback only, with the grading component undertaken using independent (e.g., graduate student) markers. Unfortunately, this approach also removes one of the main motivations for students to participate in the process. It also negates any time savings, and is likely to increase instructor workloads.

The alternative approach is to have students review several essays[1] and take the average grade. This will greatly improve reliability in the presence of isolated rogue reviewers, but still provides no incentive for students to take the review process seriously. It may also fail to produce reliable grades if even a relatively small proportion of the class act as rogues.

We believe it is important to motivate students to complete the reviews by assigning marks for this activity, and by making these marks at least coarsely reflect the review quality. If these two elements are present, students who review in a conscientious manner will be rewarded and rogue behaviour will be discouraged.

A novel grading algorithm has been developed for this purpose, and is presented in Section 2. We have evaluated the algorithm by simulating a wide variety of class conditions. The results of our simulation experiment are presented in Section 3. Section 4 is a short note on detecting plagiarism, and the paper ends with a summary of related work and our conclusions.

2 The grading algorithm

Let us begin by recalling our objective: we wish to provide a reward to reviewers who participate well by allocating a portion of the assignment mark to the review. This review mark should broadly reflect the quality of the grading.

Each reviewer is assigned a number of essays to grade. Too few reviews will reduce the reliability of the grading, while too many will create too much work for students. We suggest assigning at least five essays, with ten being ideal (although see Section 4 for a more precise recommendation). Assuming each review takes 20 minutes, ten reviews can be completed in about three and a half hours. All the experimental data presented in Section 3 are based on ten essays per reviewer.

Once the reviewing is complete, we have roughly this number of grades for each essay. No doubt some reviewers will fail to complete their allocation for various reasons, but high reviewer participation is a requirement for generating reliable grades.

The algorithm generates two quantities: a *grade* for each essay, and a *weight* for each reviewer. The essay grades can be used directly, just as you would an averaged grade. The reviewer weights need to be interpreted by the instructor, and converted into review marks based on their distribution or other factors. Our experimental results (see Section 3) show that a wide range of reviewer weight distributions can arise, making it difficult to provide more specific guidance in this matter.

To compute an essay grade, we start by averaging the individual grades from the ten or so reviewers. This a "naive" averaging assigns equal weight to each of the reviewers. However, our model assumes that some reviewers will perform better than others, and that we can measure this effect by looking at the difference between the grades assigned by the reviewer and averaged grades. The larger this difference, the more out of step the reviewer is with the consensus view of the class. Accordingly, we adjust the weighting of the reviewers based on this difference. These adjusted weights can then be used to revise the grades.

The calculation of grades and weights is an iterative process. Each time the grades are calculated, the weights need to be updated, and each change in the weights affects the grades. In practice, convergence occurs quickly, typically requiring four to six iterations before a solution ("fix-point") is reached.

In general, there are many fix-points for the grades and weights. For example, setting the weighs of all but one reviewer to zero will form a solution. Starting out with the average grades generally leads to a "reasonable" fix-point, but not always. We return to this issue after presenting the basic algorithm.

2.1 Algorithm details

A more precise specification of the algorithm is as follows. Let

- g_e^r be the grade assigned by reviewer r to essay e;
- R_r be the set of essays allocated to reviewer r;
- E_e be the set of reviewers allocated to essay e;
- W_r be the current weight attached to reviewer r;
- G_e be the current grade for essay e.

We compute G_e as the weighted average of the grades assigned by the reviewers:

$$G_e = \frac{\sum_{r \in E_e} g_e^r \times W_r}{\sum_{r \in E_e} W_r}$$

We can now use G_e to compute the differences between assigned and awarded grades. Dividing by $|R_r|$ accounts for reviewers who do not complete their alloted number of reviews.

$$\Delta_r = \frac{\sum_{e \in R_r} (G_e - g_e^r)^2}{|R_r|}$$

The higher the value of Δ_r, the lower the regard we hold for the reviewer. Weights are assigned in inverse proportion:

$$W_r \propto^{-1} \Delta_r$$

[1] We use the term "essay" to denote whatever work was submitted by the student. It is intended to include computer programs, reports, etc.

Many possible solutions to this proportionality are possible, the simplest of which is to set W_r to $1/\Delta_r$. However, there is a tendency for this calculation to generate very large weights, effectively abdicating the grading to the opinions of a few. For this reason, we have preferred to "dampen" the weight calculation, by considering reviewers with weights above and below the class average.

It is easy to see that scaling every W_r by a constant has no effect on the calculation of G_e. Choosing the mean value of Δ_r gives the equivalent weight calculation

$$W'_r = \frac{\text{mean } \Delta_r}{\Delta_r}$$

The following log-dampened function seems to provide a reasonable behaviour over a wide range of conditions:

$$W_r = \begin{cases} 2 + \log(W'_r - 1) & \text{if } W'_r > 2 \\ W'_r & \text{if } W'_r \leq 2 \end{cases}$$

i.e., we allow weights to rise to twice the class average, with any further increase being awarded sparingly, at a logarithmically dampened rate.

2.2 Properties of the algorithm

To illustrate how the algorithm works, we consider a very small set of essays and reviews.

g_e^r	Reviewer (r)			
Essay (e)	a	b	c	d
1	10	10	9	5
2	3	2	4	5
3	7	4	5	5
4	6	4	5	5

Reviewers a, b and c are in broad agreement with all the essays. However, reviewer d is a "rogue," assigning a median grade (5) to everything.

The weights after running the algorithm are shown in Figure 1. Reviewer c is identified by the algorithm as the most accurate, followed by a and b, and lastly d. The "rogue" has ended up contributing around half the weight of reviewers a and b, and about one sixth of the weight of reviewer c.

Reviewer	Weight
a	1.1
b	1.1
c	3.6
d	0.6

Figure 1: Calculated weights

The marks are fairly close to what they would have been if all the reviews from d were omitted. Figure 2 shows the calculated grades, along with the "naive" averaging and the grades calculated from just reviewers a, b and c. Not much should be read into this data, other than to observe that the grading algorithm is able to both increase and decrease the grades over the naive averaging.

The rogue was perhaps fortunate with this allocation of essays. A few better or poorer essays would have separated the weights further. However, as rogue strategies go, choosing a median grade is usually better than the main alternatives. These include:

- give everyone the maximum possible grade (MAX);
- give everyone the minimum possible grade (MIN);
- give everyone the median grade (MED);
- grade randomly (RND).

Essay	Grade	Naive	a-c only
1	8.9	8.5	9.4
2	3.6	3.5	3.3
3	5.2	5.3	5.2
4	5.0	5.0	5.0

Figure 2: Calculated grade, and two alternatives

If the rogue marker assigns zero marks to each of the four essays, the grades and weightings change to those shown in Figure 3. The rogue now contributes less than one quarter the weight of any other reviewer.

Essay	Grade
1	9.0
2	2.9
3	4.9
4	4.6

Reviewer	Weight
a	2.2
b	2.9
c	3.4
d	0.5

Figure 3: Weights and grades with a MAX rogue

3 Experimental evaluation

In order to explore the performance of the grading algorithm over a range of likely conditions, we performed a simulation. The simulation included a model of student performance and rogue reviewer behaviour.

3.1 Simulation parameters

For the model of student performance, we assigned each essay a "target" grade out of 10, according to an arbitrary but typical grade distribution (see Figure 4). I.e., around 6% of the essays will be worth full marks, 10% will be worth 9/10, 20% will be worth 8/10, etc.

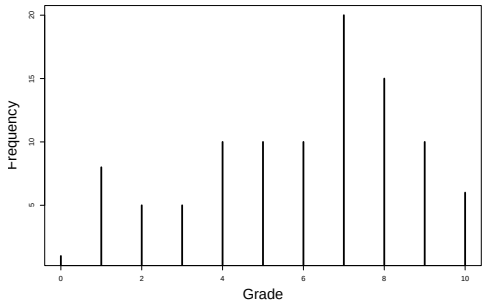

Figure 4: Essay "target" grade distribution

We modelled the performance of reviewers by assuming each to have an intrinsic "variability" between zero and five (see Figure 5). A zero means the reviewer is always accurate — the review will match the "target" grade in all cases. In general, a variability of n means the assigned grade will be randomly chosen between the target plus n and the target minus n, each being equally likely.

In all of the experiments we ran the simulation with 100 essays, with each reviewer being randomly allocated 10 essays.

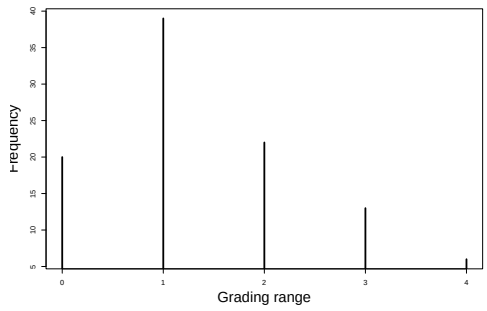

Figure 5: Reviewer variability

3.2 The effect of dampening on accuracy

In this section, we address the question:

> "Does dampening the weights affect the accuracy of the calculated grades?"

We conclude

> "Yes, but not significantly."

The experiment consisted of 200 trials, where each trial calculated the mean difference between "target" and calculated grades for a simulation with a given number of rogue reviewers. This was repeated with undampened and then log-dampened weight calculations.

The box-and-whisker plot in Figure 6 shows the expected range of the average grading error. The box (thick line) is drawn to ± one standard deviation, and the whisker (thin line) to ± three standard deviations of the mean[2]. The plot to the left of each pair is for the undampened calculation, and the plot to the right uses the log-dampened calculation. Overall, the log-dampened weights result in slightly less accurate grading, but (curiously) contribute a little more consistency.

The graph also gives an indication of how much grading reliability is lost as the proportion of rogue reviewers increases. In a class with 40% of the students grading with no regard to the essay quality, about five percent "noise" is added to each essay mark.

Note that this is a just the mean difference over the 100 essays. Individual essays may be marked considerably less accurately. The standard deviation for individual grades difference[3] varies from around 0.2 (5% rogues) to 0.7 (40% rogues).

3.3 Distinguishing rogues from non-rogues

We are most interested in the conditions and extent to which the algorithm is able to distinguish between "rogue" reviewers (who grade with no regard to the essay quality) and conscientious reviewers (who grade to within a few points of the target mark).

Our experiment assumes that each rogue will adopt one of the four "rogue strategies" — MAX, MIN, MID and RND — and that each of these strategies is equally likely. We also repeated the experiment using a fixed strategy for each rogue, and obtained

[2]A quick plot of grading error shows a clear bell-shaped curve, strongly suggesting this distribution is normal. This is supported by the p-values for the Shapiro-Wilk normality test, which range from 0.15 to 0.77.

[3]Again, this distribution appear to be normal.

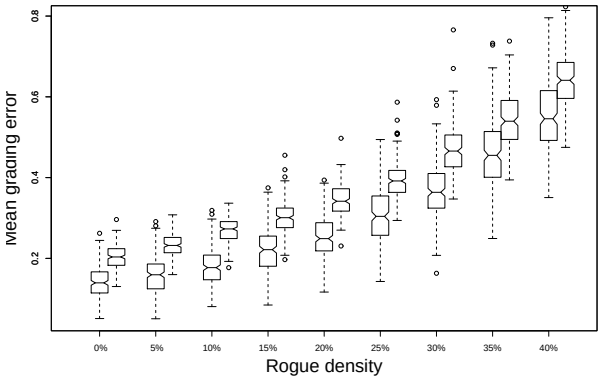

Figure 6: Mean grading error by rogue density. The undampened figures (left of each pair) show a slight improvement over the log-dampened figures (right of each pair)

similar results[4]. It seems more reasonable to assume that rogues will behave independently.

The box plots in Figure 7 show the distributions of reviewer weight for rogue and non-rogue students when between 5 and 50% of the class behave as rogues. At the 5% level, the rogues dominate the lowest weightings. Few of the rogues even reach the "mean" weight of 1.0.

At the 50% level, the majority of the rogues are still at the lowest end of the graphs. As the box plots show, this area is also largely free of conscientious reviewers. As should be expected, a number of rogues are being rewarded with high weightings, but they comprise a small minority (only outlier rogues exceed the non-rogue mean weight).

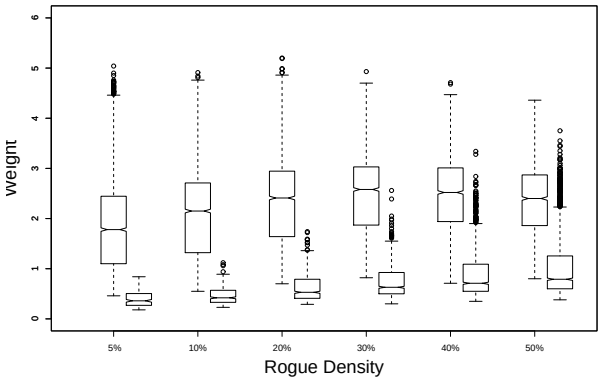

Figure 7: Reviewer weighting by rogue density

In all seriousness, any class that has half the students treating the review exercise in a derisory manner would be foolish to attempt any form of peer assessment. Nevertheless, it is encouraging to think that even under such extreme conditions, the grading algorithm is still largely able to differentiate rogue from non-rogue.

To the extent that these results mimic classroom behaviour, it appears that peer assessment can produce accurate grades. Most poor graders are reliably

[4]The MID strategy is the best choice for a rogue, but even under extreme conditions (e.g., 50% or more of the class adopting the same strategy) conscientious reviewers perform better.

identified and their contribution diminished in favour of accurate reviewers.

4 Preventing plagiarism

Plagiarism, in the form of two or more students submitting identical work, has been a longstanding concern in most institutions. We consider only this form of plagiarism here, and do not address, for example, the individual student copying material from the Internet or engaging the services of a senior student.

The *opportunity* to detect copying is assured when one individual does all the marking. Actual detection is, of course, dependent on the attentiveness of the marker. In the analysis that follows, we will assume that (a) peer markers are more or less equally likely to detect and report plagiarism as independent markers; and (b) actual detection is much more likely to occur when markers have a small number of essays.

Under these conditions, peer marking can be at least as effective at detecting copying in large classes as an independent marker.

For a class of n students, there are $n(n-1)/2$ potentially plagiarised pairs of essays. Even when n is fairly small, this number is quite large. When $n = 30$ there are 435 pairs. When $n = 100$ there are nearly 5,000 pairs. In contrast, a student peer marking ten essays need only consider 45 potential instances of plagiarism.

In practice, $n \approx 100$ is an upper limit for the number of assignments that one individual can mark. Classes larger than this require several markers, and consequentially the opportunity for detecting plagiarism drops. With two markers, half the pairs are only seen by one marker. With ten markers, cheats have at 90% chance of avoiding detection, even when the markers are perfectly attentive.

The analysis for peer marking is somewhat different. Instead of n essays and one or a few markers, we have n reviewers each with a bundle of b essays. Each bundle contains $b(b-1)/2$ pairs of essays, and so there are up to $nb(b-1)/2$ pairs available for detection[5]. The number of pairs that escape the attention of any single reviewer is

$$\frac{n(n-1)}{2} - \frac{nb(b-1)}{2}$$
$$= \frac{n}{2}\left(n - (b^2 - b + 1)\right)$$

i.e., when $n \leq b^2 - b + 1$ every pair can be seen by at least one reviewer, but when $n > b^2 - b + 1$ some pairs pass through unchecked. Setting the number of essays to review to at least $\sqrt{n}$ will ensure that nearly all cases of plagiarism have the potential to be detected.

This coverage is sensitive to small changes in bundle size. For a class of 100, a bundle size of eleven is sufficient to ensure all pairs can be covered. However, reducing the bundle size to ten allows 9% to escape unchecked. This grows to 27% for bundles of nine, 43% for bundles of eight, down to 88% for bundles of four (see Figure 8).

Overall, it appears that peer marking can offer at least as many opportunities for detecting incidents of copying as an individual marker, provided the bundle size is set appropriately. Any reduction in motivation or skill by peer markers is offset by the far smaller number of cases each peer marker needs to consider.

[5] The current allocation algorithm makes no attempt to ensure all possible pairs are included, although it could be made to do so. In practice, a small number of pairs do end up assigned to two reviewers (hence displacing other pairs), but this is rare, accounting for no more than a fraction of one percent of the total.

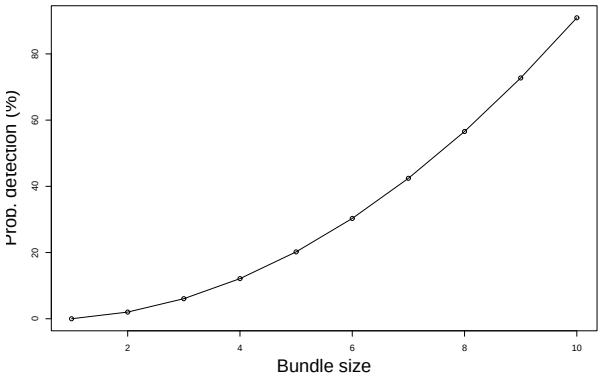

Figure 8: Coverage of pairs with a class of 100

5 Related work

Topping (1998) provides a major survey of the literature on peer assessment prior to 1997. Unfortunately, most work on web-based systems has occurred after this date. No comprehensive survey of web-based systems has been compiled, and given the very broad subject area our own literature search is unfortunately quite fragmentary.

The most widely used web-based system appears to be *calibrated peer review*™(CPR) (Chapman 2001), which has been adopted by more than 300 institutions. In common with our system, CPR uses a grading rubric. The unique feature of CPR is a "calibration" step. Instructors prepare three sample submissions, of high, medium and low quality. Before peer reviewing, students are asked to practice grading the samples until they are consistently able to reproduce the target marks. It is assumed that no "rogue" reviewers remain after calibration. Another report of CPR by Wise & Kim (2004) uses the calibration phase to calculate a "competency index," akin to our reviewer weights. This index, which varies between 1 and 6, is used to weight the grades assigned by the student.

The problem of motivating students to take the marking seriously is also addressed by Sitthiworachart & Joy (2004). Their approach is to have students mark the feedback they receive. Students thus interact in three roles: author, marker, and feedback marker. They claim this helps students develop critical judgement skills, and found that the feedback marks improved over time as students gained experience.

Gehringer (2001) reported a web-based peer review system with similar objectives to our own. His system is distinctive in allowing students to submit arbitrary sets of web pages, thus allowing the inclusion of diagrams and multi-media elements. Feedback is a central feature of his system. After submitting their work, an initial review phase is used to give students an opportunity to make revisions. This is followed by a grading review, which contributes part of the final marks for the assignment. Both the initial and grading review are unstructured comments. Finally, students are given the opportunity to *review the review*. Marks from the review of the review also

counted towards the final assignment mark.

All these systems reflect a concern for the quality of the reviews, and illustrate that a variety of check-and-balance mechanisms can be employed.

6 Conclusions

Peer assessment is on the brink of entering mainstream use in undergraduate courses. The potential benefits to learning are generally accepted, and fit naturally into the research-based orientation of University teaching.

Many of the obstacles to adopting peer assessment in large classes are overcome by the use of electronic submission and web-based support software. We have addressed a major remaining problem of grading reliability in the presence of an unknown and potentially large proportion of "rogue" reviews. Our experimental simulations indicate that our grading algorithm provides a robust solution under a wide variety of conditions.

Finally, a straightforward analysis suggests that peer assessment can provide similar opportunities for detecting plagiarism as an independent marker.

References

Ballantyne, R., Hughes, K. & Mylonas, A. (2002), 'Developing procedures for implementing peer assessment in large classes using an action research process', *Assessment & Evaluation in Higher Education* **27**(5), 427–441.

Bloom, B. S., ed. (1956), *Taxonomy of educational objectives: The classification of educational goals. Handbook I, cognitive domain*, Longmans, Green, New York; Toronto.

Chapman, O. L. (2001), 'Calibrated peer review^{TM}', *http:// cpr.molsci.ucla.edu*.

Gehringer, E. F. (2001), Electronic peer review and peer grading in computer science courses, *in* 'Proceedings of the Thirty-Second SIGCSE Technical Symposium on Computer Science Education', ACM Press, pp. 139–143.

Moskal, B., Miller, K. & King, L. A. S. (2002), Grading essays in computer ethics: Rubrics considered helpful, *in* 'Proceedings of the 33rd SIGCSE Technical Symposium on Computer Science Education', ACM Press, pp. 101–105.

Sitthiworachart, J. & Joy, M. (2004), Effective peer assessment for learning computer programming, *in* '9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education', ACM, Leeds, United Kingdom, pp. 122–126.

Topping, K. (1998), 'Peer assessment between students in colleges and universities', *Review of Education Research* **68**(3), 249–276.

Wise, J. C. & Kim, S. (2004), Better understanding through writing: Investigating calibrated peer review^{TM}, *in* 'Proceedings of the 2004 American Society for Engineering Education Annual Conference'.