Assessing the reliability of self- and peer rating in student group work

Bo Zhang \textsuperscript{a}, Lucy Johnston \textsuperscript{b}, Gulsen Bagci Kilic \textsuperscript{c}

\textsuperscript{a} University of Wisconsin - Milwaukee, USA
\textsuperscript{b} University of Canterbury, New Zealand
\textsuperscript{c} Abant Izzet Baysal University, Turkey

Online Publication Date: 01 June 2008


URL: http://dx.doi.org/10.1080/02602930701293181
RESEARCH PAPER

Assessing the reliability of self- and peer rating in student group work

Bo Zhang*a, Lucy Johnstonb and Gulsen Bagci Kilicc

aUniversity of Wisconsin – Milwaukee, USA; bUniversity of Canterbury, New Zealand; cAbant Izzet Baysal University, Turkey

Peer and self-ratings have been strongly recommended as the means to adjust individual contributions to group work. To evaluate the quality of student ratings, previous research has primarily explored the validity of these ratings, as indicated by the degree of agreement between student and teacher ratings. This research describes a Generalizability Theory framework to evaluate the reliability of student ratings in terms of the degree of consistency among students themselves, as well as group and rater effects. Ratings from two group projects are analyzed to illustrate how this method can be applied. The reliability of student ratings differs for the two group projects considered in this research. While a strong group effect is present in both projects, the rater effect is different. Implications of this research for classroom assessment practice are discussed.

Introduction

While group work is a popular instructional tool at various levels of teaching (Slavin 1995; Johnson & Johnson 1996; Lejk et al. 1996), grading group work for evaluation and reporting purposes is often not the most enjoyable part of a teacher’s everyday activities. Even after long training sessions, many teachers still experience a high level of uneasiness and frustration (Ross et al. 1998). Apparently, assigning all group members the same score, regardless of how much each member has contributed, seems unfair and thoughtless and this practice does not motivate most students (Kagan 1995). Therefore, the major concern in evaluating group work is how to account for individual contribution difference (Gillies & Ashman 2003). This, unfortunately, is not an easy undertaking in most educational settings, where teachers only have the final product from group work, which represents a collective effort of all group members. Without additional information indicative of the group process, it is difficult for teachers to decompose that group effort into individual efforts.

Most methods proposed to account for individual efforts in group work rely on peer and self-rating of contribution to group work by students themselves (Conway et al. 1993; Lejk et al. 1996; Lejk & Wyvill 2001, 2002; Johnston & Miles 2004). Although teachers may not be fully cognizant of the individual contribution of each group member, students are well aware of how much each individual has contributed to the success or failure of their group. These self- and peer ratings contain the vital, if not the only, source of information available to gauge the individual contribution. Among the nine methods reviewed by Lejk et al. (1996), at least seven require peer reviews. For example, in a typical group project assignment, all group members might work together without being evaluated on their progress by the teacher. In this case, an overall evaluation of the group might then be obtained by having students rate each group member’s contribution, including their own. An individual student’s score, which is reflective of his/her

*Corresponding author. Email: boz@uwm.edu
contribution to the overall group project, can then be obtained by summing or averaging the ratings across all group members. If the group project also has an individual student work component, students’ final scores can reflect a combination of their group and individual efforts, with the group effort component reflective of the differential contribution among group members.

Methods that rely on peer and/or self-ratings to account for an individual’s efforts in a group project must consider the quality of these ratings themselves. Imagine the consequences if students have rated each other’s contribution unfairly: any adjustment to a student’s individual score would be biased. In other words, to adjust students individual scores, these ratings must be valid and reliable. To be valid these ratings must accurately reflect the unique contribution of each individual group member. To be reliable these ratings must be fairly consistent across different group members.

Past research on the quality of peer and self-ratings has primarily focused on the validity of these scores. Although concerns have been raised with regard to the ability of students to rate each other fairly (Burke 1969), a large body of research shows a high degree of agreement between student and teacher ratings in various disciplines of higher education, as reviewed by Falchikov and Boud (1989), and Falchikov and Goldfinch (2000). Research on the reliability of peer and self-rating in group work, however, is limited. In other words, most research pertaining to the quality of peer and self-rating has been devoted to the agreement between student and teacher ratings, as opposed to the consistency of ratings among students themselves. One possible reason for this lack of research is that traditional reliability indices, such as the correlation coefficient, the percentage of agreement, and the Kappa statistic (Cohen 1960), cannot be easily applied in this situation. In group work, each student receives ratings only from other members of that group, thus it is hard to calculate traditional reliability indices across all groups. Moreover, if these indices are computed at the group level then they will most probably be unstable due to the small sample size. Finally, it is not convenient for these statistics to accommodate the multiple ratings for each member by group members.

This research explored the reliability of peer and self-rating in group work in two ways. First, it demonstrated how the reliability of peer and self-rating obtained from group work can be estimated under a Generalizability Theory framework (Cronbach et al. 1972; Shavelson & Webb 1991). Second, factors that might affect peer and self-rating in group work were examined. To that end, these specific research questions were addressed:

(1) How reliable are self- and peer ratings from group work?
(2) To what extent does the inclusion of students’ self-rating affect the reliability of peer and self-ratings?
(3) How large are the group effect and the rater effect?
(4) How reliable are the ratings on specific contribution indicators, such as motivation or communication?

Answers to these questions provide important empirical information on how to make use of peer and self-ratings to obtain individual student scores that are reflective of their contribution to group work. The reliability coefficients computed to answer the first two questions will directly show whether peer and/or self-ratings are reliable enough to be taken advantage of. The third question has practical significance. If a large rater effect is found then student training may be necessary to enable students to fairly rate themselves and their peers, in terms of their group contribution. Meanwhile, if a large group effect is found then it may be necessary to form groups in a different manner. Results from answering the fourth question will show what indicators are reliable when multiple indicators are employed in measuring individual contribution.
Peer and self-rating under Generalizability Theory framework

Generalizability Theory (G theory) addresses the weakness in classical test theory in which the error term is an undifferentiated mixture of multiple error sources (Shavelson & Webb 1991). G theory decomposes the systematic and unsystematic sources of error into multiple variance components, by considering different sources of errors that may have contributed to the error term in a measurement design. These sources are referred as facets in G theory terminology. When considering peer and self-rating in group work, the major facets are rater effect and group effect. The rater effect indicates the variability in scores due to students that are generally stricter or more lenient in their ratings than others. The group effect, on the other hand, indicates the variability in scores due to particular groups being different in their ratings overall. It is also possible that an interaction exists between the group and rater effects. This would occur if students’ rating depends on the group they were assigned to. As each group comprises multiple raters, the rater facet is considered to be nested within the group facet, marked as rater:group. Contingent on the measurement objective, the group and rater facets may be considered as random or fixed effects. A random effect would imply the raters or the groups are a random sample of the universe. They may also be considered as a random effect if the universe of generalization is sufficiently large (Shavelson & Webb 1991). A fixed effect, on the other hand, limits the generalizability of results to the investigated sample of the raters or groups only. In studying the reliability of peer and self-ratings from group work, instructors are interested in whether the rating of each student is reliable, regardless of who the raters are, or which group each student is in. Therefore, group and rater effects are better treated as random than as fixed, as exemplified by the two studies in this research.

A Generalizability Theory analysis is often conducted in two stages (Shavelson & Webb 1991; Brennan 2001). The first stage is the generalizability study, referred to as the G study. It defines the variance components and estimates their magnitude. Traditional analysis of variance (ANOVA) methods can be utilized to obtain these estimates. However, for unconventional designs such as the peer and self-rating in group work, the use of alternative estimation methods is often preferred. The second stage is the decision study, referred to as the D study. In this stage, information from the first stage is used to compute summary reliability coefficients. Researchers may also optimize the possible resources to set up the best design in achieving the desired reliability.

The self- and peer ratings obtained from group work pose some unique challenges in studying reliability using Generalizability Theory. The major difficulty is that raters cannot be separated from ratees (persons to be rated), resulting in an atypical multilateral rating situation. This uniqueness may have an impact on how each rater assigns grades because during the rating process student raters are fully aware of the fact that they are being rated by peers simultaneously. Consequently, more able students could deliberately deflate their self-rating (Krause & Popovich 1996; Lejk & Wyvill 2001) and the so-called ‘hitchhikers’ or ‘free-riders’ (Kerr & Bruun 1983)—group members who failed to do their duties but received the same score as other members—may intentionally inflate self-rating and/or peer rating (Falchikov & Goldfinch 2000).

The inseparability of raters from ratees also affects the computation of the reliability coefficient. In evaluating the reliability of peer and self-ratings, the object of measurement should be the rating of persons on their contribution to group work. The G-study design is a random-effect nested design. Specifically, it is the person facet (p) crossed with the rater (r) facet, but both are nested within the group (g) facet, or (p x r):g in G-theory notation, as demonstrated in Figure 1. Both the group and rater facets are random-effect facets as they are treated as random samples from all the possible groups and raters (i.e. the group and rater universe).
As shown in Figure 1, the total variance in rating each person \( \sigma^2(X_{gp}) \) is decomposed into four components:

\[
\sigma^2(X_{gp}) = \sigma^2_g + \sigma^2_{p,p_g} + \sigma^2_{r,rg} + \sigma^2_{p_r,p_g,e},
\]  

where \( \sigma^2_g \) is the variance of rating across different groups; \( \sigma^2_{p,p_g} \) is the variance due to person and person–group interaction; \( \sigma^2_{r,rg} \) is the variance due to rater and rater–group interaction; and \( \sigma^2_{p_r,p_g,e} \) is the variance due to rater–person interaction, rater–group–person three-way interaction, and other error terms not accounted for by the facets built into this design. The higher the proportion of the total variance \( \sigma^2(X_{gp}) \) accounted for by the person variance \( \sigma^2_g \), the higher the reliability will be. In other words, the peer and self-rating will be reliable if the different ratings reflect the different contributions of students, rather than different rating strategies among raters or groups.

An independent person variance component is not extractable from this design as it is always confounded with the person-nested-in-the-group effect, which certainly is a disadvantage. This issue can be handled in two ways. The first option is to ignore the group effect and change the design into a one-facet person by rater crossed design. However, as each person is rated by raters from one’s own group only, this change will result in a large amount of missing data, amounting to \( (n_g - 1)/n_g \), where \( n_g \) is the number of groups. Moreover, in classroom assessment situations, groups are often self-selected. There is little ground to assume that the groups are equal. Ignoring the group effect would simply confound it with other effects in the model, making them hard to interpret or control. The second option is to treat person nested within group as the object of measurement, assuming that persons are always nested in groups. As this assumption will always hold in the peer and self-rating of group work situations, the analysis of two group projects in this research followed this option.

Another unique aspect to this design is its high likelihood to be unbalanced due to different group sizes. For example, for the projects studied in this research, group size ranges from 2 to 5. Coupled with the limited sample size frequently encountered in classroom assessment, the traditional least-square estimation of variance components can be inaccurate. Instead, the maximum likelihood (ML) or the restricted maximum likelihood (REML) procedures need to be considered. The ML method maximizes the likelihood function over the parameter space in the model as expressed in Equation 1. The restricted ML method separates the likelihood function into the fixed and the random parts and maximizes the random part only. For the exact form of the likelihood function under these two methods, refer to Swallow and Monahan (1984). These methods
produce estimators that are normally distributed (Searle 1987). Compared with the ANOVA method, the REML method provides non-negative variance. As the negative variance estimates are usually related to small sample size, this method is more robust for the sample size requirement. In addition, it also provides more stable estimates as its standard error of parameter estimates is smaller than other methods (Marcoulides 1990). Accordingly, in this research, variance estimation was conducted by the REML method, available in the SAS PROC VARCOMP procedure (SAS Institute, Cary, NC, USA).

With the variance components estimated, two reliability coefficients can be computed to reflect the consistency of rating in the D-study stage. These coefficients reflect the proportion of the expected observed score variance explained by the universe-score variance. They can be interpreted in a similar manner to the reliability coefficient in the classical true-score model theory, such as coefficient alpha (Cronbach 1951). While the appropriate degree of reliability depends on the purpose of measurement, as a general guideline, the commonly accepted lower limit for alpha is 0.7 (DeVellis 1991; Nunnally 1978). This value is selected so that the standard error of measurement is small enough compared with the standard deviation of the test score.

The generalizability coefficient (\( \xi^{2} \)) addresses the relative error in making decisions. This type of error comes from the variance components related to the relative rank ordering of individual students. It consists of all interaction terms with the person facet. The generalizability coefficient for the two studies in this research is computed as:

\[
\xi^{2} = \frac{\sigma_{pg}^{2}}{\sigma_{pg}^{2} + \sigma_{rel}^{2}} = \frac{\sigma_{pg}^{2}}{\sigma_{pg}^{2} + \frac{\sigma_{pg,pp}^{2} + \sigma_{pg,pe}^{2}}{n_r}}
\]

(2)

where \( \sigma_{rel}^{2} \) is the relative error; \( n_r \) is the number of raters; and other terms share the same meaning as in Equation 1. The index of dependability (\( \phi \)) (Brennan & Kane 1977) measures the consistency in making absolute decisions. An absolute decision focuses on the absolute level of an individual student’s performance with no reference to others. The absolute error term includes all the variance components in the model except the object of measurement. The index of dependability is obtained by:

\[
\phi = \frac{\sigma_{pg}^{2}}{\sigma_{pg}^{2} + \sigma_{abs}^{2}} = \frac{\sigma_{p}^{2}}{\sigma_{p}^{2} + \frac{\sigma_{pg,pp}^{2} + \sigma_{pg,pe}^{2}}{n_r} + \frac{\sigma_{rg}^{2}}{n_r} + \frac{\sigma_{g}^{2}}{n_g}}
\]

(3)

where \( \sigma_{abs}^{2} \) is absolute error and \( n_g \) refers to the number of groups.

In classroom assessment settings, instructors are usually more interested in judging whether an individual student has met a learning standard or not than in comparing this student’s performance against others. In that sense, teachers are more likely to make absolute decisions. Following that logic, the index of dependability was reported for the two studies investigated in this research.

**Study 1**

**Method**

This study used the data from Bagci Kilic and Cakan (2006). Data were collected in one semester. Two group projects (referred as Group Work I and Group Work II hereafter) were assigned in the
semester. These projects were independent of each other. Each group was required to jointly plan a science lesson and team teach it. Subjects were 134 students enrolled in a teacher education program. They were assigned to three classes taught by a single instructor. In each class, students selected their own groups. Group size ranged from three to five. Sample size is quite limited for all three classes. The distribution of students into classes and groups is as follows: 39 students divided into nine groups in class 1, 30 students into eight groups in class 2, and 26 students into eight groups in class 3. Note that only groups with complete ratings were included in this study.

Peer and self-rating of contribution to the group work were requested after students finished team teaching. The ratings were based on an analytic scoring rubric (Goldfinch 1994; Lejk & Wyvill 2001), which considered the following six indicators of quality team work: (1) motivation/responsibility/time management, (2) adaptability to other team members, (3) creativity, (4) communication skills, (5) general team skills, and (6) technical skills. These factors reflect group work effort (indicators 1, 2, 4 and 5) and academic contribution to the final product out of the team work (indicators 3 and 6). A six-point scale was used to rate each student on each indicator: 4 for outstanding; 3 for better than most of the other members; 2 for about average; 1 for not as good as most of the other members; 0 for no help at all; and −1 for hindrance to the group.

The G-study used the two-facet random-effect design, as illustrated in Figure 1. Data from each class under each project were analyzed separately. The variance component was computed by the REML method. In calculating the generalizability coefficients, conditions in the original design, such as number of groups and number of raters in each group, were utilized to show the reliability of the peer and self-ratings.

Results

Table 1 presents the percentage of each variance component in the total variance of both peer and self-ratings combined. This analysis is based on the sum of all six indicators. The percentage of variance components differs across three classes. Even within each class, the percentages sometimes changed considerably from Group Work I to Group Work II. For example, in class 3, group variance increased from 6% to 35%. Still, some general trends over all three classes can be observed. The rater effect is the weakest among the four effects in the design. It never accounts for more than 12% of the total variance. Conceptually, this relatively small variance indicates that students are actually quite consistent in rating the contribution of group members.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Person (%) (σ²_p)</th>
<th>Rater (%) (σ²_r)</th>
<th>Group (%) (σ²_g)</th>
<th>Error (%) (σ²_e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Work I:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 1</td>
<td>42</td>
<td>2</td>
<td>42</td>
<td>13</td>
</tr>
<tr>
<td>Class 2</td>
<td>39</td>
<td>9</td>
<td>33</td>
<td>19</td>
</tr>
<tr>
<td>Class 3</td>
<td>64</td>
<td>4</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>Mean</td>
<td>49</td>
<td>5</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>Group Work II:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 1</td>
<td>33</td>
<td>12</td>
<td>21</td>
<td>34</td>
</tr>
<tr>
<td>Class 2</td>
<td>48</td>
<td>2</td>
<td>42</td>
<td>9</td>
</tr>
<tr>
<td>Class 3</td>
<td>36</td>
<td>10</td>
<td>35</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>40</td>
<td>8</td>
<td>32</td>
<td>21</td>
</tr>
</tbody>
</table>
Second, there is a strong group effect. With the exception of the first group project (i.e. Group Work I) for the third class the variance component for the group effect accounts for approximately one-third of the total variance. Finally, all the error variance terms increased from Group Work I to Group Work II.

Table 2 presents the dependability index for the above rating. The without-self-rating column shows the reliability obtained when considering only peer ratings, not self-ratings. In general, the reliability is acceptable for almost all conditions. The first group work rating is more reliable than that of the second group work. The reliability obtained when considering only peer ratings is less than when both peer and self ratings are utilized combined, but the difference is small. In this study, the inclusion of self-ratings did not negatively bias the overall reliability.

Table 3 depicts the reliability of the two distinct indicators evaluated by peer and self-ratings: group work effort and academic contribution. The group work effort indicates how members were evaluated in terms of behaving as a responsible group member while the academic contribution factors reflects how members were rated in terms of their contribution to the academic quality of the final group product. The group effort component consists of four such indicators: motivation management, adaptability, communication skills and general team skills. The academic achievement component includes the remaining two indicators—creativity/originality and technical

Table 2. Dependability Index for the overall rating in Study 1.

<table>
<thead>
<tr>
<th>Classes</th>
<th>With self-rating</th>
<th>Without self-rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Work I:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 1</td>
<td>0.85</td>
<td>0.86</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>Class 3</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>Mean</td>
<td>0.83</td>
<td>0.81</td>
</tr>
<tr>
<td>Group Work II:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 1</td>
<td>0.71</td>
<td>0.63</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>Class 3</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td>Mean</td>
<td>0.77</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3. Dependability Index for the effort and academic contribution components in Study 1.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Group effort</th>
<th>Academic contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Work I:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 1</td>
<td>0.85</td>
<td>0.71</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>Class 3</td>
<td>0.87</td>
<td>0.78</td>
</tr>
<tr>
<td>Mean</td>
<td>0.86</td>
<td>0.77</td>
</tr>
<tr>
<td>Group Work II:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 1</td>
<td>0.80</td>
<td>0.64</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td>Class 3</td>
<td>0.69</td>
<td>0.57</td>
</tr>
<tr>
<td>Mean</td>
<td>0.72</td>
<td>0.61</td>
</tr>
</tbody>
</table>
skills. Overall, ratings on the group effort factor were found to be more reliable than those related to an individual’s academic contribution and once again the reliability was found to decrease from first group work to second group work on both factors.

Table 4 displays the reliability of the ratings obtained from the six individual indicators included in the scoring rubric. As the results indicate, there were large discrepancies in the reliability indices obtained across the three different classes. However, within each class, the ratings across the different indicators are quite similar. This seems to suggest that an overall rating tendency affected the ratings on all the indicators. For example, for Class 2 under Group Work I and Class 3 under Group Work 2, the ratings for all indicators were quite low. When comparing the results in Table 3 to the results in Tables 1 and 2, the dependability indices obtained for the individual indicators seem quite low: many are below 0.5. Therefore these ratings are probably not reliable enough to be used individually.

### Study 2

#### Method

This study analyzed the data collected in Johnston and Miles (2004). Sixty-one undergraduate students in one class were divided into 15 groups. Ratings from one student were not used in this study as they did not follow the scoring rubric. Group size ranged from three to five members. Like Study 1, students selected their own groups. The group work was to complete a piece of social psychological research, including research design, data collection and data analysis. For this group project, each individual was required to write up the results of the project based on the group work. This differed from the first study in which each student received a group score. The group project took 10 weeks.

Both peer and self-ratings that reflected group members’ contribution to the project work were requested. Ratings were based on an analytical rating scheme that included seven indicators of an individual’s contribution, six of which were related to the academic contribution. For each indicator, the following rating scale was applied: 3 for major contribution; 2 for some contribution; 1 for minor contribution; 0 for no contribution; and −1 for a hindrance to the group. The major difference between this scale and the scale used in the first study is that the four-point category for outstanding contribution was not used.
The G-study design utilized in this study is exactly the same as that used in the first. However, as ratings on individual indicators are not available in this study, all analyses were based on the sum of the seven indicators.

### Results

Table 5 presents the variance components when both self- and peer ratings are considered. As the results indicate, in this study the rater effect variance component was largest, accounting for about one-third of the total variance. Unlike the students in the first study, students in this study seemed less capable of consistently rating each other’s contribution. The variance component due to the group effect, while smaller than that from the first study, is still quite large, accounting for about one quarter of the total variance. One major difference from the first study is the effect of including a student’s self-rating, which increases the error variance considerably. However, including a student’s self-rating did not seem to affect the estimates of rater and group variance components. In other words, that increase in error variance mainly comes from the decrease in person variance. Nevertheless, the dependability index for the combined peer and self-rating is quite low, 0.36, and when a student’s self-rating is removed it increases to 0.63. The finding that self-rating is quite different from peer rating is consistent with previous research on the effect of self-rating (Falchikov 1991; Lejk & Wyvill 2001). In this case, students actually inflated their self-ratings of their contribution to the group project (Johnston & Miles 2004).

### Discussion

This research described a Generalizability Theory framework to evaluate the reliability of peer and self-rating in grading group projects. The design is unique in the sense that a pure person variance component cannot be identified. This research suggested using person nested in group as the object of measurement. Also, due to unbalanced design and small sample size, the maximum likelihood estimation method is recommended to obtain the variance component estimation.

Both studies in this research showed a strong group effect. In other words, although the instructors had prepared a detailed analytical scoring rubric, students in different groups still rated each other’s contribution using different standards. Some groups were more lenient or stringent in grading each other’s contribution. This group effect might be due to the voluntary nature of the group formation in the two studies. To boost student motivation and involvement, instructors in both studies allowed students to form their own groups. One way to reduce the group effect is to randomly assign students to groups, which, unfortunately, seems to be at odds with the effort to get students more engaged and interested in the assigned group project.

The group effect seemed to increase as group members stay together longer, as shown in the increased group variance from Group Work 1 to Group Work 2 in Study 1. The familiarity with other group members may have made it more difficult for students to rate each other objectively.
Thus, in classroom assessment, reshuffling group membership periodically may be necessary to achieve higher rating reliability.

The difference in the rater effect between these two studies may have resulted from the different background of the participants. It is likely that the student teachers in the first study had received more training on how to use an analytical scoring rubric than did the undergraduate psychology majors in the second study. This was also reflected in the results obtained when a student’s self-rating was included, which greatly reduced the dependability of overall scores in the second study. Meanwhile, the nature of the group work itself may also have had an impact on self-rating. As individual work was requested in the second study, students might spuriously believe that they had done more than their peers in the group work. However, these two studies were conducted in different disciplines and cultural settings. The present study investigates the reliability of peer and self-ratings from the final results of rating only. What exactly has caused the difference in these ratings requires more in-depth research. In addition, this research estimates the reliability of peer and self-ratings at the group level. To find out why one individual rater has behaved differently from his/her peers, qualitative analysis is needed.

To adjust individual contributions, it is necessary that reliable ratings are used. In the first study, multiple ratings were found to be reliable and therefore they can be used to generate scores with different interpretations and for different purposes. For example, an individual score can be derived based on how he/she was rated by their peers and how he/she rated him/herself in terms of his/her academic contribution only. This score can then be interpreted as the level of achievement of each individual student in the group work. However, for the second study, utilizing a student’s self-rating had a significant negative effect on the reliability. In terms of reliability, these ratings should probably not be used for a student’s overall score. As each student’s score from the group project considered in the second study had two parts, an individual and group component, it would make sense to use peer ratings for the group component only. On the other hand, there are reported benefits of self-ratings (Goldfinch 1994) and so excluding them might not always be the best idea.

The similarities and differences between these two studies point out some general trends in the reliability of peer and self-rating of the contribution to group work. One can use a Generalizability Theory framework to evaluate whether students’ ratings are reliable, as demonstrated by these two examples. This study utilized the univariate framework to estimate the reliability of peer and self-rating. In situations when an analytical rating scale is applied, such as in Study 1, a multivariate analysis (Brennan 2001) may also be conducted to estimate the reliability of rating on multiple indicators. In any case, only when reliable ratings are achieved can they be used in any of the methods designed to adjust the group grades to reflect individual contribution.

Finally, considering the potentially strong group effect, some random element should be built into the group formation process in classroom assessment whenever possible. Instead of simply letting students form their groups, teachers may ask students to randomly draw their group membership from a pool of numbers prepared. Post hoc adjustments may be necessary to increase the involvement of all students in the group work. To reduce the rater effect, training seems imperative. This may be particularly effective for subjects with little rating experience. This research also suggests that self-rating in group work should be securitized carefully. Dependent on the nature of the group work, students can rate themselves very differently from the way they rate others.

Notes on contributors
Bo Zhang is an assistant professor in School of Education at University of Wisconsin – Milwaukee, USA. His main research interests are model data fit, model parameter estimation, and missing data treatment in the field of educational measurement.
Lucy Johnston is an Associate Professor of Social Psychology at the University of Canterbury, New Zealand. Her main research and teaching interests are in stereotyping, stigmatization and non-verbal behavior in social interactions.

Gulsen Bagci Kilic is an associate professor in the School of Education at the Abant Izzet Baysal University, Turkey. Her main interests are in elementary science teacher education, with particular reference to self- and peer assessment of student performance in different contexts.

References


B. Zhang et al.