Justice in the Classroom: Peer Assessment of Contributions in Group Projects

Dale Parsons
Department of IT and Electrotechnology
Otago Polytechnic
Dunedin, NZ
dale@tekotago.ac.nz

Lecturers in the Bachelor of Information Technology programme at Otago Polytechnic have been exploring various approaches to assessing an individual's contribution to group work. While there are sound educational and vocational reasons for using group work in higher education, and particularly in information technology related programmes, assessing group work always presents lecturers with certain challenges.

The ultimate aim of this study was to provide a method to acknowledge the quality, impact and efficacy of individual students' contributions to the group, without introducing competition within the group.

1. INTRODUCTION

There are sound educational and vocational reasons for using group work in higher education. Employers value teamwork skills and as part of professional practice a graduate needs to be able to critically evaluate their own work (self assessment) and that of their peers. It is also important that students learn about their effectiveness in a group setting (Lejk *et al.* 1996; Boud et al. 1999; Rafiq & Fullerton, 1996).

Educationally, group work is seen as providing a vehicle through which students can be involved in deep learning, developing their skills experientially and contributing to the skills they will need for life long learning (Bourner *et al.* 2001). Group work emphasises co-operation over competition and can promote respect for other group members' experiences and values (Boud *et al.* 1999).

Unfortunately, introducing group work means introducing group assessment, an area rife with difficulty. A fair mark for a given student should reflect that individual's effort and abilities. When students are assessed on work performed as a group it can be extremely difficult to determine the exact contribution of each group member.

Problems arise when either students or tutors perceive that a student's mark does not truly reflect his accomplishments and understanding. The most common concern is that some students will be "passengers" – students who benefit from a group project without making a sufficient contribution to the work (Bourner and Bourner, 2001). The current study was in fact inspired by a case where a poor student was teamed with a strong student for a group assignment. The assignment – and consequently both group members – received a mark of 92%. There were indications that the stronger student has done the bulk of the work on the assignment. On a later individual assessment, the weaker student scored only 5%. Clearly, allowing this student to receive 92% for work that was not his own is fair neither to the student who carried him, nor to the other students in the class.

There are a variety of approaches to insuring equity in group assessment. Most focus on the need for feedback from the group members, rather than an attempt by the lecturer to determine precise contributions for each individual (Lejk, Wyell and Farrow, 1997; Boud, Cohen and Sampson, 1999; Conway, Kember, Sivan and Wu, 1993). This paper discusses a series of group assessment methods used in courses at the Otago Polytechnic Bachelor of Information Technology program. Each method is an extension of the one before it, modified to accommodate student feedback and tutor concerns. The final methodology provides a relatively sound metric for group assessment, and is currently being used in a number of courses in the BIT.

1.1 TYPES OF EXISTING GROUP MARKING METHODS

There are six main methods currently in use for allocating group marks :

■ Equal Marks: Each group member receives the same mark. (Gibbs, 1992)

■ Task Splitting: Each student contracts to do one of the tasks that make up the group activity and each student is marked individually. (Gibbs, 1992)

■ Pool of Marks: The group members distribute the marks amongst themselves by a process of negotiation. (Lejk *et al.* 1996)

■Base Mark Plus or Minus Contribution Mark: Each student is given a lecturer-generated base mark and a peer generated effort mark, which is added to or subtracted from the base mark. (Conway *et al.* 1993)

■Multiplication by Weighting Factor: The tutor-assigned group mark is multiplied by a peer assessed weighting factor. (Conway *et al. 1*993)

■Holistic Peer Assessment: Each student awards one grade to each of the other group members. This grade reflects their overall impression of that student's contribution to the group effort. (Lejk & Wyvill, 2001)

■Separation of Process and Product: The tutor assesses the product and the peers assess the process and the final mark is a combination of the two with any desired weighting. (Lejk *et al.* 1996)

2. THE FOUR METHODS

2.1 Method 1

We began with a simple Base Mark Plus or Minus Contribution Mark group assessment method. Each student was asked to rate his fellow group members on six different project tasks: researching the topic, report writing, ideas and suggestions, presentation preparation, referencing and handouts and preparing demonstrations.

The required rating scheme is shown in Table 1. For each task, students assigned each group member (including themselves) a score between -2% and 2%; the assigned scores had to sum to zero across all group members. Each student submitted his or her table separately and they did not see what marks their peers had allocated to them.

An example of the rating worksheet with one group's allocated ratings is shown in Table 2.

The group shown in Table 2 is basically in agreement as to which student has contributed the most and least to the assignment. Their order of students is the same; , with all students identifying student 3 as making less of a contribution than students 1 and 2. Student 3'sone

Table 1: Description of each markMethod 1 Rating Scheme

Didn't contribute in this way	Willing but not very successful	Average	Above average	Outstanding
-2%	-1%	0%	1%	2%

student's marks have a lesser smaller range than the other two.

The student's ratings were added to the overall project mark to produce an individual score for each student. In the above example, Students 1 and 2 would have received an additional 1.3%, while Student 3 would have lost 2.6%.

This first method gave students the possibility of being rewarded for a greater contribution to the group. It was relatively simple for the lecturer to implement as there is one product to mark per group, with that result modified by the student's peer assessment ratings.

There were however, several problems with this assessment method. The major problem was the apparent difficulty of the peer assessment task itself. Nearly 20% of the groups had members who either submitted incorrect tables or did not submit a table. Clearly, the task needs to be simplified or automated.

A second weakness was the lack of inter-rater consistency. That is, some groups were in disagreement as to whose contribution was the highest and the lowest. Further consideration had to be given to ways of dealing with such cases.

A third problem was the tendency of students to refuse to rank their group, simply assigning a mark of zero to all members. Figure 1 shows the distribution of marks across all students. As can be seen, the commonly assigned mark was 0, indicating no difference in member contribution. It seems unlikely that group contribution was always as even as reflected by this pattern.

2.2 Method 2

Method 2 used a more holistic approach, in an attempt to correct some of the problems encountered with Method 1. Each student was required to give a single percentage mark to each member of their group, including themselves. This mark could range from +12% to –12%. 0% was defined as meaning " this student was an average contributor". It was a zero-sum system -- all the scores for a group had to add to zero. Therefore the person who contributed the most should have the highest mark and the person who contributed the least should have a negative mark. Each student's mark for the group project was adjusted by the mean of the ratings they received in the peer assessment. Table 3 shows an example peer assessment using Method 2.

Table 2: Method 1 - Multiple Criterion Zero Sum System

Assessor: Student one	Student one	Student two	Student three	Total
Researching the topic	0	0	0	0
report writing	0.5	0.5	-1	0
Ideas and suggestions	0	0	0	0
Presentation preparation	0	0	0	0
Referencing and handouts	1	1	-2	0
Preparing demonstrations	0	0	0	0
Total:	1.5	1.5	-3	0
Assessor: Student two	**Student one**	**Student two**	**Student three**	**Total**
Researching the topic	0.5	0.5	-1	0
report writing	0.5	0.5	-1	0
Ideas and suggestions	0	0	0	0
Presentation preparation	0	0	0	0
referencing and handouts	0.5	0.5	-1	0
Preparing demonstrations	0	0	0	0
Total:	1.5	1.5	-3	0
Assessor: Student three	**Student one**	**Student two**	**Student three**	**Total**
Researching the topic	0	0	0	0
report writing	1	1	-2	0
Ideas and suggestions	0	0	0	0
Presentation preparation	0	0	0	0
referencing and handouts	0	0	0	0
Preparing demonstrations	0	0	0	0
Total:	1	1	-2	0

Method 2 eliminated some of the clerical difficulties encountered with Method 1, but was still flawed. In particular, students reported a reluctance to give negative marks, this meant that they were unable to give a sufficiently wide range of marks.

Students also seemed to be confusing effort and contribution. Students were defining effort as the time spent on the project and assigning marks on that basis, rather than on actual contribution to the final product.

2.3 Method 3

Method 3 was designed to overcome the problems caused by the student's attitude toward negative marks used in Method 2. We moved from a zero sum system to what we have termed a "60-Sum system". In this system, students were free to allocate marks in any way they wished, as long as the total across all group members was 60. Students could thus give a clear penalty to a non-performing student without actually having to give a negative mark.

Each student's Contribution Mark was then converted to a zero sum mark by the lecturer. Their final mark was the overall project mark plus their Contribution Mark. Table 4 shows an example peer assessment using the 60-Sum system.

In this group, Student 2 did very little work towards the group project, as reflected in the marks of Students 3 and 4, and those of student 2 himself. Student 2 has shared the rest of the marks equally between the other

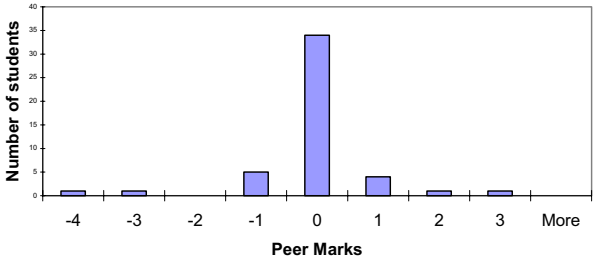

Figure 1: Averaged peer marks

Table 3: Method 2: Single Score Zero Sum System

Marker	Student one	Student two	Student three
Student one	-5	10	-5
Student two	-2	3	-1
Student three	-5	10	-5
Final mark	**-4**	**8**	**-4**

group members, making no attempt to distinguish between the contribution levels of the other students.

Student 1, who was the next weakest student, has awarded the base mark to all group members, even to Student 2. Again, no attempt has been made to use the peer assessment exercise to reflect accurately individual member contributions.

This pattern of marking raises several questions: Should the marks of Student 1 be omitted from the peer assessment computation? Should Student 2 even have been allowed to remain in the group?

Students who used Method 2 voiced concerns about the relationship between the number of group members and the range of marks. They felt that in the 60-sum method, a small group would be likely to opt for a greater spread of marks than a large group. Tables 5 and 6 illustrate this problem.

Objectively, a reward of 5 marks is the same regardless of group size, but the students felt very strongly that this was not the case. They considered the mark was 5 out of 30 for a group of two, and 5 out of 15 for a group of four. Thus 5 was a greater contribution mark for the larger group. Students were also concerned that larger groups were forced to allocate a smaller range of marks.

2.4 Method 4

Student feedback about Method 3 led us to further modify the system, producing what we called the "Flexi-Sum" method. In this system 20 was the base mark, regardless of how many students were in the group. A student who was rewarded for their extra contribution

Table 4: Method 3: 60-Sum System.

Marker	Student one	Student two	Student three	Student four
Student one	15	15	15	15
Student two	20	0	20	20
Student three	10	5	22.5	22.5
Student four	10	5	22.5	22.5
Average	13.75	6.25	20	20
Base mark	15	15	15	15
Final contribution mark	-1.25	-8.75	5	5

Table 5 & 6: A group of two, and A group of four

	A	B
A gives	35	25

	A	B	C	D
A gives	10	20	15	15

would receive a mark of greater than 20. A student who had contributed less than his or her peers would receive a mark of less than 20. A group that deemed all group members had contributed equally would allocate each member a mark of 20. Thus a group of two would allocate a total of 40 marks, a group of four would allocate a total of 80. Table 7 shows an example peer assessment with this method for a group of four. In this example, Student D is perceived as the least contributing member by Students A, B and C. Student D has rated all members as contributing equally. Clearly, Student D's marks are suspect.

Method 4 solved several of the problems identified in the three earlier peer assessment methods. However, some thorny issues still remain. Primary among them is the question of mark variability. How close must the agreement be among members of the group? At what point should the lecturer intervene, and ask the group members to reassess?

Table 8 shows an example peer assessment, which is in order agreement but not magnitude agreement. That is, all members agree that Student C was the weakest contributor, but Student A sees the difference as being very large, and Student C sees it as being fairly minor. Clearly we need some method for quantifying this inter-ranker variability, and identifying when it has reached a level where one or more group members must be seen as biased. Identification of such bias would indicate the need for lecturer intervention.

In parallel with the class using Method 4, another class was using the Group Marking project as a software development exercise. These students were to implement a system for automating the entry, computation

Table 7: Method 4 - Flexi-Sum Method

	A	B	C	D	Total
A gives	28	22	18	12	80
B gives	26	22	18	14	80
C gives	26	22	18	14	80
D gives	20	20	20	20	80

Table 8: Showing a group in order agreement

	A	B	C	Total
A gives	30	20	10	60
B gives	25	20	15	60
C gives	22	20	18	60

Table 9: Standard Deviations Of A Group Not In Agreement

	A	B	C	D	Std dev
A gives	23	17	13	7	**5.83**
B gives	21	17	13	9	**4.47**
C gives	21	17	13	9	**4.47**
D gives	15	15	15	15	**0.00**
Std dev	**3.00**	**0.87**	**0.87**	**3.00**	

Table 10: Standard Deviations Of A Group In Agreement

	A	B	C	D	Std dev
A gives	23	17	13	7	**5.830952**
B gives	21	17	13	9	**4.472136**
C gives	21	17	13	9	**4.472136**
D gives	22	16	14	8	**5**
Std dev	**0.83**	**0.43**	**0.43**	**0.83**	

and management of the Group Marking technique. Since computerisation of the system greatly facilitated computation, we used this second project to explore measures of bias. As a first approach, we computed the row and column standard deviations for each group. Row standard deviations show the variability in an individual student's rankings; column standard deviations show the variability between students. Table 9 shows an example Flexi-Sum assessment with standard deviations computed.

In table 9, the first three row standard deviations are acceptably close, but the last is not. This signifies that student D is not in marking agreement with the other three group members. The column standard deviations will all be close to 0 if the marks allocated to each student are roughly the same. Table 10 shows a sample marking where all students are in approximate agreement.

What values are large enough to raise the bias flag? In practice, I found that individual bias was indicated when any column standard deviation was greater than 2.0.

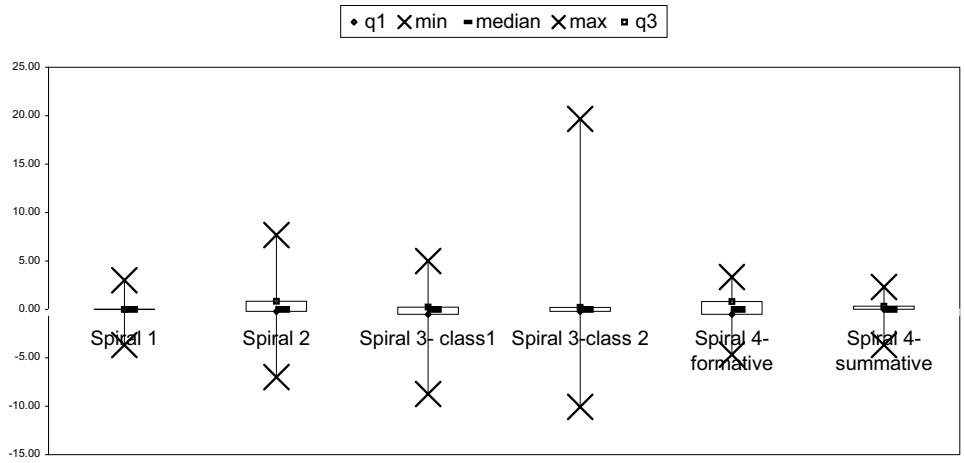

Figure 2: Variability Of Contribution Marks

Interpretation of the row standard deviation flag is more problematic. Row standard deviations could take on any values, they just need to be comparable to indicate general marking agreement. However, since the standard deviation measures only variability and is not sensitive to order, even equal column standard deviations do not guarantee an absence of bias. Further study is needed to establish an appropriate measure of student bias for group assessment

3. GENERAL DISCUSSION

In response to student feedback we progressed from Method 1 to Method 4, with each system providing a more accurate and practical method for peer assessment. While each method resolved some of the problems encountered with the previous ones, difficulties remain. In this section I discuss general issues which apply to all peer assessment methodologies.

3.1. Responses of Participants

On the whole the students were very accepting of having a contribution component as part of their group mark. Most saw the contribution component as a fairer method of distributing group marks than the shared mark method. In Method 2 there were two top students in mixed ability groups, who felt their final mark was lower than they would normally receive. It is a measure of success that in the last method all the students were happy with their individual contribution component. More importantly, when given the option of being marked on their individual effort or the group's product, all the students chose to be marked on their group's product.

3.2 Spread of Marks

Most groups did not give extreme contribution marks. The majority of groups gave scores near zero, effectively agreeing to the same group mark. Figure 2 shows the range of marks for each method. The large range in Method Three was due to a group where two of the three students dropped out. The marks for that group were –10.07, 19.67 and –9.6. Without this group the maximum for that class would have been 4.77, and the minimum –6.13.

In Figure 2 the boxes represent the upper and lower quartiles of the distribution of marks. Thus 50% of the marks for each method lie within the boxes. This indicates that half the members of each class are giving contribution marks clumped around zero. In Figure 3 the near outliers are denoted with the + symbol, and the far outliers with the o symbol. While the bulk of each class is clumped around zero, there were small numbers of quite extreme contribution marks. Thus the system effectively distinguishes very good and very bad performances. This is comparable to the results of Lejk and Wyvill (2001).

3.3 Metric Reliability

One of the vexing questions asked of group assessments are whether "the marks for the product produced by the group can be validly used to infer the competence of the individuals within that group?" (Lejk *et al.* p.276, 1996).

One of the techniques for assessing this is to look at the relationship between each student's individual marks and their group work mark. This statistic is only valid to the extent that the individual assessment is directly comparable to the group assessment. It is simplistic to suggest that a student would perform at the same level regardless of the assessment activity. Nonetheless, it is

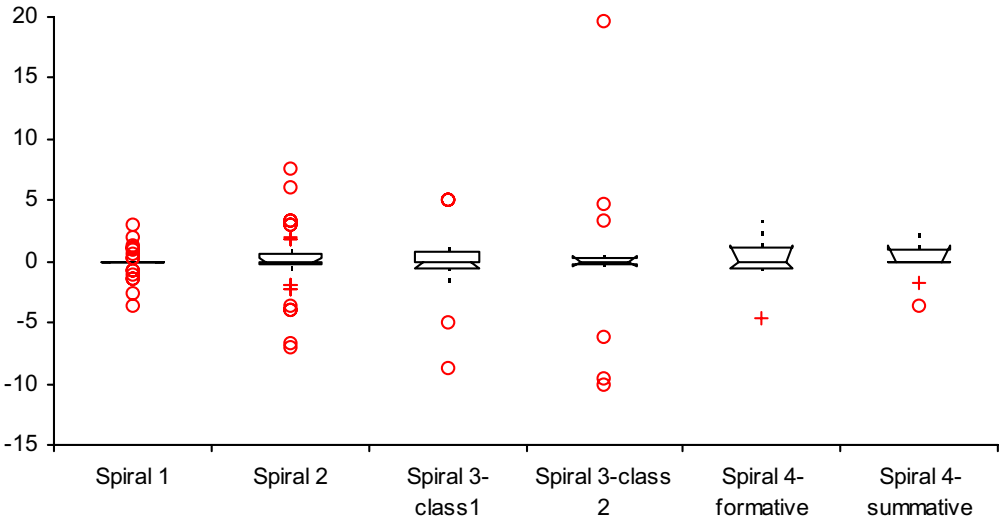

Figure 3: Outliers for each class

Table 11: Correlation Analysis On Final Marks For Method 1

	Asgn 1	Asgn 2	Test 1	Test 2
Final	0.40	0.74	0.90	0.86

Table 12: Correlation analysis from the previous year. Project is marked using Shared Group Mark method

	Exam	Essay	Review	Presntn
Group Project	0.46	0.34	0.47	0.37

Table 13: Correlation analysis on finals marks for current year. Project is marked using Method 3: 60-Sum System.

	Exam	Essay	Review	Reading
Group Project	0.68	0.52	0.79	0.48

informative to consider the correlations between individual and group work for the various methods.

Table 11 shows the relationship between marks on individual assessments and the final course mark for the first class in the study. Assignment 1 was a group assignment that used a Shared Group Mark method (i.e. all students in a group received identical marks for the assignment). Assignment 2 used Method 1, the basic Contribution Mark method. Test 1 and Test 2 were individual assessments. The assignment that used the contribution method (Assignment 2) has a higher correlation with the final mark than does the Assignment that used the Shared Group Mark method. We can thus assume that the marks students received on Assignment 2 more accurately reflected their actual competence with the course material.

An interesting analysis is available for Method 3, the "60-Sum System", where the same assignments were used in a previous instantiation of the course. In this earlier course, the project was marked using the Shared Group Mark method. The correlations between group and individual assessments for the two classes are shown in Tables 12 and 13.

In all cases, the correlation between individual assessment mark and group project mark is higher when the 60-Sum system is used than when the Shared Group Mark system is used. This is compelling evidence for the value of incorporating peer evaluations into marks for group work.

3.4 Marker Agreement

As discussed earlier, marker agreement is an area of concern for peer assessment methodologies. In an attempt to measure the magnitude of the problem, I have summarised the results of the four Methods in terms of top student agreement. That is, for each group I have determined whether all members have identified the same student as the top student in the group. One would expect groups to agree on which student contributed the most, even if there is disagreement about the precise degree of the contribution differential between group members. Table 14 presents the results of the top student agreement analysis for the four peer assessment methods.

The level of top student agreement ranges from 77% to 50% across the four methods. If you exclude the

Table 14: Number of groups in order agreement

	Method 1	Method 2	Method 3 class 1	Method 3 class 2	Method 4 formative	Method 4 summative
Top student agreement	77%	75%	63%	80%	50%	50%
Number of groups giving non-zero scores	38%	58%	63%	60%	75%	50%
Of the non zero groups- the number in agreement	40%	57%	40%	67%	33%	0%

Table 15: Group with similar contribution levels.

Marker	Student one	Student two	Student three	Std dev
Student one	20	20	20	0
Student two	19	20	21	0.82
Student three	21	19	20	0.82
Final	20	19.67	20.33	0.38

Table 16: Group Agreement Under the Standard Deviation Criterion

	Method 1	Method 2	Method 3 class 1	Method 3 class 2	Method 4 formative	Method 4 summative
Groups not in agreement	7%	25%	13%	40%	0%	25%

all-zero groups (i.e. where group members gave each other all equal contribution marks) the level of agreement is considerably less. Perhaps top student agreement is too sensitive a measure. Consider the following group.

Although the group in Table 15 would not be considered in top marker agreement, there is, in reality little difference in the contribution marks of these three students. The standard deviation analysis discussed earlier would not flag this group as suffering from bias. Using that analysis (i.e. where standard deviations below 2.0 constitute sufficient agreement), the allocated marks are much more consistent. This analysis is shown in Table 16.

The fundamental problem with marking agreement is the perception students have of what contribution is and how to quantify it. If students are not judging contribution by the same criteria they are not going to be awarding the same marks for it.

4. CONCLUSION

Assessing group work remains problematic. Threats to validity and reliability need to be examined. Problems with bias, student perception and subtle social pressures need to be resolved. However, I am confident that a method of awarding an individual contribution component to a group mark is a fairer and more valid system than simply giving each student in a group the same mark.

To successfully implement this system you need convincing reasons for making an assessment a group task. Once you have this, there needs to be an element of student buy-in. Students need to be convinced that having a contribution component is fairer than the single mark system.

A strategy for dealing with non-performing students needs to be developed. I gave students the authority to expel non-performing group members. This may be necessary since the lowest possible contribution component in the methods presented here was –20%.

Clearly, a group mark minus 20% is still too high a mark for a student who has not contributed anything.

The system worked very well when it was used as formative assessment first. This gave students experience at using the marking system. It also gave students an indication of how their contribution was perceived by the rest of the group. Students were subsequently more comfortable with the system when it was applied to a summative assessment.

Students struggled with defining contribution. They also found it difficult to turn this concept of contribution into a numerical mark. I would recommend that each student writes a list of what they have contributed, which they then show to all their team members. A table made up of a list of categories that define contribution could be useful.

A strategy for dealing with marker disagreement needs to be decided on and communicated to the students. Students should be asked to reassess their marking and in extreme cases the lecturer needs to be able to exclude a biased student's marks. I found a bias flag based on mark standard deviation to be a useful tool for an objective view of marker agreement.

On a personal note I have enjoyed implementing this study. I feel happier now about using group assessments in my papers. This study has made me more informed, critical and thorough in all my assessment practices.

References

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braum (Eds.), Test validity (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum.

Boud, D., Cohen, R. & Sampson, J. (1999). Peer Learning and Assessment. Assessment & Evaluation in Higher Education, 24(4), pp. 413-426.

Bourner, J., Hughes, M. & Bourner, T. (2001). First-year Undergraduate Experiences of Group Project work. Assessment & Evaluation in Higher Education, 26(1), pp. 19-39.

Conway, R., Kember, D., Sivan, A. & Wu, M. (1993). Peer Assessment of an Individual's Contribution to a Group Project. Assessment & Evaluation in Higher Education, 18(1), pp. 45-54.

Gibbs, G. (1992). Booklet 4: Assessing More Students – The Teaching More Students Project. Oxford, The Polytechnics & Colleges Funding Council.

Lejk, M., & Wyvill, M. (2001). Peer Assessment of Contributions to a Group Project: a comparison of holistic and category-based approaches. Assessment & Evaluation in Higher Education, 26(1), pp. 61-72.

Lejk, M., Wyvill, M., Farrow, S. (1996). A Survey of Methods of Deriving Individual Grades from Group Assessments. Assessment & Evaluation in Higher Education, 21(3), pp. 267-280.

Lejk, M., Wyvill, M., Farrow, S. (1997). Group Learning and Group Assessment on Undergraduate Computing Courses in Higher Education in the UK: results of a survey. Assessment & Evaluation in Higher Education, 22(1), pp. 81-91.

Rafiq, Y. & Fullerton, H. (1996). Peer Assessment of Group Projects in Civil Engineering. Assessment & Evaluation in Higher Education, 21(1), pp. 69-81.