Classroom

In this section of Resonance, we invite readers to pose questions likely to be raised in a classroom situation. We may suggest strategies for dealing with them, or invite responses, or both. “Classroom” is equally a forum for raising broader issues and sharing personal experiences and viewpoints on matters related to teaching and learning science.

How to Stop Collusion in Peer Review Exercises: Evidence From the Classroom*

Within classrooms, reciprocal rating or a two-way rating system is frequently used in peer reviews of students’ performances in group projects or assignments. Intentional distortion of assessment can be achieved in such scenarios through careful contracting between participants. This article shows how a modified score function can be used to stop such kind of collusion between students.

In my teaching career for the last eight years, I have frequently used peer review as an effective tool for teaching and learning in the classrooms. Peer review has always resulted in increased participation in the classroom and an increased level of interest in the subject. However, the practice of peer review is not without its problems. It is observed that, if not all, many students often collude between themselves to give higher marks to each other to intentionally distort their final marks towards the higher side. Although this practice may not hinder their intellectual capability of evaluative judgment, it creates problems when the scores matter for a student’s future endeavors. The evaluation of the critical ability of students to judge others’ work also becomes nearly

*Vol.27, No.5, DOI: https://doi.org/10.1007/s12045-022-1379-1

Keywords
Peer review, collusion, score function, assessment.
impossible if collusion takes place between students. To find a way out of this, I improvised how the aggregate marks would be calculated in a peer review exercise and introduced a new score function in the classroom. There was a curious difference in how the students graded each other between when the score function was provided and when it was not. The difference is noteworthy because the score function’s knowledge almost eliminated the collusion between students.

1. Introduction: Reviews and Ratings in Non-sequential and Independent Interactions

There are many instances where interactions or transactions in a market do not follow a sequential process. Hence, they are not dependent on the history of previous interactions. One of the most popular examples of such transaction is the two-way feedback system implemented in many corporate entities [1] for strengthening their across-rank feedback mechanism. Another similar example is the peer review system among students in a classroom activity. The following sections discuss the issue of peer review in the student community in detail, but all the discussions can readily be transported to the office space to connect it with the general two-way feedback system.

1.1 Bias in Peer Review of Group Assignment

In this article, I talk about assessing the intellectual or academic standard of a project. For this exercise to be effective, I assume that the evaluators, students, in this case, have the intellectual ability to make relatively accurate judgments so that the assessment is fair and actually useful as a metric of the academic quality.

1.1.1 Reliability and Accuracy Concerns

Despite the benefits of a peer assessment, the use of self and peer assessment for summative purposes, where students evalu-
ate each other based on their observed performance on the completion of the project [2], is often critiqued as being problematic due to concerns about the validity and reliability of the assessment [3]. Though one of the primary concerns regarding peer assessment for summative purposes is the students’ ability to assess other students’ work reliably, in this article, I assume that the students are capable enough to carry out that task. This ensures the students’ understanding of quality [4] and the capacity for objective evaluation [5]. However, peer pressure within the student community can often lead to biased peer-grading distortion. Such social obligations and the resulted distortions, if not controlled from a design perspective, reduce the acceptability of the process both to the student and the educator community. As a safeguard against social obligations, Vickerman [6] suggests the process of anonymous marking to enhance accuracy, as students might be assured that they will not be identified if they grade low for some under-performers who might carry a significant amount of social capital. Considerable work has been done on designing algorithms and creating options for effective adjustment of scores to increase fairness in peer marking, which seeks to enhance reliability, accuracy, and consistency in peer marking [1, 7, 8]. Conway et al. [7] had proposed the method of ‘individual weight factor’ (IWF) to instill fairness in grading. Sung et al. [8] developed an algorithm to assess each student’s reliability, and based on that, the grade provided by the student is altered to maintain fairness. However, none of these methods address the issue of collusion between two students within a group, provided that all the students are equally capable of evaluative judgment and fairness in judging all others except themselves. In other words, these algorithms are not strategy-proof. These algorithms do not invest in using students’ rationality to stop them from making contracts with each other to achieve an unbiased final rating/grade.

In this article, I assume a scenario where several students will review each others’ work. I don’t want them to evaluate others based on personal affinity or apathy but on an observed idea of merit. In other words, I don’t want any sort of contract to develop
between them to maximize their final marks. I developed a utility function to ensure that no one is benefited by ‘mutual contracts’.

2. Model

Let us say we have \( N \) students. Each student will evaluate every other student and will be evaluated by every other student on the same scale. The final marks obtained by the student will be defined by an aggregate function.

Student \( i \), when evaluated by student \( j \), gets a score \( S_{i,j} \). Now, we define a function ‘score distance’ of student \( i \), \( SD_i \) as follows

\[
SD_i = \sqrt{\sum_{k \in (N \setminus i)} (S_{i,k} - S_{k,i})^2}.
\]  

(2.1)

We construct the final utility (score) of a student \( i \), \( U_i \) as

\[
U_i = Agg_i + SD_i,
\]  

(2.2)

where \( Agg_i = \frac{1}{N-1} \sum_{k \in (N \setminus i)} S_{i,k} \).

(2.3)

SD function is simply a summation of the traditional aggregate scores and the score distance defined in (2.1). The SD function ensures that any collusion or ‘contracting’ action is punished. We call this utility function the SD function. SD function is simply a summation of the traditional aggregate scores and the score distance defined in (2.1). The SD function ensures that any collusion or ‘contracting’ action is punished. For example, if cooperation (collusion), anti-cooperation (defection), and honesty are the three choices for a person, the utility function defined above ensures that if one person wants to collude, the best response of the other person will be to defect, and vice-versa. By cooperation as a choice, I denote the strategy of a student or a group of students to mark the other group with the highest possible marks in the hope of reciprocity. By anti-cooperation or defection, I denote the strategy where a group of students marks the other group with the lowest possible marks to elevate their rank within the class. Finally, by honesty, I go by the traditional meaning of being honest about the grading and not strategizing about the marks.

In the following section, I demonstrate the efficacy of this score function in reducing any kind of cooperative behavior using two classroom experiments.
3. Classroom Experiment

For this article, I conducted two sets of classroom experiments, each with two separate groups of students in my MBA class on game theory. Both the groups of students are familiar with the idea of ‘strategizing’, and hence, understanding the score function was not much of a difficulty for them. The same groups participated once in the experiment without the score function and once with it.

3.1 Peer Review Without SD Function

In this experiment, there were 15 groups in one class, and in another, there were 10 groups who peer-reviewed each others’ assignment submissions. Their final score was calculated using the function $Agg$, defined earlier. I have calculated the score difference between each pair of groups, and the graphs are shown here Figure 1 and Figure 2.

3.2 Peer Review With SD Function

For this experiment, 15 groups graded each other in one class, and in another class, there were 14 groups for the peer review activity. The distribution of the ‘score distance’ between each couple
### Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>$A_{cor}$</th>
<th>$B_{cor}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>0.955907621</td>
<td>0.6311572</td>
</tr>
<tr>
<td>Group 2</td>
<td>0.97532485</td>
<td>0.929750627</td>
</tr>
<tr>
<td>Group 3</td>
<td>0.958272656</td>
<td>0.896310802</td>
</tr>
<tr>
<td>Group 4</td>
<td>0.909114582</td>
<td>0.985333916</td>
</tr>
<tr>
<td>Group 5</td>
<td>0.935611038</td>
<td>0.914646145</td>
</tr>
<tr>
<td>Group 6</td>
<td>0.968374148</td>
<td>0.939211655</td>
</tr>
<tr>
<td>Group 7</td>
<td>0.964079092</td>
<td>0.664853314</td>
</tr>
<tr>
<td>Group 8</td>
<td>0.956151389</td>
<td>0.849432306</td>
</tr>
<tr>
<td>Group 9</td>
<td></td>
<td>0.963717297</td>
</tr>
<tr>
<td>Group 10</td>
<td>0.920056831</td>
<td>0.954920677</td>
</tr>
<tr>
<td>Group 11</td>
<td>0.931358964</td>
<td>0.924384227</td>
</tr>
<tr>
<td>Group 12</td>
<td>0.952046088</td>
<td>0.923442652</td>
</tr>
<tr>
<td>Group 13</td>
<td>0.960511871</td>
<td>0.780277729</td>
</tr>
<tr>
<td>Group 14</td>
<td>0.942983038</td>
<td></td>
</tr>
<tr>
<td>Group 15</td>
<td>0.870701359</td>
<td>0.486388921</td>
</tr>
</tbody>
</table>

Correlation between my grades and students’ peer grading.

of students for each class is shown in Figures 3 and 4. Also, the correlation between their scores for each group and mine is shown in Table 1. Out of 15 groups in Class A, one group marked everyone 0 to relatively increase their own marks, so the correlation could not be computed in Table 1.

### 4. Discussion of Results

The experiment results show a significant difference between the scoring patterns between two kinds of utility functions. From the histogram of Figure 4, we can clearly see that the score differences are not concentrated around 0, and this shows us that collusion was not seen in the peer review exercise in the classroom when I used the SD utility. Histogram of group A, in Figure 3, however, leaves space for some doubt as there is a good amount of concentration around lower score distance. We cannot rule out evidence of collusion just by looking at the histogram. However, in Table 1, it is evident that they have a very high correlation with the grades that I gave independently. Hence, although the score
distances are low, a high correlation with faculty grades clearly eliminates any scope for collusion. Thus, for both groups, colluding between two groups was clearly not an option when they knew the algorithm behind their grading.

Whereas, it is interesting to note that, when aggregate scoring, instead of SD utility, was used as the scoring method in peer review, Figure 1 & Figure 2 shows that there are incredibly high levels of collusion in the marks. The score differences range from 0 to 10 for both classes. This shows that in the absence of any strategy-
Figure 4. Score differences of Group B with SD utility.

proof scoring solution, rational players will naturally collude with each other to secure high marks.

To reiterate the premise of the article, peer review systems in classrooms are often used to inculcate a sense of evaluatory and critical judgment in the students. It also, theoretically, removes the bias from evaluation by a single person, i.e., the teacher. However, one significant challenge in administering peer review exercises is the issue of collusion between the students. As seen in earlier sections, students often collude among themselves to ensure high scores. In this article, I have presented a score function that effectively prevents collusion between two students, given that the students are rational and increasing the total marks is their objective.

References


