Evaluating Peer Contributions to Group Work

Eric Tollefson

This paper was completed and submitted in partial fulfillment of the Master Teacher Program, a 2-year faculty professional development program conducted by the Center for Faculty Excellence, United States Military Academy, West Point, NY, 2015.

Introduction

Collaborative group project work in the classroom is common in many disciplines, because it teaches, encourages, and reinforces numerous skills and competencies that are desirable in the job market. However, group work brings with it a number of implementation challenges, one of the most important of which is varying contribution levels among group members. Peer evaluations, or assessments, allow teachers to understand and evaluate information about the individual efforts of team members that would otherwise be unavailable. There is a tremendous amount of literature on the topic of peer evaluation. This review surveys a sampling of the relevant literature with respect to peer evaluation development, administration, and challenges, as well as some alternative techniques and considerations.

Background

Collaborative group project work comes in a variety of forms. Martinazzi (1998) describes the concept of student learning teams (SLT), which are fashioned after similar teams in industry and emphasize familiarization with the challenging dynamics of being a member of a team. Mozas-Calvache and Barba-Colmenero (2013) describe the use of project-based learning (PBL) and problem-oriented learning (POL) in engineering education. They point out that engineers typically work in multidisciplinary teams, and thus both PBL and POL reinforce skills that students will have to possess or develop after graduation. Noonan (2013) provides a good overview of the project method and group work from her perspective in nursing education, citing sources that describe, for example, how it impacts student interest, independence, and resourcefulness; fosters a number of different learning strategies; and encourages active citizenship. These papers all agree that collaborative group project work has a significant number of benefits.

On the other hand, group work also has its challenges. Marks and O’Connor (2013) note that there are many students who are frustrated by group work due to differing ability levels and work ethics, as well as interpersonal conflicts. In their paper, they set out to gain insight into the attitudes of their business students concerning group work, and to determine if their attitudes differ from non-business students. They designed a survey to understand student perceptions of five main issues: organization of group work; value of group work; motivation of instructors; experiences with group work; and opinions of group work. Their analysis of the survey results lead them to conclude “that students generally understand the motivation of group work [and] see it as an important skill” (Marks and O’Connor, 2013, p. 157). But students also distrust teachers’ motivations for assigning group work. In general, they are interested in selecting their own groups and group leaders, and they also prefer smaller groups. Finally, and most
Important for this review, students generally prefer an individually tailored grade through the use of peer reviews (Marks and O’Connor, 2013). As matter of fact, the premise of nearly every article reviewed in this literature review is that the prevention of ‘social loafing’ or ‘free riding’ (terms used for group members who do not contribute as much to group products as the other team members) is essential in group work, and that peer evaluations are the best way to do that. However, a few alternatives to peer evaluations are presented later in the paper.

In addition to being useful for dealing with free riding, peer evaluations have also been shown to improve group performance in general. Brutus and Donia (2010) explore the effects of peer evaluations in improving group effectiveness over time. Specifically, the authors examine a group of undergraduate business students over the course of two semesters. All students used the same web-based peer evaluation system, and the authors chose both an experimental group and comparison (control) group. Their results indicate that group effectiveness did in fact improve over the course of two semesters, and that their performance was significantly higher than the control group who did not use the peer evaluation system.

In this review, the term ‘peer evaluation’ refers to the assessment of a group member’s contribution to group products, whether it be a self-evaluation, or an evaluation of another group member; the term will be used interchangeably with ‘peer assessment.’ The remainder of this paper will discuss the overall peer evaluation process; the development of peer evaluation instruments; some numerical considerations for integrating peer evaluations into an individual grade; peer evaluation bias; and alternatives to peer evaluations.

**Process Considerations**

There are a number of components to the process of obtaining and acting upon peer evaluation input. Gueldenzoph and May (2002) provide a very good overview of best practices concerning peer evaluations for group assessments. They break the process into the following steps:

- Building the foundation;
- Creating effective evaluation tools;
- Articulating the evaluation criteria;
- Ensuring student participation;
- Using formative evaluation;
- Using comprehensive summative evaluation; and,
- Assessing the evaluation process.

Their important contribution is a holistic perspective of the process, instead of a focus on a particular step. For example, they point out that it is important to ensure that students understand the evaluation process from the beginning; that students should have a role in developing their own evaluation criteria; and that the peer evaluation process itself should be evaluated at the end of the course. They also highlight that the evaluation process should be formative, i.e., that it should be administered in such a way that students have time to act on the feedback throughout the conduct of the project and to improve their performance. Jin (2012), Dyrud (2001), and Mentzer (2014) also reinforce the importance of a formative, rather than just summative, peer evaluation process.

Noonan (2013) also examines the overall peer evaluation process, but from an ethical
perspective. She addresses the ethical implications of the determination of group composition, prevention of ‘free-riding’, preparation of the students for group work, communication of guidelines for group work, assessment of group projects, and the evaluation of individual contributions. Many of her points align with the best practices of Gueldenzoph and May (2002). She states:

The key to group assessments is that assessments are well developed and address all the complexities inherent in group work. Students must be aware and involved in the development of group assessments to achieve an optimal outcome for this type of work. (Noonan, 2013, p. 1422)

Her focus is on group assessments overall—peer assessments that evaluate individual contributions are a critical subset of those. The author also specifically addresses the ethical role of the teacher in the process, an important reminder that group assessments are not just a student issue.

A particular aspect of the peer evaluation process that must be considered is its ease of use. To that end, a number of schools have developed web-based tools for designing, administering, analyzing, and providing feedback. Mentzer (2014) proposes a web-based peer evaluation tool called the Comprehensive Assessment for Team-Member Effectiveness (CATME), developed by the School of Engineering at Purdue University. The tool uses a rating scale which is behaviorally anchored (essential for subjective rating scales) and can be calibrated, which can help remove bias in individual responses. It allows evaluations in any of the following areas: team interaction, project work contribution, timeline management, expectation of quality, and knowledge and skills related to the material. CATME facilitates the assignment of an individual grade, as well as the administration of peer feedback for formative purposes. The author does not describe the complex algorithm used to calculate the contribution score. He remarks that more than half of the students indicated that the tool was beneficial. Other papers surveyed, such as that of Brutus and Donia (2010) and Mozas-Calvache and Barba-Colmenero (2013), describe their use of web-based peer evaluation systems in the context of their studies. Such web-based applications provide tremendous flexibility and significant time savings to the peer evaluation process.

There are some other interesting procedural variations with respect to peer evaluation administration. Most papers surveyed required participants to complete the peer evaluation surveys anonymously and individually. Dyrud (2001), on the other hand, requires the final evaluation (of three for the course) to be completed by the group, as a group. The first two evaluations in the course are based upon Martinazzi’s (1998) model and are completed anonymously for all members of the group. Those evaluations are used strictly for formative feedback and do not affect the students’ grades. For the third and final evaluation, a form is also filled out for each individual member of the team, but it is completed as a group and is signed by all team members. This evaluation counts for 30% of the project grade. The author claims that requiring the group to complete the form as a team forces them to work through any disagreements before submitting the form, and that this technique has virtually eliminated group dysfunction in her courses. Other variations include whether to require all students to complete the evaluation forms, or some subset. As part of a larger study, Jin (2012) creates a peer
evaluation system in which students only complete a peer assessment if they feel that an individual in the group failed to contribute. For that assessment, the student is asked to recommend a percentage of the group grade that the particular individual should receive. The teacher verifies the validity of the complaint, and, if justified, and uses the average recommendation to adjust the underperforming student’s grade.

Instrument Development

The development of the criteria against which teams and individuals will be evaluated is a significant aspect of peer evaluations. Martinazzi (1998) recommends that students be allowed to provide substantial input to the process. In fact, he begins by soliciting from his students how the instrument should be developed in the first place. Allowing his students to lead the process, they identify 39 areas to be evaluated; prioritize and select the most important areas; develop the actual questions and rating scale; determine the frequency of administration; and decide how to incorporate the results into individual grades. The paper’s primary contributions are the knowledge elicitation process itself, the resulting list of evaluation questions and rating scales, and the idea of gaining student buy-in into the peer evaluations through participation in the development process.

A characteristic of instrument development that must be considered is its complexity. In Jin’s (2012) experiment, he applies a different evaluation method to each of two courses. In one, he asks students to evaluate their peers based upon a number of criteria. Individual grades are then adjusted according to the technique developed by Conway, Kember, Sivan, and Wu (1993), which will be discussed later in this review. In the other course, the students only complete an evaluation if they feel that another team member did not contribute. As discussed previously, the evaluation simply asks for the percentage of the group grade that the individual member should get. Thus, the latter technique is simpler and more straightforward than the former method. When queried about the fairness of the assignment of individual grades at the end of the course, the students who used the simpler method responded more favorably. The author concludes that the fairness of the evaluation instrument and approach is not necessarily a function of its complexity, i.e., that developing more complex instruments that account for numerous dimensions does not necessarily increase students’ perceptions of fairness.

Individual Grade Assignment

The paper of Conway, Kember, Sivan, and Wu (1993) is a foundational, oft-cited work that comprehensively addresses individual grade adjustment for peer evaluations. They begin by reviewing and categorizing existing schemes into three bins. The ‘pool of marks’ technique allows the members of the group to distribute the total points of the group among its members. For example, if a group of four gets a group grade of 80% on their report, they would divide the 4 x 80 = 320 points among the team members. The authors chose not to pursue this method because of the burden placed on the team members and the risk of provoking competition among the group members. In ‘base plus/minus contribution mark’ techniques, students receive a base mark for the group product as well as individual contribution marks. These contribution marks are either added to or subtracted from the base mark to create the final grade. The primary concern for this category of techniques is the relative weighting of the two scores. The third
category includes ‘multiplication by weighting factor’ methods. These methods also involve both a base and a contribution grade, with the former being weighted, or scaled, by the later.

Conway, Kember, Sivan, and Wu (1993) choose to pursue techniques in the third category. They develop a two-part weighting factor technique that they subsequently discard, but then settle on a technique involving Individual Weighting Factors (IWF) that are multiplied by the overall group grade to determine individual grades. The IWF is calculated as follows. Each individual is evaluated numerically for a number of criteria. Those criteria ratings are added together to create an Individual Effort Rating. The Individual Effort Ratings for the members of the group are averaged to create the Average Effort Rating. The IWF, then, is simply the Individual Effort Rating divided by the Average Effort Rating. The authors also explore ways to reduce extreme deviations by scaling the IWF by an additional factor. Their feedback from students indicate that the technique was highly accepted.

Ko (2015) examines Conway, Kember, Sivan, and Wu’s (1993) IWF method and describes some of its drawbacks, such as the potential for a very high spread in score, and its inability to deal with variability in assessment scores. He also reviews newer techniques built upon the original IWF concept. He addresses the issue of whether or not to include self-assessments in IWF calculations, and concludes that self-assessments should be considered. He then addresses and demonstrates two fairness issues related to assessor reliability and assessee score variability. He then recommends a new, iterative IWF method that mitigates the identified fairness issues. The author’s demonstration of his technique on standard datasets shows its efficacy; however, because it is an iterative method, it cannot be calculated by means of a closed-form equation. It must be calculated through computer code, which can be implemented via macro in most common spreadsheet software packages.

Mozas-Calvache and Barba-Colmenero (2013) propose a base plus/minus contribution mark technique in which the contribution mark can be related to the individual’s contribution relative to the average group effort using either a linear or an exponential function. The latter allows the teacher to magnify deviations from average group performance, resulting in higher repercussions for free riding. As discussed previously, Jin (2012) recommends a simple method of directly eliciting from team members the recommended multiplicative grade adjustment to the group grade.

Bias Considerations

Peer evaluations are not without their potential drawbacks. One such challenge is bias. There are many possible types of bias. Dingel and Wei (2014) claim that there is little information in the literature about the potential biases inherent in peer evaluations with respect to gender, race, course performance and leadership role. Therefore, they design an experiment to investigate the effects of these variables on peer evaluations, and use multiple linear regression, analysis of variance, and other statistical methods to determine which variables have a significant effect on peer evaluation scores. The authors find that leadership and course grades are highly correlated with peer evaluations (supporting the validity of peer evaluations), as is race (undermining the validity of peer evaluations). However, due to the small proportion of “students of colour” (Dingel and Wei, 2014, p. 731) in their sample, they recommend that racial bias be explored
further in future studies. Their analyses do not identify gender as a significant factor in peer evaluations.

Tucker (2014) addresses gender bias specifically. His work spans six case studies involving two universities and three educational disciplines. He finds that women receive higher peer assessment scores than men, and that men tend to give higher scores in general. Nonetheless, although he finds a statistically significant difference between mixed- and same-gender assessments (favoring same-gender), the effect is quite small. Thus, he concludes that there is little evidence that gender leads to bias in peer assessments.

Mozas-Calvache and Barba-Colmenero (2013) develop a web-based peer evaluation system and apply it to 20 groups of four. In their analysis of the resulting peer evaluations, they draw some conclusions with respect to bias. First, they note that self-evaluations are typically higher than evaluations of an individual by other team members. Second, they show that the effects of gender and race are reduced in most of the dimensions examined. Third, they find that leadership dimensions receive lower scores in general, but that the peer evaluations are able to identify those who have been leaders.

**Alternatives**

Although peer evaluations seem to be the most common method to evaluate individual effort and to prevent free riding, there are other alternatives in the literature. Lambert, Carter, and Lightbody (2014) use a wiki concept to counter the biases inherent in peer evaluations. The wiki solution allows the teacher to see each individual member’s contributions to the group effort in order to facilitate a more objective assessment. The teacher can track not only the source of the contributions, but also the history of the changes and communications between the team members. The ability of the teacher to see the inner workings of the group obviates the need for relying on students to provide that information through the lenses of their own biases. The authors also recommend a step-by-step grading process which they demonstrate on an actual example. Their results show that the time required to assign initial grades to the projects using the wiki method is similar to that required for traditional methods.

Another technique to prevent free riding, developed by Swaray (2012), focuses on the design of the group project itself. In his example, the author forms groups through random selection. He then designs the project so that one of the primary outputs of the group project, a group presentation, is briefed by a randomly chosen group member. Although briefed by one group member, the grade for that part of the project is assigned to the entire group. The remaining parts of the projects are a combination of individual and group requirements. Based upon questionnaires that the author administered to assess the students’ opinions of the group project, he concludes that the random selection in group design and their assessments reduced free-riding and had other positive effects.

A third method to improve group dynamics involves an understanding of the effects of personality within groups. Rodríguez Montequín, Mesa Fernández, Balsera, and García Nieto (2013) examine the correlation between the Myers-Briggs Type Indicator (MBTI) and project group success. The MTBI is used to measure psychological preferences in perception and
decision making to determine personality type. The authors conclude that there are particular personality types associated with group success, both in terms of the mix of personality types and the personality type of the group leader. Their results provide great insight into team formation, assuming that MTBI results are available, or that it can be administered at the beginning of a course.

**Recommendations**

Based upon the existing literature, there are a number of things to consider when developing group projects and evaluating individual contributions. The following are some recommendations.

- Be aware of the ethical implications of assigning group work.
- Involve students in the development of the group work evaluation process.
- Ensure that students understand the criteria against which their peers will evaluate them, as well as methodology for adjusting their grades based upon those evaluations.
- Evaluate the peer evaluation process at the end of the course.
- Consider using other means to assess individual contributions to supplement, or even replace, peer evaluations.
- Use peer evaluations multiple times during the course so that it serves both formative and summative purposes.
- Make peer evaluation tools as simple as possible.
- Mitigate against potential biases in peer evaluations when possible.
- Consider web-based tools to reduce the burden of administering peer evaluations.
- Pay special attention to group formation. Students prefer to select their own groups; however, there may be reasons to direct membership. Personality type, such as that identified by the MTBI, is an additional aspect that can be considered.
- Consider the implications of how the project design affects group dynamics.

**References (Annotated):**


This paper explores the effects of peer evaluations in improving group effectiveness over time. Specifically, the authors examine a group of undergraduate business students over the course of two semesters. All students used the same web-based peer evaluation system, and the authors chose both an experimental group and comparison (control) group. Their results indicate that group effectiveness did in fact improve over the course of two semesters, and that their performance was significantly higher than the control group who did not use the peer evaluation system.


This paper provides a good survey of peer assessment methods used to adjust individual grades. The authors settle upon a multiplicative factor scheme which they apply in the classroom.
Unsatisfied with a number of the aspects of the technique, in particular, its complexity, the authors modify it. Their new method, using what they call the Individual Weighting Factor (IWF), bases an individual’s grade on the assessment of his/her peers relative to the average effort of the group. The authors also consider a common scaling factor to reduce the large deviations of individual scores within a group. In addition to addressing individual assessments, the authors also discuss peer assessments of the presentations of other groups. The IWF method discussed here has been referenced often and has resulted in a number of related techniques.


In this paper, the authors note that there is little information in the literature about the potential biases inherent in peer evaluations with respect to gender, race, course performance and leadership role. Therefore, they design an experiment to investigate the effect of these variables on peer evaluations, and use multiple linear regression, analysis of variance and other statistical methods to determine which variables have a significant effect on peer evaluation scores. The authors find that leadership and course grades are highly correlated with peer evaluations (supporting the validity of peer evaluations), as is race (undermining the validity of peer evaluations). However, due to the small proportion of ‘students of colour’ in their sample, they recommend that racial bias be explored further in future studies. Their analyses do not identify gender as a significant factor in peer evaluations.


This paper discusses the development and administration of a peer evaluation tool to counteract student group dysfunction in the author’s business classes. While she uses the results of Martinazzi (1998) to develop a peer evaluation tool, she departs from his example by administering the assessment three times per quarter, but only counting the last for grade. The first two peer evaluations are used for formative purposes to help students understand their own beneficial and detrimental behaviors. The last evaluation is completed by the group, as a group, and signed by all team members, and counts for 30% of the project grade. Requiring the group to complete the form as a team forces them to work through any disagreements before submitting the form. The author claims that this technique has virtually eliminated group dysfunction in her courses.


This paper is a survey of best practices concerning peer evaluations for group assessments. Especially valuable is the authors’ consideration of the entire process of peer evaluation development, including the involvement of the students prior to execution; the thoughtful creation of measureable evaluation criteria; the assurance of student participation; the implementation of a formative process instead of just a summative one; and the assessment of the peer evaluation process at the conclusion of the course.

This paper discusses the application of peer assessments in two different construction management courses that emphasize project-based, collaborative learning. In one course, the author applies the peer evaluation method of Conway, Kember, Sivan, and Wu (1993), which uses the Individual Weighting Factor (IWF) to adjust individual grades based upon peer assessments completed by all group members. In the other course, students only completed a peer assessment if they felt that an individual in the group failed to contribute. That assessment asked students to recommend a percentage of the group grade that the particular individual should receive, and uses that result to adjust the grade. Their primary conclusions are that more complex peer assessment instruments do not necessarily improve fairness, and that peer assessments should be used in a formative role, especially earlier in the course.


This paper examines the Individual Weighting Factor (IWF) method developed initially by Conway, Kember, Sivan, and Wu (1993), as well as a number of the related techniques that it generated. The author identifies two primary issues with existing IWF-based techniques. First, he addresses the issue of whether or not to include self-assessments in IWF calculations, and concludes that self-assessments should be considered. Second, he addresses and demonstrates two fairness issues related to assessor reliability and assessee score variability. He then recommends a new, iterative IWF method that mitigates the identified fairness issues. The author’s demonstration of his technique against standard data sets shows its efficacy.


This paper uses a wiki concept to counter the biases inherent to peer evaluations. The wiki solution allows the instructor to see each individual member’s contributions to the group effort in order to facilitate a more objective assessment. The instructor can track not only the source of the contributions, but also the history of the changes and communications between the team members. The authors also recommend a step-by-step grading process which they demonstrate on an actual example. Their results show that the time to assign initial grades to the projects using the wiki method is no longer than that required for traditional methods.


The authors of this paper focus on the attitudes of students concerning group work. To explore this question, they develop a survey administered to both business and non-business students. Their survey addresses student perceptions of five main issues: the organization of group work; the value of group work; motivation of instructors; experiences with group work; and opinions of group work. Overall, they conclude that there is no consensus concerning a preference for group work; however, students do prefer to choose their own teammates and leader (if required). They also demonstrate support for the use of peer evaluations.

The author discusses the value of “student learning teams” (SLT) for teaching students the advantages and challenges of being on a team. The implementation of SLTs in the author’s classes led his students to request an assessment of individual team members through peer evaluations that are tied to each member’s course grade. This paper describes the process that he used to elicit student input to develop a peer evaluation instrument. He discusses question development, prioritization, and selection; rating scale development; assessment frequency; and course grade integration. The author also describes the average results by question for three different courses. The paper is a very interesting example of allowing students to develop their evaluation tools.


This paper proposes a web-based peer evaluation tool called the Comprehensive Assessment for Team-Member Effectiveness (CATME), which considers evaluations in any or all of the following areas: team interaction, project work contribution, timeline management, expectation of quality, and knowledge and skills related to the material. The author’s use of CATME considers not only the assignment of an individual grade based upon contribution, but also the tool’s role in motivating the students to contribute through formative feedback in class. Importantly, the author also describes the tool’s behavioral benchmarking and calibration process which can help remove bias in individual responses. He points out that more than half of the students indicated that the tool was beneficial.


In this paper, the authors design an application of project-based learning in their engineering (surveying) course, with particular focus on evaluating individual contributions to group work. For grade assignment purposes, the authors use an additive model that adds or subtracts points from a baseline group grade based upon each individual’s peer evaluation. They develop a web-based peer evaluation system and apply it to 20 groups of four. Their primary contribution is the analysis of the peer evaluation results. They describe a number of results with respect to the effect of having groups whose team members have worked together previously; the difference between self-evaluation and the evaluations of other team members; and the assessment of individual team member leadership.


This paper addresses the ethical implications of group work overall, including determination of group composition, prevention of ‘free-riding’, preparation of the students for group work, communication of guidelines for group work, assessment of group projects, and the evaluation of individual contributions. The paper provides an important perspective of the role of evaluating individual contributions within the larger concept of group work in general. The author also specifically addresses the ethical role of the lecturer in the process. The paper is a very good survey of primary ethical concerns related to group work.

This paper examines correlation between the Myers-Briggs Type Indicator (MBTI) and project group success. The MTBI is used to measure psychological preferences in perception and decision making to determine personality type. The authors conclude that there are particular personality types associated with group success, both in terms of the mix of personality types and the personality type of the group leader. Their results provide great insight into team formation, assuming MTBI results are available, or that it can be administered at the beginning of a course.


This paper attempts to solve the issue of free-riding through the entirety of the group project design. Groups were formed by the author through random selection. The outputs of the group project include a group presentation that is briefed by a randomly chosen group member, a group report, a short-answer essay completed individually by each member of the group, and individual reflections. After the conclusion of the projects, the author administered questionnaires to the class to assess the students’ opinions of the group project. The author concludes that the random selection in group design and their assessments reduced free-riding and had other positive effects.


The author builds upon previous studies that set out to determine whether there is a gender bias in peer evaluations. His work spans six case studies involving two universities and three educational disciplines. His primary conclusion, as the title suggests, is that there is little evidence that gender leads to bias in peer assessments. He does find that women receive higher peer assessment scores than men, and that men tend to give higher scores. Nonetheless, although he finds a significant difference between mixed- and same-gender assessments (favoring same-gender), the effect is quite small.