The assessment of group work: lessons from the literature

Professor Graham Gibbs

1 Introduction

Judging from the number of articles on the topic there has been an explosion in the volume of 'group-based' learning, and especially of group project work of various kinds, since a survey of group assessment in higher education the mid 1990s found very modest levels of group learning and assessment (Lejk and Wyvill, 1997). Group project work appears to offer teachers an effective way to engage students, to increase the complexity and challenge of the tasks that students gain experience of working on, to offer students the opportunity for collaborative working, to ameliorate the effects of inadequate libraries through students sharing resources, and to offer the possibility of greatly reduced marking loads, especially where lengthy and complex products emerge from project work. But not everything is positive and, in particular, assessment of groups seems to be a perennial problem. The purpose of this paper is to inform the way group work is assessed. It attempts to offer evidence-based guidance on a range of practical decisions that teachers need to make in designing and carrying out group assessment. It draws on published literature of two main kinds: individual studies of the effects of specific practices on student learning behaviour and performance or marks, and reviews and meta-analyses of collections of such studies that each address a key issue or collection of related issues.

There is a body of literature concerning group work and its assessment that consists of descriptive case studies of practice, with practitioner observations and student feedback on their experience of group work. This literature seldom provides a clear empirical basis for informing teaching decisions but it does provide a range of options for assessment practices designed to address problems with the assessment of groups. There is also a substantial body of social psychology and organisational psychology literature which concerns group behaviour which helps to explain why students behave as they do (e.g. Johnson et al 1981) but which rarely addresses the assessment issues identified here.

There is less empirical evidence available concerning some assessment issues than others, and no evidence of value concerning a few issues. But a number key questions have been addressed well in the literature, some backed by substantial numbers of empirical studies that together paint a coherent picture of the consequences of going about the assessment of groups in particular ways and which therefore provide clear pointers to teachers.

Summary

- Group work has the potential measurably to improve student engagement, performance, marks and retention and usually succeeds in achieving this potential provided that there are associated assessment mechanisms that leverage appropriate student learning behaviour. In the absence of such assessment mechanisms these benefits may well not materialise.

- Allocating a single group mark to all members of a group rarely leads to appropriate student learning behaviour, frequently leads to freeloading, and so the potential learning benefits of group work are likely to be lost, and in addition students may, quite reasonably, perceive their marks as unfair.

- There are a range of mechanisms that allocate differential marks to individual students within a group which are perceived as fair, and which result in more appropriate student learning behaviour and so have the potential to reap the educational benefits of group work. These mechanisms are described in detail in the literature, including detailed variations. Teachers new to assessing group work have plenty of well documented options, including on-line systems, and do not need to invent their own mechanisms.

- Clear allocation of understandable sub-components of the group task to individuals, by the teacher or through students having the project management skills to do this for themselves, improves individual responsibility for these sub-components.

- If groups are too large, individual motivation and effort can be less than if students had studied alone, and associated mechanisms such as peer assessment are likely to be conducted with less seriousness and reliability.

- Streaming groups raises group marks for the groups of better students and reduces the marks of the groups of weaker students, including on subsequent individual assessments. So it is better to create mixed ability groups provided assessment works in such a way that the better students can benefit from their greater contribution.

- If group tasks are short in duration, culturally homogeneous groups perform better. With longer and more complex tasks cultural heterogeneity leads to equivalent or better group performance. Home students' marks are not dragged down by having international students in their group. Group marks in heterogeneous groups are best predicted by the ability of the best student in the group, rather than by the weakest student or by the average ability of students in the group.

- Gender effects on student behaviour and performance with assessed group work have been identified, but most of these effects are inconsistent across different studies, with the exception that female students on average outperform male students at 'group maintenance' functions.

- Groups produce better work than do individuals but whether group work produces inappropriately high average marks depends on context variables, especially the experience of the teacher in setting a group task of appropriately calibrated difficulty and the experience of the marker in calibrating standards that take the greater achievement of groups into account. Higher marks for group work are likely but not inevitable.

- Group work often produces a narrow range of marks for both groups and individual students but there are several mechanisms which can contribute to producing a reasonable spread of marks for individuals, including peer assessment undertaken anonymously.

- Peer assessment of the contribution of others within a group can be reasonably reliable but is better restricted to global judgements of conventional academic tasks rather than multiple judgements against detailed criteria or judgements of professional skills. Peer marks are more reliable (and with a wider spread) when the marker knows the student being marked but the student being marked does not know who is marking them. Students learn from peer assessment.

- There are simple mechanisms which avoid almost all of the potential pitfalls of group based assessment, especially through separating formative assessment of group work from subsequent summative assessment of individuals.

- The extent to which group assessment problems are evident, the extent to which assessment mechanisms have the desired positive effects on student learning behaviour, and the seriousness with which peer marking is undertaken in implementing these mechanisms, are all in part a consequence of the learning milieu the teacher has created

and the extent to which students react responsibly within this milieu. In a healthy learning milieu there may be no group assessment problem that requires attention.

2 Realising the benefits of group based learning

The benefits to students of group based study and group project work have been comprehensively demonstrated both in general (Johnson et al 1991) and in many varied specific contexts. Meta-analyses of large numbers of studies of the implementation of small group learning within individual discipline areas usually show large positive impacts on student performance, marks, attitudes towards learning and persistence or retention. For example Springer et al, (1999) reviewed 383 studies and on the basis of the evidence supported "widespread implementation of small-group learning" (p21) in undergraduate education in the three disciplines studied. Problem based learning (PBL) also invariably uses learning in groups and again meta-analyses of studies comparing PBL with other pedagogies show consistent benefits to student learning processes (such as the extent to which students adopt a deep approach) and outcomes (such as marks) from PBL (see Dochy et al, 2003, for a recent review).

Conclusions from studies of the educational benefits of group based study of various kinds, where students are assessed individually, are so consistent that details of the studies from which these conclusions are drawn need not be discussed here. Rather this review focuses on the assessment challenges that confront teachers once they have decided to use forms of group based study or group project work in order to reap its educational benefits, informed by the available empirical evidence concerning attempts to overcome or by-pass the potential pitfalls of assessment involving groups. These potential pitfalls can be severe and are reported with monotonous regularity in descriptive accounts of assessment of group work. They usually require attention if group work is not to produce worse rather than better educational outcomes for at least some members of the group. For example there can be a sizable decrease in individual effort when working in groups compared with individual work, where there is no marking mechanism to identify the contribution of individuals (Latane et al 1979).

Such problems are not inevitable. For example White et al (2005) report the development and use of three questionnaires to evaluate assessment of group work: Feelings Towards Group Work (Cantwell and Andrews, 2002); Attitudes Towards Peer Evaluation, and Attitudes Towards Group Assessment. Students' attitudes were found to become more positive after experiencing group project work in which all students received an un-moderated group mark. As White et al conclude, this is not the negative picture which is commonly reported (e.g. Barfield, 2003; Gatfield, 1999) and possible explanations for good outcomes without elaborate assessment methods are discussed below.

The problem of allocating fair marks to individuals within groups in ways which leverage appropriate group learning behaviour, and so maximise the potential educational gains of group work, is the focus of a large proportion of the higher education literature on assessing group work. Some of the most common approaches to tackling this issue, such as using peer assessment within groups, have been the subject of reviews of a range of studies.

Most of the literature on the benefits of forms of group based learning other than group project work involves assessment of individuals after the group based learning is complete, and so avoid most of the potential problems associated with assessing the product of group work (such as a group report or design) and then redistributing marks to individuals in some way afterwards.

3 Assessment of individuals within groups

It is rare that assessment of groups is for the purpose of arriving at a mark for the group. It is individuals who graduate and gain qualifications and this review will focus on those forms of assessment that are designed to produce an appropriate mark for an individual who has been involved in group work. The main practical issues and options are summarised here and then the subsequent review of the literature addresses issues common to many of these options. This summary draws on Lejk et al (1996) which in turn draws on Gibbs (1995a). The main potential problems that assessment of group work is likely to face are discussed below.

The construction of student groups

The way student groups are constructed has a marked impact on the quality of the end product that is assessed. High ability students gain higher grades when in 'streamed' groups of similarly high ability students than when they are in mixed ability groups. The reverse is the case for low ability students: they benefit from working in mixed ability groups and suffer from being in streamed low ability groups. Furthermore, low ability students will suffer in subsequent examination when they have been working in streamed low ability groups, compared with having worked in mixed ability groups (Lejk et al 1999). Allowing students to form their own groups is likely to have a similar impact as streaming – the high ability students will tend to form groups with each other and the low ability students will be left with other low ability students to work with. The fairest option is therefore to construct mixed ability groups but to make sure that high ability students who contribute more have their greater contribution recognised in their individual mark so that they are not unfairly penalised by being obliged to work with lower ability students.

Culturally homogeneous groups outperform heterogeneous groups if the group task is of short duration, presumably as they find it easier to get down to work quickly when they are familiar with the culture. As the duration of the group task increases, the performance gap between homogeneous and heterogeneous groups decreases and by four months the gap disappears (Watson et al 1993), presumably because students eventually develop a way of working effectively together despite cultural differences. Culturally heterogeneous groups can offer many potential benefits over homogeneous groups, including increased numbers of ideas, greater creativity and flexibility (Kirchmeyer, 1993) provided students have the skills and time to manage the group process (Watson et al, 1993).

As group size increases, individual motivation decreases (Kerr & Bruun, 1983). Students also lack the group management and facilitation skills to cope well with large groups or to cope with the greater likelihood of 'social loafing' in groups where students can hide more easily. Four to six students seems to be ideal, with groups of eight or more creating significant problems. Slightly larger groups may help where creativity or sheer workload is an issue, but slows down progress due to the greater difficulties of reaching decisions, allocating tasks, monitoring progress, and pulling components to together. Group size also affects the seriousness with which group members undertake peer assessment in order to moderate or distribute grades and, as mentioned below, larger groups are as a consequence less reliable in their peer assessment than small groups.

Practice in group work

Students usually have a number of opportunities for practice at producing the products of individual study that are assessed, such as through writing essays, before the first time they are marked and have the marks contribute in any significant way to their qualifications. Students may have far less experience with group work, and the scale of the product may be sufficiently large (for example contributing 100% of the marks for a course) that group work can become a high risk activity for students. A study of students' learning responses to assessment regimes at three universities found ample evidence of students finding assessment techniques they had not previously encountered (such as group presentations) anxiety provoking, disorienting and, when they contributed substantially to their module marks, unfair (Gibbs and Dunbar-Goddet, 2007). An issue of fairness, and acceptability to students, is therefore how students come to learn how to operate

effectively in groups in such a way that their marks are a fair indication of what they have learnt rather than only a reflection of their level of familiarity with group work. Gibbs (1995b) provides practical guidance for groups encountering their first short group project and Gibbs (1995c) provides a comprehensive set of exercises groups can run for themselves to support them in tackling larger, more complex and longer duration projects, and to develop their group skills.

Differential contributions

As a precursor to the sections that follow, this section introduces the issue of differential contributions within a group. Individual students within a group will almost inevitably have contributed different things to the writing of the report or other product that is to be assessed. This is not simply a matter of different levels of effort. Different components of the group work undertaken by different members, such as reviewing the literature, collecting data and creating a powerpoint presentation for the final report, may not be of equivalent academic value, may be conducted to varying standards, and may require varying levels of assistance from other members of the group. Even if all group members have put in similar numbers of hours, an unlikely scenario, this is not normally a basis for allocating marks. It is the level of challenge and standard of the work of individuals which normally determines their marks rather than their effort. The issue to be addressed by the teacher is how to distinguish marks for individuals within a group who have made different kinds of contributions. The issue to be addressed by the students is how to perform group maintenance functions in such a way that the range of differential contributions is managed within reasonable and fair limits.

Freeloading and appropriate group behaviour

As mentioned above, there is a sizable decrease in individual effort when working in groups compared with individual work, where there is no mechanism to identify the contribution of individuals (Latane et al, 1979). The level of freeloading increases when tasks are not identified clearly (George, 1992). It is clear from many ‘case study’ accounts of group work assessment (e.g. Bacon and Stewart, 1999) that allocating all individuals within each group the same un-moderated group mark results in a variety of problems, and should be avoided. If an individual student knows that the mark they get will be largely the consequence of the levels of effort and achievement of other students, there is little incentive to put in much or any work themselves and little disincentive to allocate available study hours to other courses with individual assessment where personal effort is more likely to produce personal benefits in terms of improving marks. Pitt (2000) has applied game theory to explain such student behaviour in assessed group work, arguing that the best strategy for individual students may not involve promoting teamwork or even co-operation. ‘Freeloading’ is a potential problem with all group work and a major focus of student discontent, both because freeloaders gain marks they do not deserve, and because if overall effort is lower then the overall group mark is likely to be lower and will disadvantage those who did pull their weight. When the group task is poorly defined, feeloading increases (George, 1992). Houldsworth and Matthews (2000) also describe a ‘sucker effect’ in which the most hardworking student gradually reduces their effort in order to avoid being taken advantage of by the freeloaders. Freeloading also causes organisational problems and delays if allocated tasks are not completed. Assessment should be designed in such a way that it results in appropriate student behaviour because individuals will see that their effort will be rewarded and their lack of effort punished. Much of the literature on group assessment concentrates on how to achieve this. What follows is a summary of the available options and, where available, evidence of their impact.

Options for reducing problems associated with freeloading

- **Limiting the emphasis on group marks**

  Here the tactic is to allocate a significant proportion of marks for an individual on a course to assignments or tests other than the group project, so that even if there are problems with the allocation of marks to individuals for the group project component, this will not matter so much.

- **Assessing the outcomes of group work with individual assignments or examinations**

Here the tactic is to undertake the group work, including the production of a group report, presentation, or other product, with formative feedback but no allocation of marks. There is then a separate assessment, such as an individual report based on the preceding group project work (cf Hindle, 1993), or an unseen examination with questions about the preceding group project work, with marks being allocated entirely to individuals. The crucial issue with such a tactic is how to design the group work and individual assessment in such a way that being totally involved in the group work is the best, or even the only, way for a student to prepare for the subsequent individual assessment. Any student who did not pull their weight in the group should, quite appropriately, not score well on the individual assessment. Characteristically the exams only include questions that relate directly to the preceding group work, for example asking about methodological issues encountered or changing a key variable in the problem scenario the groups tackled and asking how this change would have affected the group's conclusions about how to tackle the scenario. This complete separation of the learning benefits of group work from the summative rigour of individual assessment is the simplest solution to most group assessment problems, but relies on the form of questions in the project exam and on students' perceptions of the demands of these questions. If students perceive that they would be better off bailing out of the group work in order to revise for the exam, then this form of assessment will undermine group work.

- **Dividing up the group task between individuals and allocating some or all marks to component tasks**

This approach is possible where tasks can readily be divided because they have discrete components (cf Lejk and Wyvill, 2001a) or operate in an identifiable sequence. For some tasks, such as writing an extended essay, dividing the task up may be impracticable or unhelpful except for components such as sections of a reading list which are then not easy to assess separately. Where it is possible to identify components of the group task, each individual in a group can be allocated responsibility for that component and receives 50% of their marks for the quality of their own component and 50% for the entire group product, so that they pay attention to the whole task as well as to their own component of it (cf. Earl 1986). Group dynamics may work well with such an arrangement where the group will tend to ensure that all components are completed, provided responsibility for these components is clearly identified. Gibbs (1995b, 1995c) provides guidance and exercises for students to help them to make these responsibilities explicit within their group without the tutor having to divide up the task for them.

- **Teachers moderating the group mark for each individual on the basis of special knowledge about the individual**

In some circumstances, such as during supervised fieldwork or laboratory work, it will be possible for the teacher to gain an adequate impression of the relative contribution of individuals to the group work simply by informal observation. Such observation can be formalised through rating individuals on a common set of scales concerning contribution to different components of the work (such as collecting data or writing up), or different process variables (such as 'effort' or 'ideas'). In most circumstances, however, such observation would provide an incomplete basis for judging at least some individual members of most groups. An alternative approach that is fairer to students would be to collect additional evidence about each individual. Alternative ways to do this include:

  - requiring all group members to keep a project log or other portfolio that reveals individual engagement and effort
  - conducting a brief viva for each student, who would enter the viva with the group mark and leave it with the group mark plus or minus perhaps 20%, on the basis of the adequacy of their answers to probing questions about their involvement in aspects of the group project work.

While this method is mentioned in several practical guides (e.g. Gibbs, 1995a). there appears to be no published evidence about its consequences for student behaviour or performance.

- **Students moderating each other's group mark on the basis of their inside knowledge about that individual**

This has been referred to in the literature as the "Knickrehm method" (Maranto & Gresham, 1998). A crucial issue here is that it is usually the teacher who makes an expert academic judgement about the quality of the product (such as a project report) while the students peer review the quality or quantity of various types of contribution to that product. This is not the same as peer assessment of products and consequently some of the literature on the reliability of peer assessment may not apply here as that literature usually concerns students marking other students' products. A mechanism for students to moderate each others' marks starting with a group mark allocated by a tutor was first reported in 1990 (Goldfinch and Raeside (1990). Conway et al (1993) added ratings as a way of mechanising peer review. Goldfinch (1994) then simplified the rating process. A range of mechanisms for students to moderate each other's group marks were reviewed by Lejk et al (1996) and there have been many technical developments since then, including statistical transformations to overcome anomalies (Cheng and Warren, 2000; Li, 2001; Sharp, 2006) and a web-based system that ensures anonymity and reduces administration (Freeman and McKenzie, 2002). Most writers about this method claim that a wider spread of marks is achieved than by allocating individuals the group mark, that greater fairness is achieved and that such assessment is perceived by students to be fair, and consequently that there is more collaborative and responsible individual behaviour within groups. However, there is more evidence about effects on marks and on students' perceptions of the acceptability of group assessment and perceived fairness of their individual marks than there is about the impact on student behaviour.

Freeman and McKenzie (2002) have reported that use of repeated on-line self and peer assessment of group process variables as formative feedback during group work led to improvements both in group processes and in the quality of project outcomes. Falchikov (1995) has also emphasised the benefits of formative-only peer assessment for the purpose of providing feedback.

It would be difficult to recommend any of these detailed mechanisms over another without reference to details of the context but there is clearly no need for a teacher to have to invent their own mechanism.

A variation on the method of group members using ratings of some kind to moderate each other's mark from a group mark involves the teacher allocating a group mark and multiplying it by the number of students in the group and allocating this pool of marks to the group to distribute amongst its members as they see fit (Habeshaw et al, 1993). Mechanisms such as this need to be explained at the outset of the group work so that students understand the likely consequences of appropriate or inappropriate group behaviour, and it may be prudent to set limits to the extent to which students' marks can vary within a group, such as 20% between the top and bottom student. The way negotiations of allocations of marks takes place is likely to be open to both gender and cultural biases (see below) and it may be prudent to use a mechanism that allows for open discussion followed by each student proposing a fair distribution in a confidential way, with the teacher collating these 'secret' proposals that the students do not see, and the teacher then allocating the final marks.

In an even simpler method involving group members in assessment, students are allowed to make sanctions against others in their group who behave inappropriately – such as not turning up to meetings and not completing their tasks to schedule – provided there is agreement. The sanctions can consist of a proposed loss for that individual of a specified number of marks from the teacher-allocated group product mark. Sanctions may be proposed while the group is still working on the group project and may be rescinded if the individual concerned changes their behaviour appropriately and delivers their component of the group work satisfactorily. If the purpose of such assessment mechanisms is to leverage appropriate student behaviour then

such interim peer assessment, made while the group work is progressing, may have advantages, though there would need to be teacher agreement about any sanctions actually implemented. Group exercises that provide formative feedback on group functioning and group behaviour, such as those contained in Gibbs (1995c), can have the same positive effect without needing to involve either the threat of sanctions over marks or the risks of student involvement in allocating marks.

4 The use of peer assessment as part of assessment of groups

A key issue concerning the active involvement of group members in group assessment is whether peer assessment can be trusted. This section addresses a number of the key questions about the advisability of using peer assessment as part of group work assessment. Where no specific reference is cited in this section, the conclusions are drawn from a meta-analysis of 48 studies of peer assessment undertaken by Falchikov and Goldfinch (2000).

- Student peer marks are not especially reliable (Kegel-Flom (1975), but teachers' marks are not especially reliable either, and students are not much more unreliable than are teachers (Boud, 1986). Student peer assessment marks agree reasonably well with teachers' marks, with an average correlation of about +0.7, and this might be considered adequate. There are many reports of adequate agreement of peer marks and tutor marks when judging group work in different subject areas (e.g. Hughes and Large, 1993; Falchikov 1988, 1991). There is no consistent evidence that peer assessment is more or less reliable in different subject areas.

- Peer assessment which requires marking of several different dimensions (such as communication, argument, use of literature) are less reliable than peer assessment involving a single overall judgement underpinned by a set of criteria. It is also the case that students have more positive attitudes towards making global judgements about others' contributions to group work than to giving compartmentalised ratings of multiple components (Lejk & Wyvill, 2002).

- Students are more reliable when assessing traditional academic products (such as essays) than other students' performance in the context of professional practice. Falchikov and Goldfinch (2000) found average reliability coefficients from a total of 29 studies to be an acceptable r=0.75 for peer assessment of academic products but only r=0.54 for peer assessment of professional practice, which would be considered unacceptably unreliable if academics were undertaking the marking. This is easy to understand in terms of students' experience of at least some kind of membership of an academic 'community of practice' that has an implicit set of standards and criteria about common practices within that community, such as academic writing, and their corresponding lack of equivalent experience of professional practices or of the communities in which these practices take place. Academics may also lack membership of professional communities of practice and may have difficulties in communicating standards clearly to students for the marking of professional practice.

- There is equivocal evidence about whether having a number of students contribute to peer assessment improves reliability. Some studies have reported an increase in reliability with multiple peer assessment (e.g. Magin 1993). However, meta analysis of all studies on this question (Falchikov and Goldfinch ibid) has found no consistent advantage from involving a number of students in peer assessment of a single assignment compared with a single student peer assessment of the same assignment, and also found that involving a large number of peer assessors (20+) reduces reliability. Overall the evidence of any advantage gained through having more than one peer assessment in order to improve reliability is insufficient to justify the effort. It has been suggested by a number of authors that reciprocal peer assessment (in which students in pairs assess each other) would be untrustworthy, especially between friends, and would lead to higher marks than are justified. However, Magin (2001) found that such reciprocal peer assessment between those who have close

social interactions had negligible effects on marks. The key issue seems to be whether students understand and are committed to the educational rationale of group assessment and so adopt a responsible attitude to it (Boud,1986). Yan & Kember (2003) report dramatically different student learning behaviour in groups on different courses depending on the extent to which the teacher had put effort into creating a learning environment conducive to responsible collaborative learning. There may be no fool-proof technical solutions to problems such as the reliability of peer assessment in groups. Rather there are subtle cultural and social variables in the learning context, which teachers can influence, that affect the seriousness with which students engage with whatever techniques are adopted.

- Student familiarity with and 'ownership' of the criteria that are used improves the reliability of their judgements.

- Peer assessment is no more reliable on advanced levels courses than it is on introductory levels.

- A number of studies have reported gender biases in marking in the sense that males favour males and females favour females. However, first, there appears to be no consistent overall gender bias (Falchikov & Magin, 1997) and second, where such a bias has been reported, male researchers' findings tend to be biased in favour of the reliability of male markers while female researchers' findings tend to be biased in favour of the reliability of female markers (for a meta-analysis of studies of 'social influence' on judgements of others see Eagly & Carli 1981). It does seem that males and females value different criteria to different extents (Bean and Kidder, 1982) and so some effort to ensure that students are using the same criteria in a balanced way might be wise. There is clear evidence of gender differences in sensitivity to group maintenance variables and also gender differences in the value placed on such variables, so gender bias is more likely when students peer assess process than when they are assessing content or product.

- There is evidence of various social and demographic biases in some settings though effects are complex and sometimes inconsistent. Knowledge of the student is, in general, the greatest source of potential bias when academics undertake marking (Dennis et al, 1996) which is why examining is nowadays undertaken largely anonymously. However, it is not clear how peer assessment within groups could be conducted without the student undertaking peer assessment knowing who they were assessing.

- Students often believe there will be biases and can be anxious about fairness (Sherrard and Raafat, 1994) though Keaten and Richardson (1993) found that even those students who did not favour group projects as learning activities believed that peer assessment was a valid means of determining student achievement.

- Details of mechanisms for allocating marks to individuals through some kind of peer assessment, may have quite marked effects on student learning behaviour but may make almost no difference to the marks. Options, such as those outlined above, and their consequences for marks, are summarised by Lejk et al (1996).

- Marking accuracy, or reliability in terms of students producing the same mark a tutor would produce, is improved if those assessing know the individual being assessed (Kane and Lawler, 1978). However, as pointed out above, knowledge of the student increases biases, so the students and teachers here may here be producing similar grades by being biased in the same way by their knowledge of the individual being assessed.

- Secret peer assessment where an individual does not know which other student or students gave which marks to their efforts, but only the outcome of averaging others' peer assessments, produces a greater spread of marks and more distinction between individuals (Lejk & Wyvill, 2001). If peer assessment of individual contributions to group work is done in

the open, students will be more reluctant to mark others down or up based on their actual contribution. However, public feedback from peers has more effective on subsequent behaviour than from a supervisor or teacher (DeNisi et al, 1982). This illustrates well the two sometimes contrasting consequences of different group marking procedures: on marks and their perceived fairness, and on individual behaviour within the group and the quality of learning that follows from such behaviour.

5 Further issues that arise when involving students in self and peer assessment as a component of assessing groups

What kinds of ratings should be used?

As with literature on students' ratings of teachers, there is clear evidence that rating scales should be 'behaviourally anchored' in the sense that they refer clearly to behaviours that can be seen to be evident, or not, in colleagues' contributions to the group work (e.g. attended all group meetings, most meetings etc), rather than scales that cannot easily be associated with any particular behaviour that could be unambiguously seen (e.g. 'helpfulness to group') (Levi & Cadiz, 1998). Gueldenzoph (2002) provides examples of appropriate scales.

Does peer assessment promote learning?

There is a range of evidence that the act of peer assessment is, itself, a learning experience and improves student learning, independently of the fairness or usefulness of the grades that are produced (Boud, 1988; Falchikov, 1986; Magin and Churches, 1989).

Can students moderate their own group marks on the basis of self-assessment?

- Published studies offer a very mixed answer to whether students are over-generous in the marks they give themselves. Seventeen studies report students grading themselves higher than their teachers do, while 12 studies report the opposite (see Boud and Falchikov, 1989 for a review). Variations in details of the assessment procedure seem to be responsible for some of the different effects, but there are also effects due to student differences.

- Higher ability students tend to be tough on themselves and under-estimate the worth of their own work compared with the judgement of a teacher while lower ability students tend to award higher marks to themselves than their teachers believe are justified. More experienced students are either more accurate or under-estimate while inexperienced students tend to be unreliable or over-estimate (for a summary see Boud & Falchikov ibid). Interestingly the same phenomenon is evident with teachers: those teachers rated most highly by their students rate themselves as a teacher less highly than their students do, while those rated most poorly by their students rate themselves more highly than their students do. It seems that there is a general phenomenon that as one learns more and becomes more sophisticated, one applies tougher standards to oneself during self-assessment.

- There is some evidence that female students agree with their teachers more than do male students, but studies of gender effects on self-assessment are mainly inconclusive.

- Students in science self-assess more accurately (Falchikov and Boud, 1989) presumably because what they are assessing is easier to see as right or wrong and there is less subjective judgement involved.

- Students often over-estimate their marks, compared with a teacher's judgement, when their marks count (compared with when they do not), though there are studies which found exceptions to this trend (Boud and Falchikov ibid).

6 Averages and spreads of marks from group work

There is a dearth of conclusive evidence on whether group work tends to produce higher average marks and a narrower spread of marks than would be the case if the same individuals were assessed separately. Knight (2004) found average marks for group work to be 4% higher than marks for individual exercises on the same course. De Vita (2002) reported most students gaining higher marks than on previous individually assessed courses, with the weakest students experiencing the largest increase in their marks. Gray (2009) reported weaker students to gain higher marks on modules with more assessed group work while stronger students gained lower marks than they had done on other modules. However this kind of evidence on whether students gain better or worse marks than they otherwise might is rare, in part due to the difficulty of studying this phenomenon in a conclusive way. For example there is no way of judging whether in the Knight study the group and individual exercises were of equivalent difficulty.

Logically, however, if the same task is set for groups as for individuals, then group work should produce higher average marks, for a variety of reasons:

- groups have more human resources and are more productive than individuals and so should be able to read more, cite more sources, collect more data, and so on. The total 'time on task' is greater than for individual assignments.
- groups can cope with greater complexity and be more creative than can individuals and so should be able to tackle more challenging tasks, produce more solutions to problems, and produce more elaborated analyses, taking into account more issues.
- students learn more in groups than they do on their own (as reported above).

Indeed, groups would be expected to outperform individuals to such an extent that they are usually set much larger and more difficult tasks than you could reasonably set an individual, in order to provide a sufficient challenge for the group and in order to produce an 'acceptable' range of marks.

Whether the products of group work are then allocated higher or lower marks than the products of individual work will depend on the extent to which the marker (or the marking scheme) takes into account the advantages groups have and applies tougher standards in producing a mark, or takes into account the greater level of difficulty and applies more lenient standards, or balances these two effects out and marks to the same standards as for individuals. One might expect inexperienced markers, or teachers who have not used group work before, to get this adjustment wrong and produce higher average marks for groups, but for this effect to disappear over time and through marking experience and through gradually adjusting the scale and difficulty of the task. Whether or not any particular study of this phenomenon shows higher marks for groups needs to be interpreted in the light of the explicit or implicit standards being applied and the relative scale and difficulty of the tasks being set and whether they were different from the standards and levels normally applied to individual work. Variables such as 'difficulty' and 'standards' are seldom amenable to straightforward empirical study and are rarely mentioned in reports of marks for group work.

An additional complexity in addressing this question about different average marks is group size. An increased number of students increase the human capacity to cope with the scale of the task and ability to cope with complexity – but only up to a point. Large groups are very difficult to manage and get the best out of, and inexperienced students are likely to find large groups as much a drawback as a help. Again a marker would have difficulty calibrating standards to cope with different group sizes and students' varied ability to profit from group work given the group size.

The issue of spread of marks is easier to address. The range of ability and contribution of students within a group will often be similar to the range in another group – indeed groups are often set up deliberately so as to spread ability and commitment reasonably equitably between groups, and so they are relatively similar to each other. The inevitable statistical consequence of this is a narrow range of performance, and marks, between groups. It is possible to spread marks for individuals within groups out again to produce a more acceptable range (Cheng & Warren, 2000) by using one

of the techniques for distinguishing between students within a group (see above). Without such correction, and without allowing able students to form their own group (and the less able to be in a different group), the range of marks will inevitably be narrow. Also, as cited above, students produce a wider spread of peer assessed marks when peer marking is secret or anonymous.

If marking criteria involve weakly defined generic outcomes (such as 'group skills') or professional outcomes (such as competence in a complex performance e.g. 'interviewing a client') rather than familiar academic forms of performance (such as a problem sheet or essay) the range of marks is again likely to be narrow, due to the difficulty of distinguishing levels of performance. It is possible to get 0% or 100% for getting something completely wrong or completely right, but marks for 'group work skills' often have a high average and a narrow spread, with markers never allocating 0% or 100% or even anything near these extremes, especially at the bottom end of the scale. This marking phenomenon is of course evident in the range of marks allocated for individual performances, such as a seminar presentation, but such generic skills are more commonly emphasised in assessing group work and dominate peer assessment of others' contributions to the products of group work (cf Hindle, 1993).

7 Conclusions

Descriptive accounts of introducing group assessment, in which all students in a group are allocated the group mark, and which report that this brings with it many problems, are still popping up in the literature, two decades after they were first well documented. There is no need to encounter these problems. A range of practical alternative mechanisms for allocating marks to individuals within groups have been tried out, evaluated, and documented. Many studies have provided evidence concerning the consequences of these mechanisms, and variations on them, for student grades, averages and grade distributions, student behaviour and student attitudes towards the acceptability of group work and its assessment. Issues of the impact of group size and group construction on student performance and on fairness have also been studied and documented. Broader issues concerning the advisability and consequences of involving students in some form of peer assessment have been extensively researched. Much is already known on the realities of supposed biases and potential causes of unfairness. This review has attempted to bring this work together as guidance in the hope that no more naïve accounts are published announcing that group assessment is problematic.

Often buried amongst all this accumulated wisdom, technical expertise and empirical evidence, however, is the issue of local milieu. If a significant number of students want to undermine an assessment system, and behave strategically in ways that are to their advantage, in terms of marks, at the cost of their own learning and at the expense of others, even extensive and time consuming mechanisms to make marking reliable and fair may have limited success. If students understand why group work is being used, understand the assessment system, are collaborative and ethical in their behaviour and posses sophisticated group work skills, then only minimal assessment mechanisms may be necessary as safeguards. In the end it is the creation of a healthy learning milieu that can contribute most to solving group work assessment problems. If it is beyond teachers to create such a culture, then they at least have sound mechanisms to fall back on.

References

Bacon, D. R. & Stewart, K. A. (1999) Learning from the best and worst student team experiences: how a teacher can make the difference , *Journal of Management Education*, 23(5), pp. 467–489.

Barfield, R.L. (2003). Students' perceptions of and satisfaction with group grades and the group experience in the
college classroom. *Assessment & Evaluation in Higher Education, 28*, 355 – 369.

Bean, G. & Kidder, L. (1982) Helping and achieving: compatible or competing goals for men and women in medical school? *Social Sciences and Medicine*, 16, pp. 1377-1381.

Boud, D.J. (1986*) Implementing Student Self-Assessment*. Green Guide No. 5. Sydney: Higher Education Research and Development Society of Australasia.

Boud, D. (Ed.) (1988) *Developing Student Autonomy in Learning* (2nd Edn.) London: Kogan Page.

Boud, D. & Falchikov, N. (1989) Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *Higher Education*, 18, 5, pp. 529-549.

Cantwell R, H. & Andrews, B. (2002). Cognitive and Psychological Factors Underlying Secondary Students' feelings Towards Group Work. *Educational Psychology.* 22, pp. 75-91.

Cheng, W. & Warren, M. (2000) Making a difference: using peers to assess individual students' contributions to a group project. *Teaching in Higher Education*, 5, 2, pp. 243-255.

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*, 19, pp. 45–56.

Dennis, I., Newstead, S.E. & Wright, D.E. (1996) A new approach to exploring biases in educational assessment. *British Journal of Psychology*. 87,4, pp. 515-534.

DeNisi, A.S., Randolph, W.A. & Blencoe, A.G. (1982) Level and source of feedback as determinants of feedback effectiveness. *Proceedings of the Academy of Management Annual Meeting*, pp.175-179.

De Vita, G. (2002) Does assessed multi-cultural group work really pull UK students' average down? *Assessment and Evaluation in Higher Education*, 27, 2, pp. 153-161

Dochy, F., Segers, M., Van den Bossche, P. & Giibels, D. (2003) Effects of problem-based learning: a meta-analysis. *Learning and Instruction*, 13, 5, pp. 533-568.

Eagly, A.H. & Carli, L.L. (1981) Sex of researchers and sex-typed communications as determinants of sex differences in influencability: a meta-analysis of social influence studies. *Psychological Bulletin*, 90,1, pp. 1-20.

Earl, S. E. (1986) Staff and peer assessment - measuring an individual's contribution to group performance, *Assessment & Evaluation in Higher Education*, 11, pp. 60–69.

Falchikov, N. (1986) Product comparisons and process benefits of collaborative peer group and self assessments. Assessment and Evaluation in Higher Education. 11, pp. 146-166.

Falchikov, N. (1988) Self and peer assessment of a group project designed to promote the skills of capability, *Programmed Learning and Educational Technology*, 25, pp. 327–339.

Falchikov, N. (1991) Group process analysis: self and peer assessment of working together in a group, in: S. Brown & P. Dove (Eds) *Self and peer assessment.* Birmingham, SCED Publications.

Falchikov, N. (1995) Peer feedback marking: developing peer assessment. *Innovations in Education & Training International*, 32, 2, pp.175-187.

Falchikov, N. (1991) Group process analysis: self and peer assessment of working together in a group, in: S. Brown & P. Dove (Eds) *Self and peer assessment.* Birmingham, SCED Publications.

Falchokov, N. & Boud, D. (1989) Student self-assessment in higher education: a meta-analysis. *Review of Educational Research*, 59, pp. 395-430.

Falchikov, N. & Goldfinch, J. (2000) Student peer-assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70,3, pp. 287-322.

Falchikov, N. & Magin, D. (1997) detecting gender bias in peer marking of students' group process work. *Assessment and Evaluation in Higher Education*. 22,4, pp. 393-404.

Freeman, M. (1995) Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education*, 20, pp. 289–299.

Freeman, M. & McKenzie, J. (2002) SPARK, a confidential web-based template for self and peer assessment of student teamwork: benefits of evaluating across different subjects. *British Journal of Educational Technology*, 33, 5, pp. 551-569.

Gatfield, T (1999). Examining student satisfaction with group projects and peer assessment. *Assessment and Evaluation in Higher Education, 2*, 365 – 377.

Gibbs, G. (1992) *Assessing student centred courses*. Bristol, Technical and Educational Services

Gibbs, G. (1995a) *Learning in Groups: Tutor Guide*. Oxford: Oxford Centre for Staff Development.

Gibbs, G. (1995b) *Learning in Groups: Student Guide*. Oxford: Oxford Centre for Staff Development.

Gibbs, G. (1995c) *Learning in Groups: Student Manual*. Oxford: Oxford Centre for Staff Development.

Gibbs, G. & Dunbar-Goddet, H. (2007) The effects of programme assessment environments on student learning. York: Higher Education Academy http://www.heacademy.ac.uk/assets/York/documents/ourwork/research/gibbs_0506.pdf (accessed July 2009).

George, J.M. (1992) Extrinsic and intrinsic origins of perceived social loafing in organisations. *Academy of Management Journal*, 35, 7, pp. 191-202.

Goldfinch, J. (1994) Further developments in peer assessment of group projects, *Assessment & Evaluation in Higher Education*, 19, pp.29–35.

Goldfinch, J. & Raeside, R. (1990) Development of peer assessment technique for obtaining individual marks on a group project, *Assessment & Evaluation in Higher Education*, 15, pp. 210–225.

Gray, P.J. (2009) *Interpreting the numbers: what can systems data tell us about variations in marks?* Unpublished presentation. Oxford: Assessment Standards Knowledge Exchange, Oxford Brookes University.

Green, P.J. (2009) *Interpreting the* numbers*: what can systems data tell us about variations in marks?* Unpublished report. Oxford: Oxford Brookes University, Assessment Standards Knowledge Exchange.

Gueldenzoph, L.E. & Wilson, K.S. (1997) Teaching collaboration in a technology-based environment. *Instructional Strategies: an Applied Science Series*. 13, 4, pp. 1-6.

Habeshaw, S., Gibbs, G. & Habeshaw, T. (1993) *53 interesting ways to assess your students* (3rd Edn). Bristol: Technical and Educational Services.

Houldsworth, C. & Matthews, P.B. (2000) Group composition, performance and educational attainment. *Education and Traning*, 42, 1, pp. 40-53.

Hughes, I. E. & Large, B. J. (1993) Staff and peer-group assessment of oral communication skills, *Studies in Higher Education*, 18, 379–385.

Hindle, B. P. (1993) The 'Project': putting student-controlled, small-group work and transferable skills at the core of a geography course, *Journal of Geography in Higher Education*, 17, 11–20.

Johnson, D.W., Johnson, R.T. & Smith, K. (1991) *Co-operative learning: Increasing college faculty instructional productivity*. ASHE-ERIC Higher Education report No.4 Washington DC: The George Washington University School of Education and Human development.

Johnson, D.W., Maruyama, G., Johnson, R., Nelson, D & Skon, L. (1981) Effects of cooperative, competitive and individual goal structure on achievement: a meta-analysis. *Psychological Bulletin*, 89, pp. 47-62.

Kane, J.S. & Lawler, E.E. (1978) Methods of peer assessment, *Psychological Bulletin*, 85, 3, pp. 555-586.

Keaten, J. & Richardson, M.E. (1993) *A field investigation of peer assessment as part of the student grading process.* Paper presented to the Annual Meeting of the Western State Communication Association. Kansas City, MO, August 11-14.

Kegel-Flom, P. (1975) Predicting supervisor, peer and self-ratings of intern performance. *Journal of Medical Education,* 50, pp. 812-815.

Kerr, N. L., Bruun, S. E. (1983) Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*. 44,1, pp. 78-94.

Kirchmeyer, C. (1993) Multicultural task groups, *Small Group Research,* 24,1, pp127-148.

Knight, J. (2004) Comparison of Student Perception and Performance in Individual and Group Assessments in Practical Classes. *Journal of Geography in Higher Education*, 28,1, pp. 63–81.

Latane, B., Williams, K. & Harkins, S. (1979) Many hands make light the work: the causes and consequences of social loafing, *Journal of Personality and Psychology*, 37(6), pp. 822–832.

Lejk, M. & Wyvill, M. (1997). Group learning and group assessment on undergraduate computing courses in higher education in the UK: results of a survey. *Assessment and Evaluation in Higher Education*, 22, 1, pp. 81-91.

Lejk, M. & Wyvill, M. (2001a) Peer assessment of contributions to a group project: a comparison of holistic and category-based approaches. *Assessment and Evaluation in Higher Education*, 26, 1, pp. 61-72.

Lejk, M. & Wyvill, M. (2001b) The effect of inclusion of self-assessment with peer assessment of contributions to a group project: a quantitative study of secret and agreed assessments. *Assessment and Evaluation in Higher Education*, 26, 6, pp 551-561.

Lejk, M. & Wyvill, M. (2002) Peer assessment of contributions to a group project: student attitudes to holistic and category-based approaches. *Assessment and Evaluation in Higher Education*, 27, 6, pp 569-577.

Lejk, M., Wyvill, M. & Farrow, S. (1996) A survey of methods for deriving individual grades from group assessments. *Assessment and Evaluation in Higher Education*, 21, 3, pp 267-280.

Lejk, M., Wyvill, M. & Farrow, S. (1999) Group assessment in Systems Analysis and Design: a comparison of the performance of streamed and mixed-ability groups. *Assessment and Evaluation in Higher Education*, 24, 1, pp 5-14.

Levi, D. & Cadiz, D. (1998) *Evaluating team work on student projects: the use of behaviourally anchored scales to evaluate student performance*. ERIC Document Reproduction Service, ED424250.

Li, L. K. Y. (2001) Some Refinements on Peer Assessment of Group Projects. *Assessment and Evaluation in Higher Education*, 26,1, pp. 5-18.

Magin, D. (1993) Should student peer ratings be used as part of summative assessment? *Higher Education Research and Development*, 16, pp. 537-542.

Magin, D. (2001) Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education*, 26,1, pp 53 – 63.

Magin, D. & Churches, A.E. (1989) What do studens learn from self and peer assessment? In *Designing for Learning in Industry and Education.* Canberra: Australian Society for Educational Technology, pp224-233.

Maranto, R. & Gresham, A. (1998). Using “world series shares” to fight free riding in group projects. *Political Science and Politics, 31*, 789 – 791.

Pitt, M.J. (2000) The application of game theory to group project assessment. *Teaching in Higher Education*, 5,2,pp.233-241.

Li L.K.Y. (2001) Some refinements on peer assessment of group projects. *Assessment and Evaluation In Higher Education*, 26, 1, pp. 5-18.

Sharp, S. (2006) Deriving individual student marks from a tutor’s assessment of group work. *Assessment and Evaluation in Higher Education*, 31, 3, pp. 329-3433.

Sherrard, W.R. & Raafat, F. (1994) An empirical study of peer bias in evaluations: students rating students. *Journal of Education for Business*, 70, 1, pp. 43-48.

Springer, L., Stanne, M.E. & Donovan, S.S. (1999) Effects of small group learning on undergraduate Science, Mathematics, Engineering and Technology: a meta-analysis. *Review of Educational Research*, 69, 1, pp. 21-51.

Watson, W. E., Kumar, K. & Michaelsen, L. K. (1993) Cultural diversity ’s impact on group process and performance : comparing culturally homogeneou s and culturally diverse task groups, *The Academy of Management Journal*, 36(3), pp. 590–602.

White, F., Lloyd, H., Kennedy, G. & Stuart, C. (2005) *An investigation of undergraduate students’ feelings and attitudes towards group work and group assessment*. Higher Education Research and Development Society of Australasia Annual Conference, Sydney.

Yan, L. & Kember, D. (2003) Influence of the Curriculum and Learning Environment on Group Learning Approaches Outside the Classroom. *Learning Environments Research*, 6,3, pp. 285-307.

Acknowledgement

This literature review was funded by the Assessment Standards Knowledge Exchange, Centre for Excellence in Teaching and Learning in Higher Education, The Business School, Oxford Brookes University, and where Graham Gibbs is a Visiting Professor.

www.brookes.ac.uk/aske