Educational Research and Development Journal
Fall 2022, Vol. 25, No. 2, pp. 20–31

Exploring the Utility of Peer and Self Assessments in Grading Group Projects for Urban Middle School Students

Bo Zhang
Lucas Jackson
University of Wisconsin – Milwaukee

Abstract

While group projects are popular in middle school classrooms, limited research has been conducted on how to grade them. Teachers usually assign the same score to all group members, which hardly reflects the achievement level of any student. This study explores the feasibility of utilizing peer and self-assessments for grading group projects for urban middle school students. The sample includes 45 sixth graders from a public middle school in a large Midwestern metropolitan area. Students completed a group project in their English class. Results show that self and peer ratings have limited value in grading group projects in that students tend to inflate their contribution to group projects but peer ratings are more promising. Moreover, students enjoy doing group projects and prefer working with friends and others at their own levels.

Group projects are a popular instructional tool at almost all levels of education. While systematic study has yet to be conducted on the effect of group projects on student learning, it is generally believed that they are especially effective in cultivating a core set of skills that are hard to be taught by individual work, such as communication skills and responsibility taking skills (Assiter, 2017; Steensels et al., 2006). Group work also plays an important role in many other aspects of learning activities. For instance, group work is an indispensable component of team-based learning (Sweet & Michaelsen, 2007) and team-based learning has been shown to have a moderate advantage on student learning over some other instruction methods (Liu & Beaujean, 2017).

In classroom teaching, it is important to differentiate using group projects as an instruction tool from a measurement tool (Zhang, Johnston, & Gulsen, 2008). While these two roles are inseparable at times, they are not the same. Group projects enable students to apply knowledge to reallife-like problems. They are also adaptive to a range of performance levels which are usually inherent in most classes. Group projects have served as a major driving force in cooperative learning. Overall, group projects are a powerful and effective instructional tool.

Unlike the instructional function, the measurement function of group projects has been widely scrutinized. First, it is actually very hard to use group projects to measure student academic achievement. A simplistic but often adopted way is to assign the same grade to all group members. This non-discriminating scoring encourages the so-called free-rider effect in that students who have made little contribution to group work receive the same credit as those who have made major contributions (Kerr & Bruun, 1983). As many students may have already suffered from low motivation in doing group work (Kerr, 1983), some researchers have called for a ban on grading group projects (Kagan, 1995). The focus of this study is to further explore this measurement function of group projects.

Ideally, like how individual assignments are marked, teachers could assign a grade to each student based on their individual performance in a group project. To achieve that, teachers would

need to take into account each student's contribution to the group project. In practice, teachers generally don't have that group contribution information. For college students, studies on grading group work have pointed to one possible alternative: using peer and self-assessments from students (e.g., Lejk & Wyvill, 2001; Lejk, Wyvill & Farrow, 1996; Johnston & Miles, 2004; Zhang, Johnston, & Gulsen, 2008). Peer rating in groupwork assessment refers to group members rating each other's contribution to a group project. Students can also evaluate their own contribution to group work. Group contribution may be collected by a holistic or analytic approach. The holistic approach asks for an impressionistic score that reflects the inherent quality of the whole group work. The analytic approach, on the other hand, divides the whole group work into meaningful components, such as process and product, then evaluates each characteristic separately.

While the contribution information can be collected, the challenge is whether it is rigorous enough for evaluative purposes. In other words, can the ratings be trusted? Statistically, answering this question is equivalent to evaluating the reliability and validity of those ratings. In this context, high rater reliability requires that group members give consistent ratings to each member and high validity refers to that ratings accurately reflect actual contributions.

To evaluate the reliability of peer and self-ratings to group work, Generalizability Theory (G-theory) (Shavelson & Webb, 1991) has been proposed (Zhang, et al., 2008; Ohland & Layton, 2000). Compared to traditional measures (e.g., correlation coefficient and the Kappa statistic), a G-theory study is more flexible and can accommodate various research designs. Moreover, as it decomposes measurement error into multiple sources, such as those due to rater and group, actions can be taken more easily to increase reliability.

A common practice in conducting group projects is that students in each group rate all members in the group. In this case, the G-theory research design will have two facets, which are rater and group. More specifically, the G-study design is rater facet is crossed with person facet and both are nested within group facet, as illustrated in Figure 1. In the graph, the person effect (p) indicates the amount of variance in the student scores that is due to the actual contribution difference, hence, the larger the person variance is, the higher the rater reliability will be. The rater effect (r) indicates the amount of variance associated with the difference criteria (e.g., leniency and discrimination) that raters use. The group effect (g) represents the variance of student scores due to the unique formation of the groups, such as friendship effect. The last two effects are all considered measurement error, which also includes a two-way interaction term (pxr) that is undistinguishable from the error from other non-specified sources (e). Once variance components are estimated, a generalizability coefficient can be calculated. While the appropriate degree of reliability depends on the purpose of measurement, as a general guideline, the commonly accepted lower limit for reliability has been suggested at 0.7 (Nunnally, 1978) for research purpose.

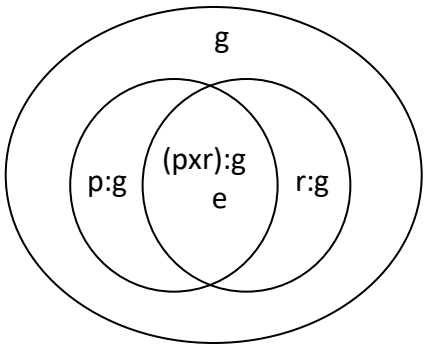

Figure 1. G-study design

For the validity of the peer and self-assessment scores, one common way is to examine the convergent validity that compares student rating to teacher rating. If there is no significant difference between them or these two types of ratings are highly correlated, one could conclude that student rating is as effective as teacher rating.

For college students, research has shown a high degree of reliability and validity for peer ratings in group work (Falchikov & Boud, 1989; Ohland & Layton, 2000; Zhang et al., 2008). Lejk, and Wyvill (2001) compared the holistic and analytic ratings and found that the holistic assessment resulted in higher inter-rater reliability. Ohland and Layton (2000) compared two peer assessment tools; one that focuses on deliverables such as quality of presentation and written work, and the other that focuses on the characteristics of group work. They concluded that focusing on identified behavioral characteristics of good teamwork improved the reliability of the peer assessment tool. Another possible reason for this relatively high reliability and validity is that college students have considerable experience with peer assessment tools and procedures (Steensels, et al., 2006).

What is unclear is whether other student populations can use peer and self-ratings to evaluate group contribution as well. While research on their value on group contribution has been mainly conducted in higher education, research on peer assessment on student learning points to its potential value for primary and secondary students as well. Double, McGrane, and Hopfenbeck (2020) showed that the effect of peer assessment on student learning does not seem to vary by educational levels. In their study of middle school students, Sadler & Good (2006) also found that peer ratings are valuable. They are highly correlated with teacher ratings with the correlation coefficients above the .9 level.

Of particular interest to the current study is urban middle school students. College students have passed through the formal stage of development. Their emotional maturity allows them to separate emotional judgments from objective contribution (Atputhasamy & Divaharan, 2002). Adolescents like middle school students are at a different stage of development. While working with peers can promote intellectual advances (Wentzel & Caldwell, 1997), the combination of a growing need for peer associations, increased self-awareness, and the introduction of social structures also create a unique situation for cooperative learning activities such as group work. Actually, the volatile nature of cognitive development and peer relationships not only have tremendous impacts for group work itself but may threaten the validity of using it for assessment (Wentzel & Caldwell, 1997; Rubin et al, 2006).

Regarding peer assessment, frequent changes in attitude and in peer relationship may contribute to students rating each other inconsistently (Lindblom-ylänne, Pihlajamäki & Kotkas, 2006). Ratings may be more about subjective social dynamics than about objective academic criteria. Additionally, peer interactions are volatile at this stage. Distinguishing among different types of peer relationships can prove a challenge for adolescents, making the educational task of rating a friend based on non-friend qualities very difficult. In other words, an overreliance on the importance of emotional social relationships can outweigh the longer-term importance of accurately assessing one's peers for academic gains (Brown & Larson, 2009). These developmental changes may invalidate the strategies and findings at the university level for adolescents. Additionally, middle school students may also have less experience interacting with their peers in such logical, objective activities as grading each other on academic performance, as not much peer rating has been conducted at the primary and secondary education levels than at college (Double, McGrane, & Hopfenbeck, 2020).

Students in urban middle school present another layer of challenges for research in the context of group work. Urban school districts deal with an increasing number of students needing special education services. In addition, suspension, truancy, and mobility rates of urban students present difficulties in gauging consistent measures in work over extended period of like, such as group projects. Peer relationships in schools can reflect external behaviors and environmental factors, which differ significantly from urban to suburban schools (Hannaway & Talbert, 1993). Hence, these peer interactions can often influence the opportunity for reliable and valid evaluation of an individual's contribution to social activities, including group work.

Research Questions

Driving the current study is the concept that adolescent development impacts group work in such a way that middle school students in urban schools face unique challenges in using peer and self-assessment tools. The current research aims to shed light on the following two questions:

1. Are student peer assessment and self-assessment reliable enough for evaluating group contributions in middle school classrooms?
2. Are student peer assessment and self-assessment valid enough for measuring individual contribution to group projects?

Rater reliability will be addressed by comparing the scores assigned by group members to each individual on group contribution. Due to the stronger group dynamics and a lack of practice using the assessment tools for middle school students, rater reliability is expected to be lower than that for college students. The validity question will be addressed by comparing the ratings from students to those from a teacher. It is hypothesized that validity of peer assessments will also be lower.

Method

Subjects

Forty-five sixth grade students were studied from an urban charter middle school in the Midwest. The students (25 female and 20 male) were observed in an English class, which they attended for one hour every day. The sample consisted of 100% African American students with 100% receiving free-and-reduced lunch. Roughly 30% of the students were labeled as receiving special education services. During the week long study, attendance was approximately 85%, with four students being suspended for one or more days.

Measures

Peer assessment (Appendix A) was adapted from a tool used in Ohland and Layton (2000), which has been shown to have acceptable rater reliability for college students. This instrument measures two aspects of student contribution: group contribution (questions 1, 7, and 8) and quality of work (questions 2, 3, 4, 5, 6, 9, and 10). Students conducted both self and peer assessments. Peer ratings were not shared with each other.

To measure student attitude towards group work, a survey was administered before and after the project. The pre-survey tapped into the feelings of working in groups as well as group relationships and dynamics. Some findings were used in forming groups. The post survey focused more on the experience of working on the group project. Survey questions are presented in Table

4. These surveys were not intended for systematically measuring the satisfaction level. Instead, they were used for collecting information on students' general attitude towards group work. No reliability or validity study was conducted on these surveys.

Procedure

The English class assignment consisted of a culminating project after the class has finished reading a novel required by the school curriculum. Students were given the task of working in groups to create brochures, with the intention that the project would allow students to demonstrate an understanding of the book and the application of the book's themes to real-world settings. The classroom teacher indicated that she had previously used group work as a means of evaluating learning in her class. In addition, students were familiar with the four-point rubric used in the general school curriculum, as well as the final assessment of the project. While the students had not previously used peer assessment tools in class, they had been exposed to self-assessment on individual projects.

The pre attitude survey was conducted before the group work started. The survey result indicated that students preferred to form their own groups and the ideal size was three. In practice, the teacher helped students organize their own groups. Although self-selection of groups may lower rater reliability (Steensels, et al, 2006), the teacher believed that random grouping or group assignment by the teacher may lead to withdrawal or less enthusiastic behaviors. Students had four days to complete the project and present it on the fifth day. After students had presented the project, they used the single-form assessment tool (Appendix A) to complete their self-assessment and a peer assessment. Their final task for the project was to complete the post attitude survey drawing on their experience with the group project. All group activities were videotaped. The teacher then used the same assessment tool to grade each student's contribution to the group work.

Estimation of Reliability and Validity

The reliability of the ratings was evaluated by applying the G-theory design as described in Figure 1. To reiterate, the design was rater crossed with person and both nested within group. The validity of the ratings was evaluated by the convergent validity between peer and self-ratings and teacher rating. Specifically, paired sample t-tests were conducted on the three types of ratings. No significant difference from teacher rating would indicate the validity of student ratings.

Control of Type I Error

One statistical issue with these multiple comparisons is the potential inflation of Type I error rate at the experiment-level. In other words, as the number of comparisons increases, so will the chance of finding a false positive be. One traditional way to control such a risk is to use the Bonferoni adjustment (Dunn, 1961), which simply divides the nominal Type I error rate by the number of comparisons. While the Bonferroni adjustment always controls the error rate at the experiment level, it can be over-conservative. In this study, as sample size is not large and lack of power is a concern, the false discovery rate (FDR) was used instead. Compared to the Bonferoni approach, this statistic provides more power while still controls the Type I error rate (Benjamini & Hochberg, 1995). Type I error rate was set at the .05 level.

Results

Table 1 gives the descriptive statistics for the peer and self-assessments. Rank order of these scores is self-rating > peer + self-rating > peer rating >teacher's rating. As teacher's rating is treated as the actual contribution level, this table indicates that self-rating and peer rating may have been inflated. Meanwhile, the larger standard deviations for the teacher and peer's ratings indicate that self-ratings are less discriminating.

Table 1.

Mean and Standard Deviation of Ratings by Self, Peer and Teacher

	N	Mean	SD
Teacher's Rating	44	2.84	.83
Peer + Self-Ratings	41	3.19	.60
Peer Ratings	44	3.00	.83
Self-Rating	41	3.37	.61

Reliability of Peer and Self Ratings

Table 2 presents the results on the reliability of peer and self-ratings. Values in the brackets are the percentages of variance each component accounts for. The first important finding is that reliability is very low when both self and peer ratings are considered. The g-coefficient as well and the d-coefficients is around .5, indicating that raters do not agree with each other in rating group member's contribution. In terms of the variance estimation, the person component accounts for only 17% of the total variance, while 49% of the variance is due to error. The 0% rater variance does not mean raters are perfectly consistent. The actual cause was a negative variance estimate. This anomaly could be due to small sample size or model data misfit, which results in the within-group variance being larger than the between-group variance (Institute, S. A. S., 2008; Briesch et al, 2014; Shavelson & Webb, 1991). In practice, negative variance is usually treated as zero. Secondly, when self-rating is removed, rater reliability for peer rating improves considerably to an acceptable level, pointing to the possibility that self-rating differs from peer rating.

Table 2.

Variance Estimates of Peer and Self Ratings

Rating	Person	Rater	Group	Error	Reliability Coefficient
Peer + Self Ratings	0.11 (17%)	0.00 (0%)	0.23 (34%)	0.33 (49%)	.48
Peer Ratings Only	0.35 (41%)	0.16 (19%)	0.29 (34%)	0.05 (6%)	.75

Validity of Self and Peer Assessments

As shown in Table 3 and Figure 2, using the 0.05 alpha level, whenever self-assessment is included, student rating is significantly different, actually higher than, the teacher's and peer's. This implies that the validity of self-assessment is quite low. With regard to the peer assessment, it is not significantly different from teacher rating. That is to say, the average of peer ratings for each student is similar to that from the teacher, which is certainly encouraging. Still, cautions should be taken in interpreting this result. As depicted in the last plot in Figure 2, the difference in that comparison still looks somehow systematic.

Table 3.

Paired t-test Comparing Self and Peer Ratings to Teacher Rating

Comparison	Mean* Difference	t	df	Original p-value	FDR p-value
Self + Peer vs. Teacher	.28	3.23	40	.002	.004
Self vs. Teacher	.46	4.65	40	.001	.004
Peer vs. Teacher	.16	1.60	43	.116	.116
Self vs. Peer	.29	2.47	40	.018	.024

* Due to missing data, some of these values are different from Table 1.

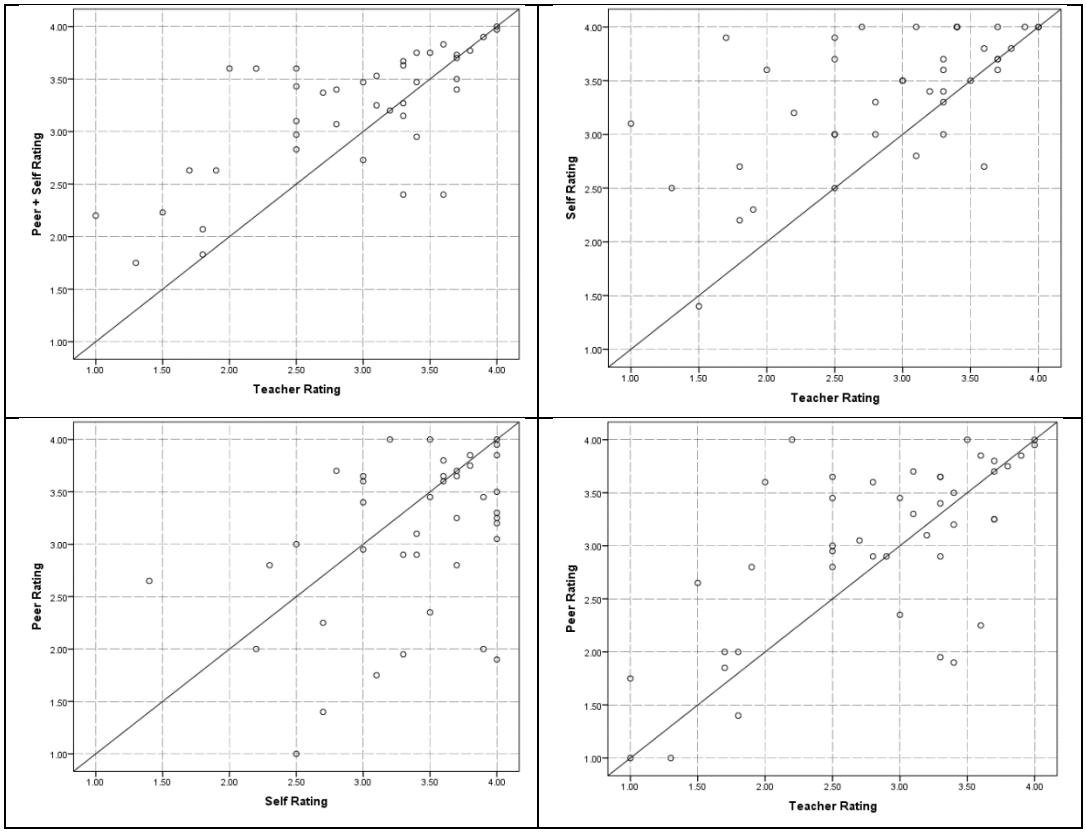

Figure 2. Relationship between self, peer, and teacher ratings.

Attitude Towards Group Work

Results on the pre and post surveys on student attitude towards group works are presented in Table 4. Overall, most students liked group work. Students were more inclined to forming their own groups, which was endorsed by the teacher in practice. In choosing who to work with in the group, students' preference clearly went to friends. Students also seemed to prefer peers at their levels. While over 80 percent students were willing to do their share of work, fewer people actually achieved that level. Also, students preferred to have more than one role in group work.

Table 4.

Pre and Post Student Attitude Towards Group Work Survey Results

Pre-survey Questions	Percentage of Endorsement		Post-survey Questions
	Pre Survey	Post Survey	
1. I like to work in groups	82	78	I liked working in a group
2. I would like to make my own group	87	42	I made my own group
3. When I work in groups, I like to work with my friends	82	78	When I worked in my group, I worked with my friends
4. When I work in groups, I like to work with the smartest students in the class	67	n/a*	
5. When I work in groups, I like to work with students who are at my level	76	n/a*	
6. I like to be the leader in groups	73	36	I was the leader in my group
7. I am willing to do my share of the work in a group	82	68	I did my share of the work in my group
8. I am willing to do more than my share in a group	62	61	I did more than my share in the group
9. I like when I have one job to do	62	n/a*	
10. I like when I can help with lots of jobs	78	59	I had more than one job in my group

* not asked

Discussion

For a classroom-based study like this, there are limitations. First, no intervention was built into the research design, making the study mostly observational. Second, the sample is quite homogeneous. Coupled with relatively small sample size, cautions should be taken to generalize the findings. As students' motivation to do group work can be quite low (Kerr, 1983), an emphasis was put on student engagement in group work, hence students were allowed to choose their own groups, which unfortunately may have resulted in a relatively large group effect. Finally, the teacher rating was treated as a measure of the actual student contribution to group work. As it is hard to capture group interaction, that rating itself might have considerable measurement error, which in turn, may have caused under- or over-estimation of the validity of student ratings.

Grading group projects poses a unique challenge to teachers as they usually don't have sufficient information on group contributions. This study explores whether middle school students can provide the necessary information on group contributions. Both the reliability and validity analyses clearly signal that self-assessment is problematic. Its reliability and validity are both unacceptably low. They tend to rate their own contribution very differently from that of their peers. On the other hand, peer rating is more promising. Its reliability is higher and it also shows more validity. Overall, these results are consistent with findings from previous research, such as De Grez, Valcke and Roozen (2012). In their study of presentation skill of college freshmen, they also showed that self-assessment scores were inflated. Similar to the qualitative findings of the Atputhasamy & Divaharan (2002), this study also demonstrates that students enjoyed doing group projects. This again attests that group projects are able to deliver developmentally appropriate learning opportunities.

To advance the research along this line and to address the above limitation, the most important future work should be to provide training on peer and self-ratings, especially on self-rating. Future studies can also explore the best way to assign group membership that will help reduce the group effect detected in this study but still not hurt group engagement. This may include studying such factors as group size, assignment method (e.g., self-self vs. random), and group composition (similar levels vs. diverse levels). Finally, future work should also study a more diverse population and use a larger sample size.

Conclusions and Implications

The ultimate goal for research along this line is to define the best practice in grading group projects for middle school students. To our best knowledge, research in this area for middle school classrooms has been extremely scarce. The current study represents the first, albeit important, step in the long journey. In a sense, it surveys the status quo situation in a typical urban middle school classroom where a group project is assigned.

This study shows that students tend to inflate their own contribution to group work while provide more valid evaluation of other student's contribution. These findings have clear implications for classroom practice. If peer and self-ratings are to be used, schools may consider providing training on these specific skills to students. They may also want to provide professional development on this for teachers so that they can help students better. Once students acquire the necessary skills for peer and self-ratings, it will not only help them with evaluating group contribution but may improve their overall learning experience, as peer and self-ratings are important for many other aspects of learning too.

References

Atputhasamy, L., & Divaharan, S. (2002). An attempt to enhance the quality of cooperative learning through peer assessment. *Journal of Educational Enquiry*. 3, No. 2: 72–83.

Assiter, A. (2017). *Transferable skills in higher education*. Routledge.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B 57*, 289–300.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of school psychology*, *52*(1), 13–35.

Brown, B. B., & Larson, J. (2009). Peer Relationships in Adolescence. Handbook of Adolescent Psychology. Chapter 3 Peer Relationships in Adolescence. 74–103.

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education*, *13*(2), 129–142.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, *56(293),* 52–64.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32(2)*, 481–509.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of educational research*, *59*(4), 395–430.

Hannaway, J., & Talbert, J. E. (1993). Bringing context into effective schools research: Urban-suburban differences. *Educational Administration Quarterly*, *29*(2), 164–186.

Institute, S. A. S. (2008). SAS/STAT 9.2 user's guide. Page 7493

Johnston, L., & Miles, L. (2004). Assessing contributions to group assignments. *Assessment & evaluation in higher education*, *29*(6), 751–768.

Kagan, S. (1995). Group Grades Miss the Mark. *Educational Leadership*, *52*(8), 68–71.

Kerr, N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of personality and social psychology, 45(4)*, 819.

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

Lejk, M., & Wyvill, M. (2001). Peer Assessment of Contributions to a Group Project: a comparison of holistic and category-based approaches. *Assessment & Evaluation in Higher Education*. 26, No. 1: 61–72

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

Lindblom-ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer-and teacher-assessment of student essays. *Active Learning in Higher Education*, *7*(1), 51–62.

Liu, S. N. C., & Beaujean, A. A. (2017). The effectiveness of team-based learning on academic outcomes: A meta-analysis. *Scholarship of teaching and learning in psychology, 3(1),* 1–14.

Nunnally, J.C. (1978) *Psychometric theory. 2nd Edition*, McGraw-Hill, New York.

Ohland, M. W., & Layton, R. A. (2000). Comparing the Reliability of Two Peer Evaluation Instruments. in proc. *2000 ASEE Annual Conf.*, St. Louis.

Rubin, Kenneth H., Stephen A. Erath, Julie C. Wojslawowicz, & Allison, A. Buskirk. (2006). "Chapter 12 Peer Relationships, Child Development, and Adjustment: A Developmental Psychopathology Perspective. "*Developmental Psychopathology, Volume 1, Theory and Method, 2nd Edition*. By Jeffrey G. Parker. 2nd ed. Vol. 1. Hoboken, NJ: John Wiley and Sons, 2006. 419–493.

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational assessment, 11(1)*, 1–31.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Prime*. Sage Publications.

Steensels, C., Leemens, L., Buelens, H., Laga, E., Lecoutere, A., Laekeman, G., & Simoens, S. (2006). Peer assessment: A valuable tool to differentiate between student contributions to group work?. *Pharmacy Education*, *6*.

Sweet, M., & Michaelsen, L. K. (2007). How group dynamics research can inform the theory and practice of postsecondary small group learning. *Educational Psychology Review, 19*, 31–47.

Wentzel, K. R., & Caldwell, K. (1997). Friendship, Peer Acceptance, and Group Membership: Relations t Academic Achievement in Middle School. *Child Development*. 68, No. 6: 1198–1209.

Zhang, B., Johnston, L., & Gulsen, B. K. (2008). Assessing the reliability of self- and peer rating in student group work. *Assessment & Evaluation in Higher Education*. 33, No. 3: 329–340.

Appendix A

Group Participation Sheet

Name_____

Block #_____

Date_____

Instructions:
Write the names of your group members in the space above.

1. For each behavior listed below, please give a number using the 1-4 scale to describe each person's contribution to the group project: **4 = Excellent, 3 = Good, 2 = Basic, and 1 = Minimal**.
2. This sheet will be used to assign a final project grade for the members in your group.
3. Your answers will be kept secret.

Group Member->	#1	#2	#3	#4
Write your name in space #1 and the name of other members in #2, #3, and #4.				
1. Attends class daily				
2. Participates in discussions				
3. Takes turns talking				
4. Listens carefully to others				
5. Takes group job seriously				
6. Accepts ideas from the rest of the group				
7. Quality of work				
8. Completes work on time				
9. Respects other group members				
10. Helps other group members				