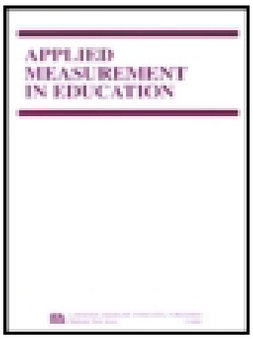

Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: www.tandfonline.com/journals/hame20

How to Assign Individualized Scores on a Group Project: An Empirical Evaluation

Bo Zhang & Matthew W. Ohland

To cite this article: Bo Zhang & Matthew W. Ohland (2009) How to Assign Individualized Scores on a Group Project: An Empirical Evaluation, Applied Measurement in Education, 22:3, 290-308, DOI: 10.1080/08957340902984075

To link to this article: https://doi.org/10.1080/08957340902984075

APPLIED MEASUREMENT IN EDUCATION, 22: 290–308, 2009

ISSN: 0895-7347 print / 1532-4818 online
DOI: 10.1080/08957340902984075

Routledge
Taylor & Francis Group

How to Assign Individualized Scores on a Group Project: An Empirical Evaluation

Bo Zhang
Department of Educational Psychology
University of Wisconsin–Milwaukee

Matthew W. Ohland
School of Engineering Education
Purdue University

One major challenge in using group projects to assess student learning is accounting for the differences of contribution among group members so that the mark assigned to each individual actually reflects their performance. This research addresses the validity of grading group projects by evaluating different methods that derive individualized scores from group work. Both Monte Carlo simulation and real test data analyses were conducted. The four investigated methods are the within-group adjustment method, the partial adjustment method, the between-group adjustment method, and the expected contribution adjustment method. For all methods, a weighting factor is computed based on the peer and self ratings of contributions to the group project by group members. This study finds that individual differences have to be taken into account if group grades are going to be assigned and utilized for evaluating individual performance at all. Adjusting contribution differences based on peer and self ratings could be an effective way to improve the validity of group grades. Among the four studied methods, adjusting both the within-group and between-group contribution differences is the most effective, and is thus recommended for classroom use.

Group work plays a very important role in making learning happen for students (Johnson & Johnson, 1999). As an essential component of cooperative learning, group projects encourage students to take more control of the learning process

Correspondence should be addressed to Bo Zhang, University of Wisconsin–Milwaukee, Department of Educational Psychology, P.O. Box 413, Milwaukee, WI 53201-0413. E-mail: boz@uwm.edu

through peer learning and self learning. As an instructional tool, group work benefits students on both cognitive and affective outcomes. Students in cooperative groups have been shown to be able to obtain significantly higher test scores than those studying alone (Sherman & Thomas, 1986; Springer, Stanne, & Donovan, 1999; Webb, 1995). In addition, group work helps increase motivation (Lourdusamy & Divaharan, 2000; Springer et al., 1999) and cultivates self-esteem and social skills (Slavin, 1995). To be effective group members, students have to learn how to work together in achieving common goals. In this sense, group work in schools provides a valuable simulation of the project-oriented teamwork in their future career.

Group work also serves as an assessment tool in schools. Teachers often evaluate student performance on group projects and use that alone or with other assessment results for grading or reporting purposes. As group projects are usually designed around real-life problems, they could be highly authentic assessment tasks. However, it is much more challenging to use group work as an assessment tool than as an instructional tool. To use grades from group projects to report the achievement level of individual students in the same manner as grades from individual assessment tasks, the assigned marks should accurately reflect each individual's knowledge or skills in the assessed content domain. This, by no means, is easy to achieve as it is hard to evaluate how much knowledge or skills one individual student has demonstrated from doing a group project.

In a typical group project assignment, all group members work together without being evaluated individually on their progress by the instructor. When a grade is assigned for the final product of the group work, all group members usually receive that same grade. As contributions generally differ among group members, the universal group grade may not reflect the actual performance of any member. Rather, it embodies the collective efforts of all group members and is indicative of the effectiveness of the overall teamwork. Without adjusting for contribution differences, both the fairness and validity of using group grade to report the achievement level of individuals would be threatened. For example, the equity of group mark would be reduced by the so-called free-riders, people who fail to contribute a fair share but receive the same score as others (Bartlett, 1995; Latane, Williams, & Harkins, 1979). The free-rider effect could be especially severe in large groups (Kerr & Bruun, 1983). For all the reasons Kagan (1995) listed that group grades should be banned in schools, the fundamental issue is the lack of accountability for individual contributions. In other words, if group grades are assigned and utilized for evaluating individual performance at all, individual contributions have to be taken into account.

In most classrooms, teachers usually have little resources to monitor the progress of all groups, thus it is hard for them to determine the exact contributions of each group member. However, students should be aware of how much each group member has contributed to the success or failure of the group project. When observation of group progress is not available otherwise, asking students to

rate their own as well as each other's contributions to the group work may provide the necessary information to gauge contribution differences. Based on the self and peer ratings on the contributions, a weighting factor may be derived for each student. For members whose contributions have been rated above the group average, their weights would favor them and, naturally, their individual scores would be higher than the universal group score.

One major criterion on whether student ratings should be used to account for individual contribution differences is the quality of the ratings. If the peer and self ratings on contributions to the group work cannot be trusted, any adjustment based on those ratings would be biased as well. Although concerns have been raised with regard to students' inability to rate their own work (Burke, 1969; Freeman, 1995), a large body of research from the field of higher education (Conway, Kember, Sivan, & Wu, 1993; Freeman, 1995; Goldfinch, 1994; Johnston & Miles, 2004; Lejk & Wyvill, 1996, 2001a, 2001b, 2002; Zhang, Johnson, & Bagci Kilic, 2008) shows that with proper training, students are capable of rating their own work. Lejk et al. (1996) reviewed nine methods used to derive individualized scores for group projects in various instructional settings. Seven of these nine methods rely on peer and self ratings of group contribution. It should be noted that most of the evidence on the validity of peer and self ratings for group work has been accumulated in the field of higher education. As the ability to conduct peer and self ratings is clearly age- and experience-dependent, relevant research on students in the elementary and secondary education is also important, which, however, is not the focus of this work.

Peer and self ratings may take on different forms in different assessment situations. They may be based on a holistic rubric with one single composite indicator (e.g., Brown, 1995; Johnston & Miles, 2004) or on an analytical rubric with multiple indicators such as attendance, cooperativeness, willingness, and academic contribution, to reflect the complex process of group work (e,g., Bagci Kilic & Cakan, 2006; Kaufman, Felder, & Fuller, 2000; Stefanou, Hood, & Stefanou, 2001). For the purpose of summative assessment, a holistic rubric has been found to be more effective (Falchikov & Goldfinch, 2000; Lejk & Wyvill, 2001b). Meanwhile, peer and self rating also depends on the function of the group work. In classroom instruction, teachers are usually interested in the efforts each member makes to the group work, such as attendance, participation, and communication with other group members. As an assessment tool, group marks should demonstrate the knowledge and skills of each student in relevant content area, thus rating of the contribution should be related to and interpreted as the academic contribution to the final product of group work. Only in that way could marks from group projects be used separately or combined with results from individual assessment tasks to report the achievement level of each student.

Being aware of the lack of individual accountability of group grades, teachers may choose to request students to hand in individual work after doing group

projects (e.g., Johnston & Miles, 2004). In that case, the assigned grade consists of two parts: one due to individual effort and the other due to group work activity. The group part still needs to be adjusted for the contribution differences.

The primary purpose of the present study was to evaluate the utility of the individualized scores derived by different methods in accounting for individual contributions. Although considerable research has been conducted on how to derive individualized scores from group projects, relatively little attention has been directed to the evaluation of these methods. As a result, it is unclear how accurately scores derived using different methods reflect the actual performance of students. In practice, faced with multiple options, classroom teachers may find it hard to choose one specific method. The three research questions for this study are:

1. Is it possible to assign a valid individualized score from doing group projects?
2. Is adjustment based on peer and self ratings helpful in grading group projects?
3. Which method should be recommended for adjusting the differential contributions to group projects?

To answer these questions, a Monte Carlo simulation was designed along with a real-data analysis. The major advantage of conducting a simulation study in this case is that student true contribution to the group work is known, thus the accuracy of the derived individual scores can be directly evaluated. Moreover, how factors such as rater and group effects affect each adjustment method can also be systematically examined. The data collected from actual group work supplements the simulation study by examining the effectiveness of different adjustment methods in a more realistic assessment condition.

METHODS TO ADJUST CONTRIBUTION DIFFERENCE

Four commonly used methods based on the peer and self ratings were studied. The first method was the autorating system method proposed by Brown (1995) (referred to as the within-group method hereafter). This approach requires each person to rate all group members (self included) on group contributions by using a single indicator. The relative contribution index is simply the average of all ratings for one person, or

$$c_{ik} = \sum_{j=1}^{N} r_{ijk} / N, \tag{1}$$

where c_{ik} is the contribution index for person i in group k; r_{ijk} is the rating of person i by rater j in group k; and N is the group size. A within-group weighting factor, w_{ikw} is the ratio of c_{ik} over the average contribution indices ($\bar{c}_k$) in group k, or

$$w_{ikw} = c_{ik} / \bar{c}_k, \tag{2}$$

Conceptually, this weight reflects how the contribution of one group member compares to the average group contribution. The individualized score for person i is the product of the within-group weight and the group mark, or

$$x_{ik} = w_{ikw} * x_k, \tag{3}$$

where x_k is the assigned group score.

The second method computes the weighting factor in the same manner as the first method, but only a proportion of the group grade is adjusted (Conway et al., 1993; Goldfinch, 1994). The rationale for this partial adjustment is that part of the group grade should directly count for the individual score regardless of the contribution level (Lejk & Wyvill, 1996). Suppose the percentage of the group grade to be adjusted is p. The individualized grade x_{ik} is derived as

$$x_{ik} = x_k * p + x_k * (100\% - p) * w_{ikw}, \tag{4}$$

where x_{ik} , x_k, and w_{ikw} are the same as defined in Equations 2 and 3. The exact value of p depends on the nature of the group project. This method is referred to as the partial-adjustment method hereafter.

The third method defines an expected proportion of contribution in deriving the weighting factor (Thompson, 1996, cited from Lejk & Wyvill, 1996). This method thus is referred to as the expected-contribution method hereafter. The expected contribution is simply the reciprocal of the group size. For example, in a group of three members, everyone is expected to contribute 1/3 to the final work. The individualized score is the group grade plus the adjustment of the deviance from the expected contribution, or

$$x_{ikw} = x_k + x_k (p_{ik} - p_{ek}), \tag{5}$$

where p_{ek} is the expected proportion. The term p_{ik} is the proportion that a group member has actually contributed. It may be computed as the ratio between the total ratings person i received (c_{ik}) and the total rating of all group members, or

$$p_{ik} = c_{ik} / \sum_{j=1}^{N} c_{ijk}. \quad (6)$$

where c_{ijk} is the total rating of each group member and N is the group size.

To boost motivation and involvement, teachers usually allow students to self-select their group members (Bacon, Stewart, & Silver, 1999). As a result, the overall ability in completing the group project probably varies across groups in practice. While students in stronger groups are more likely to obtain higher group scores, they may be more or less likely to get higher individualized scores based on the aforementioned adjustment methods. To make individualized scores from different groups comparable, an important question to address is whether students with equal contribution but in different groups actually receive the same individualized scores. To answer that question, a new method was proposed to adjust the between-group difference.

The new method utilizes a second weight to adjust for the between-group differences. The second weight compares students with equal within-group weight, or w_{ikw}. The between-group weight, w_{ikb}, is calculated as a ratio between the mean within-group adjusted score for students with equal contribution and the within-group adjusted scores for the student *i*, or

$$w_{ikb} = \sum_{m=1}^{M} x_{imw} /(M * x_{ikw}), \quad (7)$$

where M refers to the total number of student with equal w_{ikw} and x_{imw} is the within-group adjusted score for these students obtainable by Equation 3. A student's final individual group score, x_{ik} , is a product of three terms: the group score, the within-group weight, and the between-group weight.

$$x_{ik} = w_{ikw} * w_{ikb} * x_k. \quad (8)$$

For students whose within-group adjustment is lower than the average of all students at his or her contribution level, w_{ikb} will be larger than 1 and the final score will be adjusted upward. On the contrary, for students who received a high x_{ikw} due to a group effect, their final individualized scores will be lower than the within-group individual scores.

As peer and self ratings are related to the quality of group work, ratings from groups with different group scores are not directly comparable. For example, a rating of 3 for "better than most of the group members of contribution" corresponding to a group score 5 is not the same as a rating of 3 corresponding to a

group score of 1. The former simply implies more contribution to the assessed knowledge domain. Therefore, the between-group factor is derived based on groups with the same group score assigned by the instructor.

The expected-contribution method adjusts the group grade by adding a deviance from the expected contribution. Therefore, one's final individualized score will be influenced by the ratings of the contribution to only a limited degree. In contrast, the within-group method, the partial-adjustment method, and the between-group method all derive individualized scores based on the product of the group grade and the weighting factor (s). For those methods, one's individualized score could be greatly influenced by the contribution ratings. Moreover, these methods compare each group member's contribution to others in the group, making them somewhat robust to the rater or group effect. As an example, consider the friendship effect when one member deliberately inflates the contribution of all group members. As the average group rating will increase simultaneously with the individual ratings, the weighting factor for each group member may not be affected much.

Yet these three methods are not without fault. One can easily see that the scale of the derived individual scores can be quite different from that of the original group grade. In the extreme case where only one student in an *n*-member group actually made any contribution, the individualized score for that student would be *n* times larger than the group score. To alleviate this scale inflation, a constraint can be put on the maximum value of the within-group weighting factor (e.g., Kaufman, Felder & Fuller, 2000) or the derived score itself. However, what the maximum value should be is arbitrary.

METHODS

Research Design

A fully crossed design was implemented in the simulation study with the following factors: 3 (group sizes) × 2 (types of raters) × 2 (types of groups). Group size was examined at 2, 4, and 6 to emulate small, medium, and large groups in classroom settings. The quality of peer and self ratings was examined by a rater effect and a group effect. A rater effect is related to the capability of each member in working as an independent evaluator of peer and self performance. Less capable raters may be incapable of distinguishing between high and low contributors (i.e., the indiscrimination effect), excise unnecessary leniency or harshness, or rate toward the center or the extremes of the rating scale (Wolfe, Chiu, & Myford, 2000). Raters showing these effects were identified as less-capable raters and their percentage was examined at two levels: 0% and 50%. A group effect is related to how one group uses a rating criterion different from other groups. For

example, students in one group may intentionally inflate the contribution level of all members due to friendship. In the simulation, the group effect was created by setting 10% of the groups to reflect a friendship effect and another 10% to reflect an enmity effect. This resulted in the group variance representing about 20% of the total variance in the ratings, similar to what has been observed in real data (Zhang et al., 2008). Class size was fixed at 24 for all conditions to study a typical class size. For each testing condition, 200 replications were run.

Generating Self and Peer Ratings

Group ratings were generated using a two-parameter logistic rating model (Wolfe, 1997). This model factors in both the parameters from the rating task and from the individual rater to reflect a three-way interaction: the cognitive ability of students in completing the group project, their capability to rate each other's contribution, and the difficulty level of the task. This model is expressed as:

$$P(x \mid \theta, \gamma, \lambda, \delta, \tau) = \frac{\exp \sum_{j=0}^{x} \gamma_k (\theta_n - \delta_i - \lambda_k - \tau_j)}{\sum_{x=0}^{m} \exp \sum_{j=0}^{k} \gamma_k (\theta_n - \delta_i - \lambda_k - \tau_j)}, \quad (9)$$

where x refers to the assigned rating; θ_n represents the true ability to complete the project, or the true contribution of student n; δ_i refers to the difficulty level of the group project i. The term γ_k is the centrality index of rater k, which reflects the discriminating power of rater k. The term λ_k refers to the leniency level of rater k; and τ_j is the step difficulty for category j compared to category j-1.

Group ratings were simulated on a five-point scale ranging from 1 to 5 with 5 referring to the maximum contribution. The true contribution trait θ was assumed to follow a standard normal distribution in the population. The difficulty level for the group task was set at a medium level, or $\delta_i = 0$. The step difficulty of ratings was set at –3, –1, 1, and 3 so that the five categories on the rating scale would be clearly distinct. For standard raters, the centrality and leniency parameters were set at 1 and 0, respectively. For the less-discriminating raters, the centrality parameter, γ_k, was set at a lower level (i.e., 0.25 and 0.5). For lenient raters, the λ_k parameter was set –1 and –2 while for harsh raters, its value was at 1 and 2. To simulate the group effect, the above-generated ratings were artificially inflated or deflated for a subset of groups. To emulate the friendship effect, one point was added to all ratings in those groups. For the enmity effect, one point was deducted. For all cases, the original 1–5 rating scale was kept.

Generating Group Grades

Group grades were generated using the graded-response model (Samejima, 1969). This model takes a two-step approach in modifying how a group has responded to a polytomously scored task. The first step is to compute the conditional probability that group *i* will score the response category *j* and higher by the following function:

$$P_{ij}{}^{*}(\theta) = \frac{1}{1 + e^{-a_i(\theta_n - b_{ij})}}, \tag{10}$$

where p_{ij}* is the conditional probability, b_{ij} is the step difficulty, and all other terms share the same interpretation as defined in Equation 9. The conditional probability for the score category *j* is the difference between the conditional probability for the two adjacent categories:

$$P_{ij}(\theta) = P_{ij}{}^{*}(\theta) - P_{i(j+1)}{}^{*}(\theta), \tag{11}$$

In the above modeling, the discrimination parameter, a_i, was set at 1 and the four step difficulty parameters were fixed at –2, –1, 0, and 1 to generate responses with five categories. The θ term in the equation used the mean theta value of members in each group used in Equation 9. This average represented the collective contributions of all members in each group in doing the group project.

Indices in Measuring the Accuracy of the Individualized Scores

To evaluate the utility of the four methods under study, two statistics were calculated. The first one was the correlation between the true contribution trait and the derived individual group scores. A higher correlation implies that students with higher contributions tends to receive higher individualized scores, thus the method used to derive such a score shows more convergent validity. The second statistic was the root mean squared error (RMSE). It measures the absolute error in using the derived individualized scores to represent student true contribution. The group grade and the derived scores were standardized to be comparable to the scale of the simulated true contribution. The root mean square error (RMSE) was obtained by:

$$RMSE = \sqrt{\frac{\sum (x_{ik} - \theta)^2}{N}}, \tag{12}$$

where x_{ik} represents the standardized score and N is the sample size. As multiple replications were run, values reported in the Results section are the average of the correlation and RMSE values for each condition.

To investigate the impact of the controlled factors in the design, factorial ANOVA analyses were conducted using the correlation and RMSE values as the dependent variables. To meet the normality assumption, the correlation values were converted to z scores using the Fisher transformation (Fisher, 1915) and the RMSE was subjected to a log transformation. Group size, percentage of less capable raters, and percentage of non-standard groups were treated as random effects but the method factor was considered fixed.

RESULTS

Figure 1 presents the correlation between the true contribution and the derived individualized scores when no group effect is present and Figure 2 gives the corresponding RMSE statistics. Overall, the adjustment based on the peer and self ratings apparently improved the validity of group work grading. The individualized scores from all the four methods were better than the unadjusted group grade in reflecting the actual performance of students in completing the group project.

These two figures clearly show that the unadjusted group grade itself was a poor indicator of student ability. Compared to the four derived scores, its correlation was always the lowest, never higher than .4. Meanwhile, its RMSE value was always the highest. Not surprisingly, group size had a huge impact on the performance of this score. As groups became large, group grades were less informative

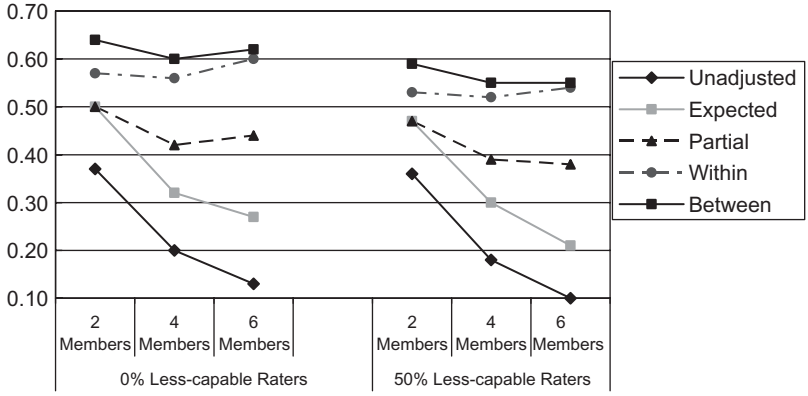

FIGURE 1 The correlation between the contribution-adjusted individualized scores and the true contribution, without group effect.

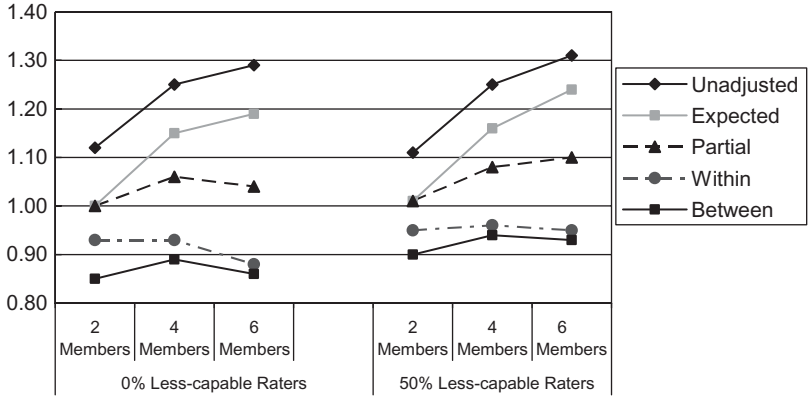

FIGURE 2 The RMSE values for the contribution-adjusted individualized scores, without group effect.

of any individual's performance. For groups with six members, the correlation level could be as low as .1. These findings confirm what many teachers and researchers (e.g., Kagan, 1995) have long believed: group grades are not suitable in reporting individual performance.

Among the four adjustment methods, the expected contribution method was the least promising. It followed the same trend as the unadjusted group grade in both Figures 1 and 2. Its correlation was considerably lower than that of other three methods. Moreover, its performance was very sensitive to group size. With the increase of group size, the performance of this method deteriorated sharply.

Adjusting part of the group grade by individual contribution gained no advantage over the adjustment of total grade. The individualized score from the partial-adjustment method demonstrated lower correlation but a higher RMSE than the within-group method for all examined conditions. However, the between-group adjustment did show additional gains over the within-group adjustment. The correlation was higher and the error was smaller. This trend was especially apparent for conditions with small groups and capable raters. When the group size increased to 6, the advantages diminished. For these three methods, the correlation remained higher than .5 even for large groups where the corresponding correlation for unadjusted group grade was very low.

The factorial ANOVA analyses of the correlation and RMSE values revealed significant interactions between the group size and the adjustment method (for r, F (8, 5,953 = 149.64, $p < .01$; for *RMSE*, F (8, 5,970) = 101.68, $p < .01$). As illustrated in the figures, this interaction indicated that as group size increased, the difference among the five methods became more pronounced. Conceptually, this implies the adjustment was more needed for larger groups. An interaction effect

was also found between the method factor and the percentage of less-capable raters (for *r*, F (4, 5,953) = 10.20, $p < .01$; for *RMSE*, F (4, 5,970) = 8.86, $p < .01$). The difference was larger among methods for conditions with all capable raters. As these interaction effects are all ordinal, a post hoc analysis was conducted on the main effect of the five scores. All the pairwise comparisons were significant, indicating that the differences observed in the figures reflected the statistically significant differences among the methods.

Figures 3 and 4 present the results from conditions where a group effect was present. The performance of the four methods was very similar to what was

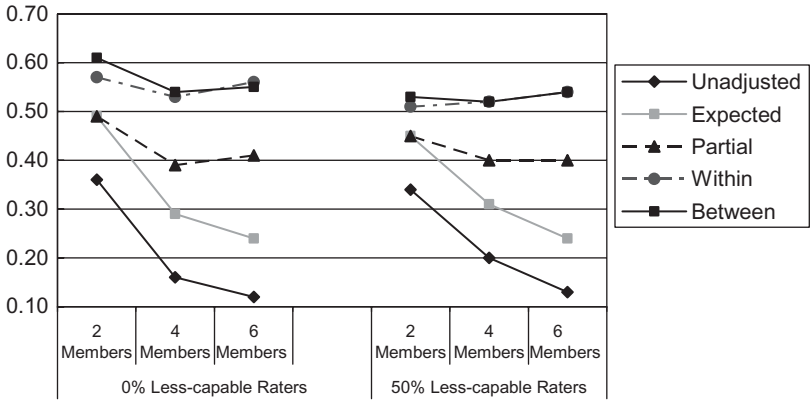

FIGURE 3 The correlation between the contribution-adjusted individualized scores and the true contribution, with group effect.

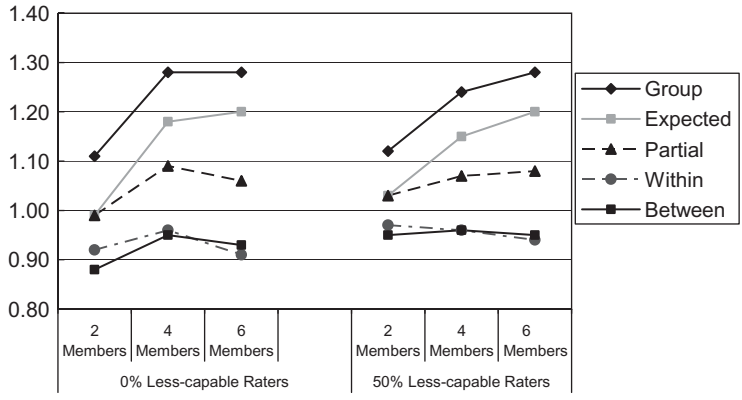

FIGURE 4 The RMSE values for the contribution-adjusted individualized scores, with group effect.

observed for conditions without the group effect. The derived scores worked much better than the unadjusted grade, especially for larger groups. The expected-contribution method was less viable than the other three adjustment methods. The interaction effects persisted between the method and the group size and between the method and the percentage of non-standard raters.

Several major differences were observed when there was a group effect. First, while the group grade itself was not affected by the group effect, the accuracy for the four derived scores decreased, as reflected by the lower correlations and higher RMSEs. Secondly, the advantage of the between-group method diminished for large groups. The hypothesis testing of the correlation and the RMSE values showed no significant differences between the between-group and within-group methods for these conditions. Finally, the disadvantage of the partial-adjustment method, while persistent at the group sizes of 2 and 4, also diminished at the group size of 6.

REAL DATA ANALYSIS

To investigate how the aforementioned contribution-adjustment methods can be applied, test data from a group work project was analyzed. The course was Principles of Management at a large university in the southeast United States. This course was taught by a permanent university lecturer who delivered lectures via university TV or Internet for two 50-minute periods per week. The students met in person in 30 sections, which were administered by teaching assistants (each TA covering six sections) for one 50-minute period per week. Each section had five teams that played a management simulation game. Simulation performance and a related team oral report accounted for 18.75% of the course grade.

The Comprehensive Assessment of Team-Member Effectiveness (www.catme.org) was administered in the paper-and-pencil form to solicit peer and self evaluation on group contribution from each team. TAs were trained to administer the instrument and used a written protocol for the administration. Students rated their own and their teammates' performance on the simulation game in these five areas of teamwork: contributing to the team's work, interacting with teammates, keeping the team on track, expecting quality, and having relevant knowledge skills and abilities. In each area, more detailed criteria were prescribed. Students conducted the ratings for each area on a 7-point Likert scale. The evaluations did not affect student grades and were not reported to the course instructor.

Ratings from one of the 30 sections were analyzed in this study. This section was chosen for the following two reasons. First, the reliability of the peer and self ratings was high. The dependability coefficient was .87. This statistics was computed by employing a random-effect nested design under the Generalizability Thoery (Cronbach, Gleser, Nanda, & Rajaratnam,1972). Specifically, it is the

person facet crossed with the rater facet, but both nested within the group facet (Zhang et al., 2008). Second, although all 30 sections had some peer and self ratings missing, this section had the fewest missing values.

The original contribution ratings and scores derived from applying various adjustment methods are given in Table 1. In the table, R1 to R4 refers to raters 1 to 4. As shown in the table, groups 2, 3, and 4 had some ratings missing. Adjustments for these groups were based on all the ratings available. The within weight is the weight computed for the within-group method and the expected weight for the expected-contribution method. The group grade is the grade assigned to each group by the TAs. Different from the aforementioned Monte Carlo simulation study, the true contribution of each student was unknown in this project. Instead, the course mark was used as a proxy. This mark was adjusted with the group grade of this section removed. Because no two groups received the same group grade, it was not possible to evaluate the between-group adjustment method by these data.

Table 2 gives the summary statistics of the derived individualized scores in comparison to the group grade. Clearly, the group grade without any adjustment would be a very poor indicator if used to reflect the actual performance of each student. It was almost uncorrelated with the final course grade. After adjusting for contribution differences, the correlation increased considerably for all the derived scores. Consistent with findings from the simulation study, the within-group method demonstrated the highest correlation while the expected-contribution method demonstrated the lowest. The RMSE statistic also showed the within-group method had the least absolute error.

The means of the three individualized scores were the same as the original group score. This is not surprising as each derived score came from the re-distribution of the original group score among group members. On the other hand, the larger standard deviation and wider range of the derived scores imply that adjusting contribution differences had increased the variability among student scores. In other words, instead of getting similar scores from the group project, students would receive quite different scores once their contribution was factored in.

This example also demonstrates the scale inflation issue discussed earlier. The maximum value for all the derived scores exceeded the maximum possible score for this project, which was 100. In practice, instructors would need to decide whether that was acceptable or a ceiling should be set instead.

DISCUSSION

Findings from this research shed some light on the use of group grades in classroom instruction. Concerning the original three research questions, both the simulation study and the real-data analysis signify that it is possible to use group work as an

TABLE 1
Peer Ratings, Weights, and the Derived Scores in the Example

Group	*Person*	*R1*	*R2*	*R3*	*R4*	*Group Grade*	*Within Weight*	*Expected Weight*	*Within Score*	*Partial Score*	*Expected Score*	*Course Mark*
1	1	6.91	5.88	6.97	7.00	85	1.00	.00	84.60	84.70	84.90	B
1	2	7.00	6.13	6.97	7.00	85	1.01	.00	85.70	85.50	85.20	A
1	3	6.91	5.97	6.97	7.00	85	1.00	.00	84.90	84.90	85.00	B
1	4	6.97	6.78	6.09	6.97	85	1.00	.00	84.80	84.80	84.90	C
2	5	6.94	5.84	–	–	78	.96	–.01	74.70	75.60	77.20	C
2	6	7.00	6.69	–	–	78	1.03	.01	80.10	79.50	78.50	A
2	7	6.53	6.66	–	–	78	.99	.00	77.10	77.40	77.80	A
2	8	7.00	6.69	–	–	78	1.03	.01	80.10	79.50	78.50	B
3	9	5.20	5.91	5.91	6.03	80	1.06	.01	84.60	83.50	81.20	A
3	10	5.47	5.91	5.91	6.03	80	1.07	.02	85.60	84.20	81.40	B
3	11	3.53	5.75	5.91	5.84	80	.97	–.01	77.20	77.90	79.30	B
3	12	6.00	3.80	5.22	4.72	80	.91	–.02	72.50	74.40	78.10	B
4	13	7.00	4.50	–	–	100	.97	–.01	97.10	98.00	99.00	A
4	14	7.00	6.19	–	–	100	1.11	.04	111.00	108.00	104.00	B
4	15	6.66	4.19	–	–	100	.92	–.03	91.60	94.40	97.20	B
5	16	7.00	6.38	–	–	93	1.37	.12	128.00	116.00	105.00	A
5	17	7.00	6.41	–	–	93	1.38	.13	128.00	116.00	105.00	B
5	18	1.25	1.19	–	–	93	.25	–.25	23.30	46.50	69.80	C

TABLE 2
Summary Statistics of the Group Grade and Derived Scores in the Example

Score	*Mean*	*Std*	*Minimum*	*Maximum*	*Correlation With the Course Mark*	*RMSE With the Course Mark*
Original group grade	86.17	8.19	78.00	100.00	–.01	1.28
Expected-contribution method	86.17	10.79	69.76	104.67	0.28	1.17
Partial-adjustment method	86.17	16.36	46.53	116.33	0.38	1.09
Within-group method	86.17	22.72	23.29	128.00	0.41	1.06

assessment tool to evaluate student learning. While assigning all group members the same score apparently does not reflect the true performance, methods based on the adjustment of individual contribution clearly have the potential to provide more valuable individualized scores. The observed medium-level correlation with the true contribution should be acceptable for the classroom assessment where results from multiple measures are often combined in reporting the final grade. In general, this research supports the use of group grades in classroom assessment with the condition that student contribution is accounted for.

Overall, methods relying on weighting factors by comparing student contributions to each other enjoy distinct advantages. These methods are less sensitive to group size. Although performance of all methods generally drops when groups get larger, the within-group and between-group methods still generate reasonable results even for large groups. For the classroom use, the between-group method should be recommended. It improves on the within-group method when group size is small.

This research focuses on the assessment function of the group project. It studies the adjustment of group grades by peer and self ratings among students. The four methods under study could easily be applied to situations where group contribution information is gathered from other sources, such as teacher observation. As the quality of teacher observation is probably higher than that of the students, the derived scores may enjoy higher validity than what observed here.

In general, classroom assessment is criterion-referenced. Teachers are less interested in knowing which students did better or worse than which students have not met the preset learning targets yet. As all the studied methods propose to derive the individualized scores based on comparing the contributions of group members, one assumption in using these methods is that adequate alignment has been established between the contribution ratings and the learning targets that the group project assesses. In other words, it is assumed that students who do better on the peer and self ratings of the contributions are more likely to meet the preset learning targets. This assumption is consistent with earlier findings by Loughry,

Ohland, and Moore (2007) that ratings are influenced by team-level norms regarding the definition of good team performance.

Doing group projects is a complex process. This study does not distinguish between self and peer rating in the simulation study. In reality, these two kinds of ratings may be quite different. In addition, while peer and self ratings are useful in adjusting group grades, they may cause discomfort in students as they perceive peer rating as criticizing their friends (William, 1992). Thus the effectiveness of these methods may well rely on the comfort level of students in conducting peer and self ratings (Gatfield, 1999). Peer grading may work best for students who actually enjoy working in groups and who accept this method of assessment (Stuart, 1994).

The present study focuses on the assessment role of group projects, which is how grades from group projects may be utilized to report individual student achievement. Accordingly, it has been argued that self and peer ratings should be related to the academic contribution to group work. However, this does not imply that the effort students put into group work is less important. While grading effort is not recommended for reporting academic achievements, effort is indispensable for attaining the instruction role of group projects. Apparently, students learn better from group work if they are more involved. A high-quality assessment framework not only aims to provide fair scores, but also encourages students to put in their best efforts. Once students realize that their contributions will be reasonably rewarded, their motivation, perception, and involvement in group projects are likely to improve (Johnston & Miles, 2004), which in turn will increase their overall learning from doing group work. Therefore, the appropriate accountability of group contributions will not only improve the validity of group grading as an assessment tool to report individual performance but also enhance the instructional function of group projects.

Findings from this research can be utilized to help teachers make better use of group projects. Although it is not realistic to ask most teachers to follow the Greek letters in the presented equations to compute different weights, the suggested methods can easily be programmed as a user-friendly program, an add-on to spreadsheet programs, or a behind-the-scenes heuristic in an automated system for collecting peer evaluation data (such as the one mentioned earlier, available at www.catme.org). Once teachers see the assigned grades for group projects as embodying individual contributions, and thus are valid in reflecting the performance of each individual student, they will have more confidence and flexibility in using group projects to serve their instructional purposes.

REFERENCES

Bacon, D. R., Stewart, K. A., & Silver, W. S. (1999). Lessons from the best and worst student team experiences: How a teacher can make the difference. *Journal of Management Education*, *23*, 5, 467–488.

Bagci Kilic, G., & Cakan, M. (2006). The analysis of the impact of individual weighting factor on individual scores. *Assessment and Evaluation in Higher Education, 31*, 6, 639–654.

Bartlett, R. L. (1995). A flip of the coin–a roll of the die: An answer to the free-rider problem in economic instruction. *Journal of Economic Education, 26*, 131–139.

Brown, R. W. (1995). Autorating: Getting individual marks from team marks and enhancing teamwork. *Proceeding from 1995 "ASEE/IEEE Frontiers in Education" Conference 2*, 3c2, 15–3c2, 18.

Conway, R., Kember, D., Sivan, A., & Wu, M. (1993). Peer assessment of an individual's contribution to a group project. *Assessment and Evaluation in Higher Education, 18*, 45–54.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

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

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika, 10*, 507–521.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education, 20*, 289–292.

Gatfield, T. (1999). Examining student satisfaction with group projects and peer assessment. *Assessment & Evaluation in Higher Education. 24*(4), 365–377.

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

Johnson, D., & Johnson, R. (1999). *Learning together and alone: Cooperative, competitive, and individualistic Learning* (5th ed.). Boston: Allyn and Bacon.

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

Johnston, L., & Miles, L. (2004). Assessing contributions to group assignments. *Assessment and Evaluation in Higher Education, 29*, 751–768.

Kagan, S. (1995). Group grades miss the mark. *Educational Leadership, 52*, 68–71.

Kaufman, D., Felder, R., & Fuller, H. (2000). Accounting for individual effort in a cooperative learning teams. *Journal of Engineering Education, 89*, 133–140.

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*, 78–94.

Lejk, M., & Wyvill, M. (1996). A survey of methods of deriving individual grades from group assessment. *Assessment and Evaluation in Higher Education, 21*, 267–280.

Lejk, M., & Wyvill, M. (2001a). The effect of the 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*, 551–561.

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

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*, 569–577.

Loughry, M. L., Ohland, M. W., &. Moore, D. D. (2007). Development of a theory-based assessment of team member effectiveness. *Educational and Psychological Measurement, 6*, 505–524.

Lourdusamy, A., & Divaharan, S. (2000). Peer assessment in higher education: students' perceptions and its reliability. *Journal of Applied Research in Education, 4*, 81–93.

Samejima, F. (1969). Estimation of latent ability using a response pattern of grade scores. *Psychometrika, Monograph Supplement, 34*(4), 100–114.

Sherman, L. W., & Thomas, M. (1986). Mathematics achievement in cooperative versus individualistic goal-structured high school classrooms. *Journal of Educational Research, 79*, 169–172.

Slavin, R. E. (1995). *Cooperative learning*. Boston: Allyn and Bacon.

Springer, L., Stanne, M., & Donovan, S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research*, *69*, 21–51

Stefanou, S., Hood, L., & Stefanou, C. (2001). Feedback and change: Assessment of individual contributions within collaborative activities in the higher education classroom. *Journal on Excellence in College Teaching*, *12*, 77–91.

Stuart, A. M. (1994). Effects of group grading on cooperation and achievement in two fourth-grade math classes. *Elementary School Journal*, *95*, 11–21.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes and outcomes. *Educational Evaluation and Policy Analysis*, *17*, 239–261.

William, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education*, *17*, 45–58.

Wolfe, E. W. (1997). *A two-parameter logistic rater model*. Paper presented a the annual meeting of the Florida Educational Research Association, Orlando, FL.

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects in simulated data with a multi-faceted Rasch rating scale model. In M. Wilson & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (vol. 5, pp. 147–164). Stamford, CT: Ablex Publishing Co.

Zhang, B., Johnson, L., & Bagci Kilic, G. (2008). Assessing the reliability of self and peer rating in student group work. *Assessment and Evaluation in Higher Education, 33*(3), 329–340.