Students’ views of group-based work and the issue of peer assessment

EDWARD EDGERTON1 AND JIM MCKECHNIE
University of Paisley

This paper evaluates the introduction of a ‘new’ form of assessment in a final year psychology undergraduate module. The assessment required students to work in groups (N = 3 or 4) to produce a conference-style poster that presented personality data collated from the whole class. The final grade for each student on this assessment reflected a group mark for the poster, a peer assessment mark from their group and their individual performance on a question and answer session on the poster. The evaluation focussed on (i) students perception of this piece of assessment, (ii) the relationship between different elements of this assessment and the module exam, and (iii) issues surrounding reciprocity in group work. The main findings suggested that generally students found the poster comparable to other forms of assessment they had experienced in terms of difficulty and effort required, and that on the whole they were positive about the assessment despite misgivings about group organisation and dynamics. In addition, the assessed elements of the poster did not correlate with performance on the exam, suggesting that they may relate to different skills. Finally, the results suggested that the measurement of reciprocity in peer assessment may be more complex where small groups are concerned.

INTRODUCTION

This paper evaluates the introduction of a ‘new’ form of assessment. In this study fourth level students were required to work in groups to produce a conference poster presenting data that had been collated as part of their Individual Differences module. The final grade achieved for each student reflected a group mark for the poster, a peer assessment mark from their group and a mark reflecting their performance in a question and answer (Q&A) session about their poster.

The drive to expand the forms of assessment employed by academics has been gaining momentum since the late 1980s. This coincided with the growth in student numbers within Higher Education and the emergence of debates about the nature of the learning experience and its aims (Livingstone and Lynch, 2000). Against this background group work has become increasingly popular. This technique is perceived as meeting the demands to develop key skills which employers have identified as important (Healey, Matthews, Livingstone and Foster, 1996). Others have emphasised group work for pedagogical reasons. Parsons and Drew (1996) argue that group learning has been used to emphasise participative and active learning. From this perspective it is argued that students are more likely to learn from experience than from the traditional approach of passing on knowledge through lecture-based forms of teaching.

While there is a growing literature on the benefits of using group work within student assessment, staff and students have expressed concerns (Healey et al., 1996; Parsons and Drew, 1996). Amongst staff, concerns include practical problems of implementing such forms of assessment through to losing sight of the individual student. In the latter case it is felt that weak students may be carried by the group. From the students’ perspective group work is not always welcomed, particularly when it includes being assessed as a group. The dominant perspective encouraged within the prevailing system is individualistic. The introduction of group work and assessment challenges this norm. The individual is now dependent on the group and the effectiveness of the team. In such cases students can, and do, perceive the system as unfair. This is accentuated by concerns over the extent to which team members ‘pull their weight’ (Parsons and Drew, 1996).

Staff concerns over weak students hiding within the group and student worries over equity of effort and commitment have been addressed by employing peer assessment. In this context, assessment by one’s peers is perceived as fairer since they have a greater knowledge of individual group members contribution (Walker, 2001). However, Abson (1994) argued that peer assessment creates an illusion of solving problems inherent within group work. For Abson, peer assessment ultimately fails because students are dissatisfied with their peer’s assessment of them and it can have a negative effect on student's interactions. However, Abson is most concerned about the validity of peer assessment. In creating groups we are also creating ‘social groups’ where aspects of social relationships start to impinge upon the validity of the assessment.

The latter concerns were at the heart of Magin’s (2001) recent work. Magin argues that the main concerns around peer assessment are reciprocity and collusion. The idea that reciprocity will emerge in peer assessment of group based work draws on the research and theories surrounding small group behaviour. In effect the assessments and judgements that are made are influenced by shared values, outlooks and beliefs; reciprocity may not always be a

---

1 Correspondence concerning this article should be addressed to the author at: Psychology Division, University of Paisley, Paisley PA1 2BE, UK. Email: edward.edgerton@paisley.ac.uk
conscious process. The nature of small group work is such that it is likely to enhance this process through mutual interdependence and task sharing (Magin, 2001). A related process is that of collusion. A number of studies have identified concerns over collusion where students in effect agree to award similar marks to each other (Parsons and Drew, 1996; Rafiq and Fullerton, 1996). The net effect of reciprocity and collusion is that they raise questions about the reliability and validity of peer assessment.

Against this background the present case study has three aims. The first is to evaluate students' views on this form of assessment and its individual elements. Evaluation of such exercises within psychology is rare. A notable exception to this pattern is Walker's (2001) recent work. In her paper Walker considers students' views on peer assessment and group work amongst a cohort of first year undergraduates. The present study focuses upon fourth level honours students. It could be argued that this group will view the introduction of such forms of assessment differently given that it is their final year of study.

Second, one of the justifications for introducing ‘new' forms of assessment is the belief that they extend students' skills. If different forms of assessment are in fact assessing different skills then student performance across the different parts of assessment should vary (Billington, 1997). In the present case study students are assessed on four elements: the group assessed poster element, the peer assessment, the Q&A assessment and the final end of module exam. The relationship between performance on each of these elements will be investigated to consider the extent to which different parts of the assessment may be evaluating different skill bases.

The third aim is to consider the issue of reciprocity within the peer assessment element. To address this issue information was collected on the marks awarded within each group by each group member.

**THE PRESENT STUDY**

The assessment required students to work in small groups (N = 3 or 4) to prepare a poster presenting their analysis of data related to the ‘big-5' theory of personality. Group membership was based on students forming their own groups; however, in a small number of cases the lecturers facilitated group formation. Workshops were held on poster design and criteria for peer assessment.

The poster constituted the coursework element of the Individual Differences module and accounted for 40% of the final module mark. There were three elements to a student's final coursework mark. The first element was a group mark awarded for the poster. This was assessed by the course tutors and weighted at 80% of the final mark. The second element was a peer assessment mark. Students rated a group member’s contribution on a scale from zero to ten. The mean of the ratings given to any one student by the other students in the group was calculated and this contributed 10% to that student's overall mark. The third element was based upon the individual student’s performance in a Q&A session based on the group's poster. This was tutor assessed and weighted at 10% of the final coursework mark. The remaining 60% of the final module grade was assessed by a class exam that could relate to any material covered in the module.

It can be argued that this form of assessment had a high novelty value for these students. While they had participated in group work before, they had not been involved in any group assessment or peer assessment exercises. Similarly while students had been involved in presenting their work in tutorial settings the idea of poster presentations was new to them.

Student views on this assessment were collected by means of questionnaire developed by the course tutors. This consisted of a mixture of closed and open questions exploring students' perception of the task's level of difficulty, comparison with other forms of assessment, mark allocation and the perceived benefits and problems of this type of assessment.

The analysis is based on two sets of data. An initial cohort of students (2000-1; N = 28) provided feedback on their perception of this assessment. This cohort of students drew attention to the issue of peer assessment. In order that we could investigate concerns over this issue the exercise was repeated with a second group of students (2001-2; N = 41) where information was gathered on each student’s assessment of their peers.

In considering peer assessment Magin (2001) proposes a specific statistical technique to investigate reciprocity, while acknowledging that this measure may also include the effect of collusion. This study adopts Magin’s analysis while raising concerns about its usefulness in addressing reciprocity. Magin (2001) argues that the results from his study and those of Montgomery (1986) demonstrate that fears about reciprocity and collusion in peer assessment are not supported. The present study mirrors Magin’s in that the groups worked together over a considerable time period. Unlike Magin, who had groups of nine to eleven students, the groups in this study consisted of three or four students. It is conceivable that group size may influence the peer assessment process.

**Student feedback about the poster assessment**

*Weightings of coursework and poster elements*

Students in the initial cohort were asked to indicate what they thought the weightings of these three elements should be. It can be seen from table 1 that for all three elements, the majority of students agreed with the actual weightings used. Where differences were expressed, there was a bias towards a lower weighting for the poster and a higher weighting towards the Q&A session. The answers to the open-ended questions indicated that students felt the Q&A session was a better reflection of individual knowledge and effort.
Table 1
Student responses to weighting of different coursework elements

<table>
<thead>
<tr>
<th></th>
<th>Lower weighting</th>
<th>Same weighting</th>
<th>Higher weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poster</td>
<td>9 (32.1%)</td>
<td>17 (60.7%)</td>
<td>2 (7.2%)</td>
</tr>
<tr>
<td>Q&amp;A session</td>
<td>1 (3.7%)</td>
<td>15 (55.6%)</td>
<td>11 (40.7%)</td>
</tr>
<tr>
<td>Peer assessment</td>
<td>7 (25%)</td>
<td>16 (57.1%)</td>
<td>5 (17.9%)</td>
</tr>
</tbody>
</table>

Comparison of poster with other types of coursework
From table 2, it can be seen that the poster was perceived as requiring more effort than essays, critical reviews and presentations and less effort than individual projects. It should be noted however, that with respect to essays, the majority of students did not find that the poster required more effort. For example, 50% of students thought the poster required the same or less effort than essays. Finally, as expected, the poster was considered comparable to the effort involved in other group work.

Students were also asked to compare the poster with other types of coursework in terms of the degree of difficulty. Table 3 demonstrates that the majority of students did not perceive the assessment as being ‘more’ or ‘much more’ difficult than other types of coursework. The most appropriate summary would be that posters are perceived as being comparable to other types of assessment in terms of difficulty.

Table 2
Comparison of poster with other forms of assessment in terms of the amount of effort involved

<table>
<thead>
<tr>
<th></th>
<th>Much more effort</th>
<th>More effort</th>
<th>The same</th>
<th>Less effort</th>
<th>Much less effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essays</td>
<td>6 (21.4%)</td>
<td>8 (28.6%)</td>
<td>6 (21.4%)</td>
<td>8 (28.6%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Critical Reviews</td>
<td>6 (24%)</td>
<td>7 (28%)</td>
<td>5 (20%)</td>
<td>5 (20%)</td>
<td>2 (8%)</td>
</tr>
<tr>
<td>Presentations</td>
<td>7 (25%)</td>
<td>9 (32.1%)</td>
<td>9 (32.1%)</td>
<td>3 (10.7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Group Projects</td>
<td>2 (7.4%)</td>
<td>6 (22.2%)</td>
<td>17 (63%)</td>
<td>2 (7.4%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Individual projects</td>
<td>3 (10.7%)</td>
<td>8 (28.6%)</td>
<td>5 (17.9%)</td>
<td>11 (39.3%)</td>
<td>1 (3.6%)</td>
</tr>
</tbody>
</table>

Table 3
Comparison of poster with other forms of assessment in terms of degree of difficulty

<table>
<thead>
<tr>
<th></th>
<th>Much more difficult</th>
<th>More difficult</th>
<th>The same</th>
<th>Less difficult</th>
<th>Much less difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essays</td>
<td>5 (17.9%)</td>
<td>6 (21.4%)</td>
<td>9 (32.1%)</td>
<td>8 (28.6%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Critical Reviews</td>
<td>4 (16.7%)</td>
<td>4 (16.7%)</td>
<td>11 (45.8%)</td>
<td>3 (12.5%)</td>
<td>2 (8.3%)</td>
</tr>
<tr>
<td>Presentations</td>
<td>4 (14.8%)</td>
<td>8 (29.6%)</td>
<td>11 (40.7%)</td>
<td>4 (14.8%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Group Projects</td>
<td>1 (3.6%)</td>
<td>5 (17.9%)</td>
<td>20 (71.4%)</td>
<td>2 (7.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Individual projects</td>
<td>1 (3.6%)</td>
<td>11 (39.3%)</td>
<td>7 (25.0%)</td>
<td>9 (32.1%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

The ‘best’ and ‘worst’ aspects
In identifying the ‘best’ aspects of the assessment the majority of responses can be summarised as falling into three main categories (these accounted for nearly three-quarters of all responses):

- product of group work, for example more ideas produced (32% of all responses)
- a ‘novelty aspect’, that is doing something different from traditional assessments (24% of all responses)
- gaining a new skill/more knowledge, for example “a new way of presenting information” (16% of all responses)

The majority of responses relating to the worst aspects of the poster can also be summarised as falling into three main categories (these accounted for 82% of all responses):

- problems with group organisation and dynamics, for example arranging times to meet (36% of all responses)
- assessment issues, for example “unfair to have a group mark in honours year” (25% of all responses)
- task difficulties, for example “knowing what information to include/exclude” (21% of all responses)

The poster as a learning experience
Students were asked to rate how useful they thought the poster assessment was as a learning experience in terms of gaining an understanding of the ‘big-5’ theory of personality; this ranged from 1 “not at all useful” through to 5 “extremely useful”. The median response to this question was 4 (mean = 3.32); this indicates that students found the exercise useful in gaining an understanding of the ‘big-5’ theory of personality.
The second aim of this paper was to investigate the relationship between the three assessed elements of the poster and the exam performance. In addition the relationship between the students ‘global’ poster mark (that is, all three elements combined) and the exam was considered. The exam refers to the end of module exam which reflects the student’s performance across the whole module. Pearson’s correlation analysis indicated that the only significant relationship was between the peer assessment mark and the individual student’s assessment on the question and answer session ($r = 0.497, p = 0.002$).

Peer assessment

Students in the 2000-1 cohort raised concerns about the issue of peer assessment. This resulted in the replication of the study in 2001-2 to consider whether the peer assessment marks indicate an element of reciprocity among students.

Magin (2001) proposes an operationalisation of this using a statistical technique. Here, the mark student ‘A’ gives student ‘B’ is compared with the mark student ‘B’ gives student ‘A’. This is done by converting the score given by raters into a z-score, that is what score did they give each group member in terms of their average score and standard deviation as a rater? Once this has been converted to a z-score the reciprocal pairs can be analysed to determine the significance and magnitude of the reciprocity effect. This is done for all reciprocal pairs in a group and allows for the calculation of a ‘reciprocity effect’ in terms of the coefficient of determination ($R^2$). An example of this analysis for one of the poster groups is given in table 4.

From this table the six pairs yields a correlation coefficient ($R$) of +0.222 (where $R = \frac{Z_{ij}Z_{ji}}{N}$) and accordingly a coefficient of determination ($R^2$) of 0.049; that is the contribution of reciprocity effects to the total variance of scores in this group is only about 5%.

A major problem with this statistical operationalisation of reciprocity is that the analysis cannot be carried out when students give the same mark to all their fellow group members, that is reciprocating pairs’ coefficients cannot be calculated when the standard deviation for a rater is equal to zero. This can be seen in the case of poster group 1 (table 5).

The example in table 5 is typical of the majority of the groups in this sample, occurring in 11 out of the 13 groups. Clearly this statistical analysis is inappropriate for the present study. This may be due to the small group size (maximum of four); however, it may also be due to the way that reciprocity has been statistically defined. Magin (2001) excluded all pairs from his analysis where a group member had given the same mark to all their fellow group members. However, in excluding these cases Magin may have lost information about the way students assign marks within a peer assessment exercise.

DISCUSSION

One of the aims of this study was to consider students’ views on the assessment. The majority of students reported that this type of exercise was comparable with other forms of assessment they had experienced in terms of ‘difficulty’ and ‘effort’. However, students did have concerns over the assessment. While the majority of students agreed with the weighting of the individual elements (80:10:10) just under one third (32.1%) felt that the marks awarded for the poster should be given less weighting. Amongst this group written comments suggest concerns over the ‘group mark’ and the fact that weaker students can be ‘carried’ in such an exercise.
Similar concerns were expressed by students who wished to increase the weighting of the Q&A element (39.2%). Students indicated that this section should be increased because it ‘reflects individual ability more’ or that they felt this aspect shows ‘those who contributed most’. For some, increasing this element was viewed as a mechanism to ‘eliminate bias in peer assessment’. While the majority of students felt that the weighting of the peer assessment element was acceptable, 25% wished this element lowered. The reasons given were: ‘marking should be done by staff only’, ‘friendship bias’ and that it was ‘difficult to be honest without starting arguments’ or that they ‘weren’t comfortable assessing peers’.

Such comments resonate with Walker’s (2001) findings amongst first year psychology students. In her study Walker found that students concerns over peer assessment included the issue of bias, judging friends, confidentiality and problems of fairness. However, even with these concerns in the present study a small minority of students (17.9%) wanted to raise the weighting given to the peer assessment element. The reasons given for this included the fact that the weighting was ‘too low considering the aggravation it causes’ while others suggested that it should be raised to ‘increase the incentive to participate’.

These themes re-emerge when students were asked to identify the ‘worst’ aspect of this exercise. The ‘worst’ aspects included problems surrounding group organisation and dynamics, task difficulties and general concerns over the different elements within the assessment. The group dynamic problems are of interest since in this study students had the opportunity to select who they worked with. It would appear that this type of exercise has the potential to create strains between students who may know each other relatively well.

There is evidence that students perceived the exercise as being of some value. In identifying the ‘best’ aspects of the experience within this study students highlighted the benefit of working in a group, the novelty of the assessment and the opportunity to use a different range of skills in the poster exercise. In addition students indicated that they perceived the exercise as a good learning experience. Healey et al. (1996) also found that students perceived that a number of skills had been enhanced by their group work experience. However, such subjective interpretations need to be supported by more objective indicators of skill development.

This was part of the motivation for the second aim. To what extent do the elements in this module assess different student skills? Billington (1997) showed that student performance on an individual poster task and essay assessment were not correlated. For Billington this was evidence that these two forms of assessment were evaluating different skills. In the present study there was no consistent relationship between the constituent elements of assessment across the module. The exception was a significant correlation between peer assessment grades and the tutor assessed grade awarded for the Q&A session. One explanation for this relationship may be that students rated highly by their peers were more knowledgeable and this was reflected in the Q&A session. For Billington’s argument, the important result is the lack of relationship between the peer assessment, Q&A, exam performance and poster mark. Some degree of caution is needed in interpreting this pattern of results. For example it should be noted that the final exam was assessing a broader knowledge base and took place after the poster had been submitted. An additional caveat is that the correlation analysis only indicates the extent to which different forms of assessment may evaluate different skills. It does not demonstrate that a specific type of assessment has developed specific skills.

The final aim related to the issue of peer assessment. To what extent are student comments about ‘bias’ in peer assessment justified? Magin has argued, based on his statistical approach, that there is no basis for concerns about reciprocity. In this study the lack of variability in students ratings excluded the use of Magin’s technique. The issue then becomes one of explaining this lack of variability in students’ ratings. There are three possible explanations for this.

First, it could be seen as a form of collusion where group members agree to give each other the same marks, especially where this led to a high overall mark. Previous studies have drawn attention to this behaviour (Parsons and Drew, 1996; Rafiq and Fullerton, 1996). Second, a student may have given the same mark, whether agreed or not with the other members of their group, in order to avoid conflict/disagreement. There is some anecdotal evidence to support this scenario where students have made comments about feeling ‘uncomfortable’ or ‘intimidated’ about assessing their peers. Third, the lack of variability may reflect the groups and individual student’s lack of discrimination in applying the marking scheme. With respect to the latter point Walker (2001) had students assign marks on four dimensions. In this study students were provided with information detailing behaviours they should consider when evaluating their peers but were only instructed to assign a ‘global’ score. It is possible that students were unclear about the specific aspects of behaviour they were assessing or how to combine the constituent elements into a final grade. Magin’s analysis is unable to offer any insight into these explanations.

It is conceivable that the lack of variability found within this study is related to the group size. In contrast to Magin this study used smaller groups (N<5). Within small groups there may be more opportunity for the type of behaviour outlined in the first two explanations outlined above. In addition, these behaviours may be even more likely where group membership has a high level of student self-selection as in this study.

The existence of bias within peer assessment in the present study cannot be dismissed. However, peer assessment may be introduced for a number of reasons. For example, providing students with the
experience of assessing colleagues is, in and of itself, a justifiable activity. Similarly, having an element of peer assessment may increase students’ motivation to participate. While many of the student comments about peer assessment were negative, a minority did mention the idea of increasing the weighting of this element as it would increase students’ incentive to participate. It is apparent that further research into peer assessment is warranted. Future studies of small group work involving peer assessment should investigate the impact of randomly assigning students to groups. In addition, where students are assessing their peers consideration should be given to defining prescribed aspects of behaviour rather than adopting a ‘global’ score approach.

In conclusion this study has identified a number of issues surrounding peer assessment within small group work. In particular the statistical operationalisation of bias requires further investigation. While such issues warrant further attention they should not detract from the benefits of introducing such exercises. Students, in general, responded favourably to this assessment and acknowledged it as a positive learning experience. Furthermore the results indicate that this form of exercise may be assessing a range of skills, which traditional forms of assessment may neglect. In the context of debates around the issue of employability skills such findings are of interest and point to fruitful areas of research for psychology.

REFERENCES


Manuscript received on 5 April 2002
Revision accepted for publication on 19 June 2002