VALIDITY AND RATER RELIABILITY OF PEER AND SELF ASSESSMENTS FOR URBAN MIDDLE SCHOOL STUDENTS

by

Lucas Jackson

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Educational Psychology

at

The University of Wisconsin – Milwaukee

August 2014

ABSTRACT

VALIDITY AND RATER RELIABILITY OF PEER AND SELF ASSESSMENTS FOR URBAN MIDDLE SCHOOL STUDENTS

by

Lucas Jackson

The University of Wisconsin – Milwaukee, 2014
Under the Supervision of Professor Bo Zhang

This project studied the validity and reliability of self and peer assessments used for group work. The targeted population is middle school students in urban schools. The sample includes 45 sixth graders selected from a public middle school in a large Midwestern metropolitan area. The students worked in groups to complete a classroom project. Self and peer assessment forms were used to rate each member's contribution to the group work. A Generalizability Theory design was used to evaluate the reliability of self and peer assessments. The validity of student ratings was assessed by comparing them to those assigned by the teacher. Results of this study indicate that peer ratings were not significantly different from teacher scores, but self-assessment scores were significantly higher than teacher scores. Acceptable rater reliability was achieved only when self-assessment scores were excluded.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

A plethora of studies have demonstrated the usefulness of cooperative learning at all levels of education. However, a debate still exists over the use of group work as a means to evaluate learning. Many tools and strategies are applied to evaluate the effectiveness of group work as assessment; teacher and student rubrics, self-assessments, and peer-assessments are used regularly. While the effectiveness of peer and self assessments have been well-documented in higher education, few studies have been conducted on them for middle school students, where group work is frequently used as a form of learning as well as an assessment tool. The current study looks to fill this gap by evaluating the validity and reliability of peer and self-assessments involved in a cooperative learning project.

The current study observed a sample of 6th grade English students, in a public charter middle school in an urban setting. These students, predominately low-income minority students, represent a segment of the public school population generally overlooked in studies of the effects of widely accepted pedagogical tools (Duncan-Andrade & Morrell, 2008). It is believed that if cooperative learning strategies are to be frequently used with this population, a deeper understanding of the evaluation tools used to assess learning is necessary, as is the way in which students use the tools. Yet, the volatile nature of cognitive development and peer relationships in middle school students raises questions as to the reliability and validity of the use of certain types of assessment (Wentzel & Caldwell, 1997). Frequent changes in attitude, as well as in peer relationships, may contribute to students rating each other inconsistently on assessment tools or rating others based on popularity within peer groups (Lindblom-ylänne,

Pihlajamäki & Kotkas, 2006). Furthermore, a lack of understanding of the relationship between group work and academic achievement may also impact the ability students have to assess one another (Atputhasamy & Divaharan, 2002).

This particular sample of students presents a challenge for research, not only in the context of group work, but in all areas related to social and educational activities. Urban school districts deal with an increasing number of students that have been identified as needing special education services. In addition, suspension, truancy, and mobility rates of urban students present difficulties in gaging consistent measures in longitudinal studies. Peer relationships in schools can reflect external behaviors and environmental factors, which differ significantly from urban to suburban schools (Hannaway & Talbert, 1993). Thus, these peer interactions can often influence the opportunity for reliable and valid evaluation of an individual's contribution to social activities, including group work.

Group Work

Group work is utilized in many avenues of education. It is generally accepted that group work is beneficial to students in terms of instruction and learning. Group work is viewed as an instrument that enables students to develop a core set of skills that have been labeled in the literature as "transferable skills"(Steensels, Leemans, Buelens, Laga, Lecoutere, Laekeman, Simoens, 2006). However, group work as a means of assessing cooperative learning has challenges as group contributions, rater and group effects, and developmental issues can all impact the evaluation process (Summers & Volet, 2009). In addition, teachers rarely have the time or ability to identify contributions of individual

students in cooperative learning environments. There are many proposed solutions to this issue; perhaps the most well-known and widely used is peer assessment. Peer assessments generally take one of two forms; holistic or category-based evaluations (Lejk & Wyvill, 2001). After completing a cooperative learning project, students indicate the contribution of their classmates and/or the overall quality of their peers' work. Many studies have been conducted to examine the usefulness of peer assessment in terms of evaluating contribution (Lejk & Wyvill, 2001; Steensels et al, 2006). However, the vast majority of these studies look at college students. There are obvious benefits in using higher education samples, such as convenience, diversity, motivation, and maturity. But characteristics such as logical understanding and emotional maturity do not apply to primary and secondary school students (Piaget, 1972). Therefore, whether or not peer assessment can be used as an evaluation tool for individual contributions in group work during primary and secondary education remains unaddressed.

Adolescent Development

Adolescence is characterized by a change in relationships, moving away from parental control towards peer-centered interactions (Brown & Larson, 2009). As constructivists such as Jean Piaget and Lev Vygotsky note, adolescence is an important time for social growth as peer relationships help to foster cognitive gains (Rubin et al, 2006). This stage, coined *formal operations* by Piaget, occurs around the age of 12, when American students are entering middle school. Much of the changes characterized by this stage of development involve the acquisition and use of logical reasoning (Piaget, 1972).

The growing importance of peers in the life of adolescents supports the incorporation of cooperative learning in schools as working with peers promotes intellectual advances (Wentzel & Caldwell, 1997). However, the formation of cliques and the growing awareness of self can lead to newly experienced emotions, such as jealousy. The combination of a growing need for peer associations, increased self-awareness through formal operations, and the introduction of social structures such as gossip create a unique situation for cooperative learning activities. This tumultuous time has tremendous impacts for group work studies (Rubin et al, 2006).

Unlike college students who have passed through the formal operations stage of development, have emotional maturity and are able to separate emotional judgments from objective contribution (Atputhasamy & Divaharan, 2002), adolescents have far less experience interacting with their peers in situations categorized through logical, objective activities. When students are expected to rate their peers in any sort of activity, those ratings are often related more to subjective social dynamics than objective academic criteria. In addition, peer interactions are volatile. On a daily or even hourly basis, separate interactions between peers or groups can influence the way adolescents perceive or relate to one another (Brown & Larson, 2009). Adolescents need interactions, but they need practice at forming positive relationships. 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. (Brown & Larson, 2009).

Constructivist theorists see the above noted differences between adolescent and adult peer interactions as a natural cognitive step in development. 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). Thus, developmental changes present a problem when adapting strategies and tools, currently used at the university level, for adolescents.

Evaluation of Group Work

To explore a different element of the process, select studies have asked students for their feelings about the peer assessment process and functions prior to completing cooperative learning assignments. Atputhasamy and Divaharan (2002) explored student sentiment toward peer assessments in order to evaluate the contribution and quality of work of teacher education students in university cohorts. Students felt that peer assessment helped "encourage and accentuate the benefits of cooperative group work." However, students also felt "awkward" assessing their peers, especially when a face-to-face assessment was incorporated, a situation surely to be magnified if used with adolescents. The authors also found that peer assessment added motivation and engagement to projects, supporting the reasoning for incorporating peer assessments into education at all levels. In addition to raising the awareness of group dynamics, peer assessment reduces the presence of 'free-riders' within groups, a concern in group work (Atputhasamy & Divaharan, 2002). Free-rider effects are generally defined as a situation in which one student does not contribute equally to a group project, but due to evaluation methods, benefits equally from the group product (Zhang, Johnston, & Gulsen, 2008).

A limitation of many studies in the literature involving peer assessment as an evaluation tool in cooperative learning is the assumption of mature peer relationships (Atputhasamy & Divaharan, 2002). The use of peer assessment for cooperative learning

in higher education institutions, where students have developed the skills to base ratings on a set of academic criteria, not social relationships, limits transferring existing group work research to younger ages (Atputhasamy and Divaharan, 2002). In addition, while Atputhasamy & Divaharan's study (2002) only attempted to gain a basic understanding of the overall effect of incorporating peer assessment into cooperative learning, the results relied, almost exclusively, on anecdotal information, neglecting quantitative methods. Despite these limitations, the conclusion, that including peer assessment into the evaluation process of group work increased perceived fairness and created greater motivation amongst students, is an important confirmation of the potential implications of peer assessment in all educational settings.

Generally, to evaluate group work, students assess the finished products and/or the contributions of group members while working on projects or presentations during a given class time. However, unlike the other studies, in Steensels, et al. (2006), students were shown step by step how to complete the peer assessment tool. The authors noted the need to meet certain conditions in order for peer assessments to be effective. Those conditions, group size, an understanding of scoring criteria and practice using the tool all played key roles in the reliability and validity of rater scores (Steensels, et al, 2006). However, despite meeting the above conditions, the validity of rater scores may still be low. For example, self-assessment can differ greatly from peer assessment as students may inflate their own contribution (Zhang et al, 2008). The effectiveness and fairness of peer assessment can be evaluated by comparing them to teacher grading (Atputhasamy & Divaharan, 2002).

Recent studies on the evaluation of group contributions draw several key conclusions, pointing to the efficacy of using peer assessment as an evaluation tool. Firstly, peer assessment "can be a valuable tool to differentiate between student contributions to the group, if students acquire the necessary skills to carry out a peer assessment." (Steensels, et al, 2006). Secondly, when using of a teacher score as a proxy of the true score, there was a significant correlation between students and teacher ratings, indicating that students may be able to assess one another in the same way as a teacher would (Steensels, et al, 2006). This confirmation of validity appears in other studies as well (Zhang, et al, 2008), possibly informing future assessment applications.

Assessment Tools

As with all teaching strategies, the introduction and calibration of educational tools is imperative for successful learning. Several studies have been conducted in an attempt to understand which type of peer assessment tool leads to the greatest success for meeting the intended purpose (i.e. contribution of work or quality of product). As with other literature related to group work, most studies related to the assessment tools themselves take place in higher education, thereby presenting a limitation to the existing literature. College students usually have a basic systematic understanding of evaluation processes, either through direct practice or indirect exposure to self and peer evaluations, and thus are more prepared to face novel challenges than adolescents. This idea is in line with many other developmental views relating to adolescents, such as information processing theories, which indicate that cognitive processes may not yield retention without sufficient attention to the problem or task (LaBerge & Samuels, 1974). Thus, with limited opportunity to practice retrieval of educational strategies, and given the variations in the

formal operations stage of development, adolescents may struggle when using the same tools as adults.

Ohland and Layton (2000) studied the reliability of peer assessment by comparing two types of peer assessment tools: one that focused on deliverables such as quality of presentation and written work, and the other that focused on characteristics of group work. A generalizability study favored the instrument focusing on specific characteristics of group contribution. They concluded that focusing on identified behavioral characteristics of good teamwork can improve the reliability of the peer assessment tool. In addition, clearly defining the behavior qualities needed for group cooperation, in particular emphasizing developmental appropriateness, may lead to greater generalizability of assessment tools across settings.

In a similar study, Lejk and Wyvill (2001) compared two types of peer assessment tools, but drew a different conclusion from that of Ohland and Layton (2000). In the study by Lejk and Wyvill (2001), college students were randomly split into groups and then given two different forms of assessments. The method of randomized grouping, not used in many other studies of this nature, may help in minimizing variance and increasing the reliability of peer assessment (Zhang, et al, 2008). This study found that holistic assessments resulted in higher inter-rater reliability than assessments of group traits. This conclusion, while not in direct conflict with the Ohland and Layton study (2000), suggests that the analysis of contribution scores must take into account both rater and group dynamics.

Use of the peer assessment tool is dependent on the goal of the group work project itself. If peer and self-assessments are to be used as a way to rate a finished product, then studies appear to support the use of holistic assessment tools (Lejk & Wyvill, 2001). However, if peer and self-assessments are intended to incorporate contribution scores to an overall rating, then a category based tool, indicated by Ohland and Layton (2000) will generate higher reliability. This conclusion presents an interesting, and important element in the use of peer assessment. Using group work as an evaluative strategy depends greatly on the type of activity being assessed, either group contribution or final product. As noted by Steensels, et al (2006), the practice of using a specified peer assessment tool seems to improve the accuracy with which students can use the tool. However, because many of these studies have taken place in a cohort type system, where students work with one another for multiple semesters, free-rider effect, as well as ratings assigned based on friendship (friendship effect) seem to increase over time (Zhang, et al, 2008). Additional findings are indicated by Steensels, et al (2006), illustrating a need for supplemental ratings and a true score (from a teacher or observer) to validate student ratings. The triangulation of scores supports other findings, indicating that peer and self-assessments may not be suitable as the only means of evaluation for group contributions. Correlations to teacher ratings also indicate some variance in understanding of rating categories, adding to the importance of developmentally appropriate standards for evaluation (Steensels, et al, 2006).

Generalizibility Theory

One major challenge in using peer assessment for evaluation is rater reliability. Rater reliability refers to the consistency and agreement of raters in rating peers. If

students can reliably use peer and self-assessment tools in one setting, then through practice, students can be taught how to evaluate their peers in a variety of settings (Lindblom-ylänne, Pihlajamäki, & Kotkas, 2006). If students prove unreliable in their use of peer assessments, that method of evaluation may not be developmentally appropriate to assess group contributions.

Generalizability Theory is one method frequently used to assess rater reliability. The major benefit of using Generalizability Theory rather than classical test theory framework is in the determination of specific causes of potential error within the context of any given study. When used to determine inter-rater reliability, a Generalizability study incorporates error terms from multiple sources, such as person, rater, group, and the interaction among them (Webb & Shavelson, 2005). The sources of error are referred to as *facets*, similar to factors in an ANOVA design. Two generalizability coefficients can be calculated in determining the overall reliability of raters. The reliability coefficient (g) represents relative decisions, taking into account the error terms associated with the person facet. The reliability coefficient (d), on the other hand, includes all compounds of the error, thus more relevant for absolute decision making. The variance of various sources can be estimated by using standard ANOVA procedures.

In the case of group work, nested models (Figure 1) have been shown effective in two studies: Zhang, et al. (2008) and Ohland & Layton (2000). One of the greatest advantages of g-theory is that by separating error terms, reliability problems can be identified and addressed to improve results (Briesch, 2013). For example, in a classroom group work project, a large error term for the group facet would indicate differences between the groups, which may be addressed through a variety of strategies, such as

randomization or assigned seating (Webb & Shavelson, 2005). The nested model in figure one shows the relationship between the person (p), rater (r), and group (g) interactions.

Validity of Peer Assessments

The validity of peer assessment has been studied by comparing peer ratings to true scores assigned by a teacher or an observer. Previous studies have shown that university students are able to consistently rate each other accurately when compared to teacher scores (Ohland & Layton, 2000). However, the context of the Oholand and Layton (2000) study is specific to adult engineering students. As pointed out by Steensels, et al. (2006), frequent exposure to peer assessment tools and procedures increases the likelihood that students can accurately measure the intended constructs. An issue not being studied is what type of prerequisite skills are needed in order to understand the constructs associated with group work, and at what developmental stage these skills are acquired (Wentzel & Caldwell, 1997). This issue is particularly relevant to the current research, where adolescents are the targeted population.

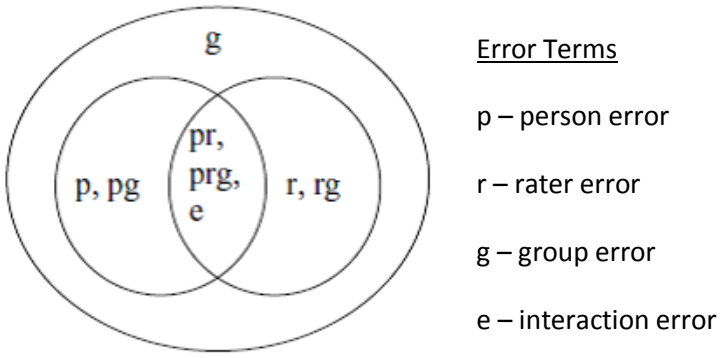

Figure 1.

Generalizability Model, Nested Design Indicating Interaction of Error Terms

Driving the current study is the concept that adolescent development impacts group work in such a way that middle school students face unique challenges in using peer and self-assessment tools that are not faced by college students, which have been the targeted population in most group-work studies. The current research aims to answer the following two questions:

1. Is peer assessment a reliable tool for evaluating group contributions in middle school classrooms?
2. Are the peer assessment and self-assessment scores valid or are they significantly different from scores assigned by teachers?

In alignment with previous research, it is hypothesized that validity of peer assessments will be highest when self-assessment scores are not included, as students may not accurately perceive their own contributions to the group (Zhang et al, 2008). Furthermore, reliability coefficients are expected to be unacceptable because of group dynamics and a lack of practice using the assessment tools. Based on evidence regarding peer relationships, student interactions with fellow group members, particularly in self-selected groups, may limit rater reliability across groups. Friendship differences between students may also relate to the reliability of peer ratings by increasing the amount of error between groups (Yugar & Shapiro, 2001).

Methods

Sample

Forty-five sixth grade students were selected from an urban charter middle school. Students were chosen as a convenience sample, based on the availability of regular access to the students and on the cooperation of school administration. The students (25 female and 20 male) were observed in an English class, which they attend 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 4 students being suspended for one or more days.

The parents and/or guardians of the students were notified by the researcher and also by the school of the observations taking place in the classroom. In addition, as a new peer assessment form was being introduced in the classroom, parents and guardians were given the opportunity to view the form before the study began. No parents/guardians chose for their students to abstain from the study. Students were told about the study in advance and were given the option not to participate in the use of any new forms or protocols. However, as the classroom assignments and projects were not introduced as part of the study, participation was required.

Measures

The peer assessment form was adapted from a tool used in a previous study that has been shown to have higher reliability (Ohland & Layton, 2000) than other types of tools. The form consisted of 10 items, measuring two constructs; group contributions and quality of work. As shown in Appendix B, the items *Attends class daily, Quality of work,*

and Completes work on time measure the quality of work construct while the items *Participates in discussions, takes turns talking, Listens carefully to others, Takes group job seriously, Accepts ideas from the rest of the group, Respects other group members, and Helps other group members* measure group contributions. The two constructs are defined in the current study as the *contribute* variable and the *quality* variable.

The measurement of reliability of the study is intended to indicate whether students were able to provide consistent ratings for each other. This is accomplished through the use of student scores from the peer and self-assessment tool. Students completed one column of the tool for each of their group mates, as well as one column (the 1st column) for self-assessment. When students had submitted the assessment form, the teacher recorded her scores in the last column. The ratings, both peer and teacher, were kept anonymous to the students. Scores for each rater were averaged to create an overall score on the 1-4 scale for each person and rater. The combination of these scores were used in the generalizability study to determine reliability coefficients. In alignment with previous research, the data from the peer assessments were analyzed for reliability using the Generalizability Theory described by Zhang, et al (2008). A general standard of .70 is acceptable for reliability coefficients in the behavioral sciences (Briesch et al. 2014). In addition, measures of error terms indicate the sources of possible error from the raters.

A pre survey (Appendix A) was given to students before the project began. As indicated by Atputhasamy and Divaharan (2002), the incorporation of peer assessment into the grading structure of cooperative learning projects has been shown to increase positive attitudes toward group work projects. The intent of the pre survey was to gage

general feelings about group work. Low reliability and validity may be accounted for if students indicated negative feelings toward group work. Thus, the survey consisted of 10 items related to the feelings of working in groups, such as *I like to work in groups*, and of group relationships and dynamics, *When I work in groups, I like to work with my friends*. The pre survey also contained one item asking students to identify their ideal group size, which the classroom teacher used to assist in the facilitation of the self-selected groups.

A post survey (Appendix C) was given to the students at the end of the study in order to provide additional data aligned with the pre survey. However, unlike the pre survey, the purpose of the post survey is to gage general student feelings about the experience of working in groups and completing the project with others. The eight items on the post survey were constructed using questions relating to attitude towards group and individual contributions, as well as one item asking whether the students *liked working in a group*. Both surveys, although more so the post survey, were designed for the specific use of this study and were intended to be used after the assessment tool had been analyzed. The goal of incorporating the pre and post survey data was to help explain the effects that may contributed to low reliability and validity. In addition, data analyzed from the pre and post survey may help shed light on possible explanations as to why students may not be reliable and accurate rater of self and peer work. The surveys aim to address developmental issues related to peer relationships and incorporate that information with the reliability and validity data.

Free-rider and friendship effects are important measures associated with rater reliability. Due to the especially important significance of peer relationships for adolescents, it was presumed that some level of friendship effect might occur within

groups (Wentzel & Caldwell, 1997). Self-selected groups, although beneficial to motivation, may be a hindrance to rater reliability (Zhang et al, 2008). See *statistics section* for discussion of free-rider and friendship effects and their respective means of measurement.

Validity of the peer and self-assessment scores was evaluated by comparing student ratings to teacher scores. Teacher scores were derived using the same assessment tool as the peer and self-assessments. Teacher scores are representative of "true scores" in the study, as a teacher's ability to assess students is accepted as objective and accurate. In general group project situations, it falls solely on the teacher to assign ratings for contribution and finished product to students. For the comparison of student and teacher scores, paired sample t-tests were conducted. T-tests were also used to measure the validity of scoring on the two sub-constructs (group contribution and quality of work). A significant difference, either higher or lower, indicates a lack of accuracy between peer generated scores and true scores.

Procedures

In the current study students were observed working on a group project during a unit activity (Appendix D) in their English class. The assignment consisted of a culminating project after the class had 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 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.

Before groups were formed, students completed the pre survey (Appendix A). Students were then asked to form groups of three; there was one group of 2 students and a group of 4 students. The teacher facilitated, but did not organize or reorganize the groups. Although self-selection has been shown to lower rater reliability (Steensels, et al, 2006), the teacher believed that random grouping might lead to withdrawal and less enthusiastic behavior (Wentzel and Caldwell (1997). Students had four days to work together to complete the project and presented the final product to the rest of the class on the fifth day. While the project was worked on within groups, the teacher was responsible for monitoring groups and making note of any abnormal behaviors (e.g. fighting).

After students presented the project, they used the single-form assessment tool to complete their self-assessment and a peer assessment for each of their group members (Appendix B). The forms were then submitted to the teacher who provided a score for each student on their respective form.

After completing the group projects and self and peer assessments each student completed the post survey. Survey information was anonymous and did not include group information.

The teacher then calculated a grade for each group member using an average of the teacher score and the self and peer assessment scores for each student. Students did

not see the ratings from their group members and were only aware of two overall scores presented by the teacher; an average contribution score and a product score, computed by averaging peer and self scores and using the school rubric respectively. These scores were not directly related to the study and were used solely for internal classroom grading.

The researcher used the raw data scores from the peer and self-assessments to calculate an average score for each student from each rater. The researcher also calculated scores for each student with and without the self-assessment scores. For the purpose of this study, teacher scores were used only as a true score in measuring validity. These scores became the data set used in determining both reliability and validity.

Results

Validity

Table 1 illustrates the mean scores for the peer and self-assessments. Student scores were averaged for each comparison between student and teacher scores based on the 1 to 4 scale used on the assessment. Comparisons were used due to the relationship of groups, each comparing the student scores to teacher scores. Comparison 1 represents the overall mean of student and teacher scores for the *peer with self* assessments. Comparison 2 represents the overall mean scores for the students and the teacher for *peer without self* scores. Comparison 3 shows the mean scores for the *quality with self* variables derived from the quality construct for students and the teacher. Comparison 4 shows the *contribute with self* variable scores from the students and teacher. Comparisons 5 and 6 illustrate the *quality* and *contribute* variables without self assessment scores. The greatest difference between student scores and teacher scores is visible in Comparison 3, where

students gave themselves, and their peers, scores on average of .77 points higher for quality of work than they were given by the teacher. Student scores were closest to teacher scores in Comparison 2 and 6, the *peer without self* scores and the *contribute without self* scores, respectively.

Table 1

Mean Scores for Peer and Self-Assessments from Students and Teacher, Including Contribution to Group Work and Quality of Work Constructs

	Assessments	Mean	N
Comparison 1	Peer with self	3.18	41
	Teacher Score	2.91	41
Comparison 2	Peer without self	3.00	44
	Teacher Score	2.84	44
Comparison 3	Quality with self - Student	3.37	41
	Quality - Teacher	2.60	41
Comparison 4	Contribution with self - Student	3.42	40
	Contribution - Teacher	3.08	40
Comparison 5	Quality without self – Student	2.94	41
	Quality Teacher	2.60	41
Comparison 6	Contribute without self – Student	2.91	41
	Contribute Teacher	3.04	41

Paired sample t-tests were used to examine the validity of student scores. Student scores, with and without self-assessments, were measured against teacher scores, which were used as "true" scores (Table 2). In addition, scores were broken between two identified constructs: *quality* and *contribute*. Comparison 1 compares the *peer with self* to the teacher scores. Comparison 2 compares *peer without self* to the teacher scores. Comparisons 3 and 4 each compare assessment constructs with self assessments, *quality with self* and *contribution with self* respectively, to the teacher scores on the same constructs. Comparisons 5 and 6 measure the construct variables without the self assessment scores against the teacher scores. Three items account for the quality construct and ten items indicate the contribution construct. T scores indicate the significance of the differences between student and teacher scores, measured at an alpha level of .05. Congruence of *peer without self* and teacher scores yielded no significant difference, $t=1.61$, $p=.12$. All scores which included the students' self-assessments were significantly different from the teacher scores $p<.05$. For all comparisons except comparison 6, students recorded higher scores for themselves and peers than the teacher awarded. The difference between *contribute without self* and teacher scores (comparison 6) was not significant, $t= -1.03$, $p=.310$.

Table 2

Paired Samples T-Test

Comparison	Variable	SE Mean	t	df	Sig. (2-tailed)
Comparison 1	Peer with Self Teacher Score	.09	3.23	40	.002*
Comparison 2	Peer without Self Teacher Score	.10	1.61	40	.116
Comparison 3	Quality with self - Student Quality Teacher	.11	7.32	40	.000*
Comparison 4	Contribute with self - Student Contribute Teacher	.12	2.82	39	.008*
Comparison 5	Quality without Self – Student Quality Teacher	.11	3.14	40	.003*
Comparison 6	Contribute without Self – Student Contribute Teacher	.13	-1.03	40	.310

**p<.05*

Reliability

A Generalizability Study design was used as the framework for evaluating the reliability of peer and self-assessments. Table 3 shows the generalizability coefficients for the overall assessment scores which excluded the self-assessments and the student

assessment scores with self-assessments. The letter *g* represents the reliability coefficient for the Generalizability Study, which is the relative decision. The *person* term indicates the amount of error associated with each person's score. Because not all students contribute the same amount to the group project, there is expected to be variance between people. The *rater* term indicates the amount of error associated with the scores given to the person by that person's group members. The *group* term illustrates the error between each group. The last *error* term represents the amount of error from non-specified sources, or the combination of errors between the person, rater, and group. These error terms represent the percentage of overall variance which can be accounted by each facet.

The reliability coefficient (*g=.95*) for the assessments without self scores indicates an acceptable level of reliability. The low rater variance coefficient of .19 for the overall score without self-assessment shows that students were reliable raters of each other's contributions within the groups, accounting for only 19% of the total variance. High group variance of 34% indicates that peer assessment scores varied by group. The dependent variable of average score, which includes the students' self-assessments, had a drastically lower g coefficient than the peer scores alone (*g=.51*). The variance coefficient value of *0.00* for the person-rater error does not signify that there was no variance in scores, but, instead indicates that the model was disrupted by negative variance, caused by the within-group variance being larger than the between-group variance, as noted by Briesch et al (2013).

Table 3

Reliability Coefficients (g) and Error Terms

Variable	G	Person	Rater	Group	Error
Peer without Self	0.95	0.41	0.19	0.34	0.06
Peer with Self	0.51	0.17	0.00	0.34	0.49

Table 4

Reliability Coefficients (g) and Error Terms for Assessment Constructs

Variable	G	Person	Rater	Group	Error
Contribution with Self	0.49	0.16	0.05	0.31	0.49
Quality with Self	0.30	0.09	0.00	0.27	0.64

Table 4 depicts the Generalizability Study results conducted for the two group work constructs; quality of work and group contribution. The first variable, Contribution to Group Work is categorized by the seven items mentioned on page 17. Similarly, the Quality of Work construct is defined by the 3 items also mentioned on page 17. The G column indicates the reliability coefficient generated by the relative decision. The four error terms, person, rater, group, and error, show the variance coefficients for each of the facets on the contribution and quality constructs. Both construct scores were averaged

using peer and self-assessments. Overall, neither g coefficients proved to contained an acceptable level of reliability with the contribution construct *g=.49* and the quality construct *g=.30.* Error terms, again, indicate large group effects, while person error accounts for only 15% in the contribution construct and 9% in the quality construct. As was the case in Table 3, the rater error term for the quality of work construct does not imply a lack of variance, but instead represents negative variance.

Furthermore, the large values of the error terms for all variables, except the overall score without self-assessment, indicate that the combined error term, variance between the person, rater, and groups, damages reliability of the peer assessment tool. These error terms, visible in Table 3 and 4, indicate that nearly 50% of error is attributed to the interaction of facets.

Survey Results

Survey results (Table 5) showed a majority of agreement for group work compared to individual assignments. An analysis of the pre and post survey data indicates that students generally enjoyed group work, both before and after engaging in the activity. Students showed the greatest percentage of agreement for items *I like to work in groups*, *I would like to make my own group,* and *When I work in groups, I like to work with friends*. Item *I am willing to do my share of work in a group* also was 82% supported, indicating that the students did not mind the idea of equal contribution to the group project. Lowest agreement came from items *I am willing to do more than my share in a group* and *I like when I have one job to do*. The pre survey also contained a final

item, for which students identified their ideal group size. The average ideal group size was 3.5, which the teacher used to guide students to for groups of threes.

Table 6 illustrates findings from the post survey, which support the results of the t-tests, indicating that students tended to inflate their sense of contribution to the group. Item *I did more than my share in the group*, resulted in 61 percent of students saying "yes." 78 percent of students indicating that *I worked with my friends*, peer assessments were consistent with teacher ratings. The high percentage of students claiming to work with their friends may explain some of the variance in the group error term from the generalizability study. Overall, after the completion of the project, 78 percent of students said that they *Like working in my group*, compared to 82% before the project began. Despite self-selecting groups, students indicated in Item *I made my own group* that 42% of the time did they feel that they made their own group members.

Overall, items from the post survey which were similar to the pre survey items, such as *I liked working in a group and I did my share (of work) in my group*, showed a decrease in students who agreed. This may indicate a lack of understanding, interest, or a dislike for the group work. However, the most consistent item, *I am willing to do more than my share/I did more than my share*, adds further support to the problems caused by self evaluations, echoing the lack of reliability and validity for ratings when self assessment was included.

Table 5

Pre Survey Items and Percent of Students Who Agree

Item Number	Item	Percent in Agreement
1	I like to work in groups	82
2	I would like to make my own group	87
3	When I work in groups, I like to work with my friends	82
4	When I work in groups, I like to work with the smartest students in the class	67
5	When I work in groups, I like to work with students who are at my level	76
6	I like to be the leader in a group	73
7	I am willing to do my share of the work in a group	82
8	I am willing to do more than my share in a group	62
9	I like when I have one job to do	62
10	I like when I can help with lots of jobs	78

Table 6

Post Survey Items and Percent of Students Who Agree

Item Number	Item	Percent in Agreement
1	I liked working in a group	78
2	I made my own group	42
3	When I worked in a my group, I worked with my friends	78
4	I was the leader in my group	36
5	I did my share of the work in my group	68
6	I did more than my share in the group	61
7	I had more than one job in my group	59
8	I would like to work in a group next time	78

Discussion

Summary

The results of the current study lead to several important findings. In both reliability and validity analyses, the inclusion of self-assessment scores gave worse

results, which is consistent with findings from previous research by Zhang et al. (2008). In the case of validity, when self-assessments were discounted, students' rating of each other's contribution was as good as the teacher's. The difference in results from the with-self and without-self scores indicates that students may not rate themselves in the same way that they rate others. Validity test results from both the contribution and work constructs indicate that students were able to rate their peers on the desired set of items with no significant difference from the teacher ratings. In addition, when students' ratings are different from the teacher's, their ratings tend to be higher.

Reliability tests also illustrated the important differences between self and peer assessments. Reliability is acceptable only when self-assessment was excluded. Back to the original research questions for this study, these findings suggest that in using students to provide group contribution information, self assessment should be avoided altogether.

Similar to the qualitative findings of the Atputhasamy and Divaharan (2002) study, results from pre and post surveys in the current study demonstrate that the inclusion of peer assessment with group work may increase student motivation. As group work is increasingly required in middle school curriculum, it is imperative to provide students with developmentally appropriate tools for assessing such work. Survey results indicate that students do prefer group work to individual work, at least in a project-based setting. Therefore, the practical application of this study, and of assessing group work as a whole, adds important information about student feelings towards group work and its assessment.

Limitations

The first limitation is the sample used. The overall sample size is relatively small (n=45). Moreover, the sample represents a highly homogeneous group in terms of ethnicity and achievement and behavioral levels. Due to the academic and behavioral deficits present in the sample of students, inferences drawn from the study may only be relevant to a similar sample. In addition, as this study was conducted during relatively short period of time, absenteeism and suspension rates may have influenced group dynamics. As indicated in Piaget's (1972) work, the use of younger samples presents a set of developmental challenges not present in studies focused on adult students. The current study shows that findings on group work drawn from university student samples are not readily transmissible to adolescents as the experiences with group work are very different.

Opportunities for Future Research

Despite the limitations of the current study, findings do indicate the potential for further research in several areas. A larger and more diverse sample may lend itself to further support the assumption that adolescents are, in fact, accurate and reliable evaluators of each other's group contributions. However, the students sampled in this study, despite representing a very specific segment of the population, are also the students with the greatest educational needs. Urban minority students are often at the bottom end of American education and serve to benefit most from studies that directly impact the way they learn and the way in which they are assessed. By incorporating the effects of social learning activities in schools that have the greatest social deficits, research can make more practical and meaningful gains.

Furthermore, types of group projects and different peer assessment tools may be more or less useful depending on the context in which they are used. As indicated in Lejk and Wyvill (2001), holistic assessment tools may be preferred in situations where students are focusing on summative classroom assessments.

Optimal statistical methods may also differ between samples and situations. While it is often accepted that Generalizability Study Designs are the most accurate and effective for evaluating multi-dimensional rater reliability, the current study should articulate the importance of carefully designed statistical procedures. Briesch et al (2013) notes that the use of a Bayesian approach, restricting negative variances, may improve reliability testing results, in particular, by reducing negative variance coefficients. Although the negative variance coefficients in the Generalizability Study were problematic when analyzing the reliability of the self and peer assessments, the results do shed light on the importance of future research. Generalizability Theory has been shown through several studies to be the most effective tool to measure inter-rater reliability (Ohland and Layton. 2000 and Zhang et al. 2008), and is typically accepted in the social sciences as reliable. However, in this particular study, due possibly to the inclusion of self-assessments in the person-rater variance, the G-study design may not be the preferred framework for analysis. Results from this study should proceed to help inform future research regarding ways to evaluate multi-faceted variances.

The importance of developmentally appropriate research is the final key component to the current study. Adolescent educational experiences differ greatly from those of university students, both academically and socially. Group work studies that emphasize different types of learners from different backgrounds may yield results

beneficial to educational decision making. When group work is used as a tool to judge learning, students need to understand the academic and social criteria for which they are accountable.

References

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

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. and Larson, J. (2009). Peer Relationships in Adolescence. Handbook of Adolescent Psychology. Chapter 3 Peer Relationships in Adolescence. 74-103.

Cohen, G. L., & Garcia, J. (2008). Identity, Belonging, and Achievement A Model, Interventions, Implications. *Current Directions in Psychological Science*,*17*(6), 365-369.

Gest, S. D. (2006). Teacher reports of children's friendships and social groups: Agreement with peer reports and implications for studying peer similarity. *Social Development*, *15*(2), 248-259.

Duncan-Andrade, J. M. R., & Morrell, E. (2008). *The art of critical pedagogy: Possibilities for moving from theory to practice in urban schools* (Vol. 285). Peter Lang.

Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). Cooperative learning methods: A meta-analysis.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing I reading. *Cognitive psychology*, *6*(2), 293-323.

Lejk, M., and 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.

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

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human development*, *15*(1), 1-12.

Rubin, Kenneth H., Stephen A. Erath, Julie C. Wojslawowicz, and 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. 419-93. Print.

Steensels, C, Leemans, 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. No.2: 111-118.

Summers, M., & Volet, S. (2010). Group work does not necessarily equal collaborative learning: evidence from observations and self-reports. *European Journal of Psychology of Education, 25*(4), 473-492.

Verhaart, M., Hagen, K. & Giles, O. (2005). Best Practice Assessment methods for evaluating an individual's performance in group work. In S. Mann, T.Clear (Eds.) Proceedings of the 18th Annual Conference of the National Advisory Committee on Computing Qualifications Conference, Tauranga, New Zealand. July 10-13. pp 113-120.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: overview. *Encyclopedia of statistics in behavioral science*. No. 36: 599-612

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

Whitcomb, G. (U.D.). Lev S. Vygotsky's Social Constructivist Theory of Cognitive Development. Ball State University.

Yugar, J. M., & Shapiro, E. S. (2001). Elementary Children's School Friendship: A Comparison of Peer Assessment Methodologies. *School Psychology Review,30*(4).

Zhang, B., Johnston, L., and 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 Survey - Pre

Name_____
_____

Block #_____

Date_____

Please circle how you feel about the following activities related to group work:

1. I like to work in groups	Yes	No
2. I would like to make my own group	Yes	No
3. When I work in groups, I like to work with my friends	Yes	No
4. When I work in groups, I like to work with the smartest students in the class	Yes	No
5. When I work in groups, I like to work with students who are at my level	Yes	No
6. I like to be the leader in groups	Yes	No
7. I am willing to do my share of the work in a group	Yes	No
8. I am willing to do more than my share in a group	Yes	No
9. I like when I have one job to do	Yes	No
10. I like when I can help with lots of jobs	Yes	No

Finally:

11. The number of group members I like is:	2 3 4 5

Appendix B

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.

Group Member->	#1	#2	#3	#4	#5

3. Your answers will be kept secret.

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					

Appendix C

Group Survey - Post

Name_____
_____

Block #_____

Date_____

Please circle how you felt about the following activities while working on your group project:

1. I liked working in a group	Yes No
2. I made my own group	Yes No

3. When I worked in my group, I worked with my friends	Yes	No
4. I was the leader in my group	Yes	No
5. I did my share of the work in my group	Yes	No
6. I did more than my share in the group	Yes	No
7. I had more than one job in my group	Yes	No

Finally:

8. How would you like to work next time?	Individually	In a group

Appendix D

Name: _____ Hour: _____ Date: _____

Depression Brochure Assignment
Due Date: Friday, March 7, 2014

One way that people learn about places, people, or things that they do not know is by reading about them. But what if they don't have time to read a whole book or they just want a quick overview of the subject? The Health Department often uses brochures to inform, educate, or encourage — quickly. They use a brochure to grab the readers' attention and get them interested enough to want to know more. Your brochure will provide facts, suggestions and pictures/graphics about STRESS.

Your Task

Your assignment is to create a **Tri-Fold Brochure about Depression** to educate fellow young people about it. **Your brochure should be designed to be displayed in a community center, hospital or school to educate the public about the causes, symptoms and treatments of depression.**

Your brochure should include graphics, related statistics and be visually attractive. It must answer or address the following issues on depression:

- What is the characterization or definition of depression?
- How many people are affected by depression?
- How many teens are affected by depression?
- What factors contribute or cause the development of depression?
- How can someone tell if they are depressed?
- What treatments exist to treat your depression? What treatments or combination of treatments are most effective? Please provide a description of each treatment.
- How can people get help for the treatment of depression?
- What resources are available to help gain knowledge about depression?

The following is a list of resources that you can use to guide your research of the topic:

- Depression and Anxiety Disorders: www.feelingblue.com
- Depression: www.fhs.mcmaster.ca/direct
- National Institute of Mental Health: www.nimh-nih.gov

Your brochure must be typed and produced using computer software.

Evaluation

The attached rubric will be used to evaluate your brochure to see how well you have presented your information. Not everyone will agree on the effectiveness of a single brochure but if you have done your job well, most readers will agree that your brochure gives them the information they want and need and is easy to follow.

Appendix E

Team participation evaluation Name _____

MEEN 440 Mechanism Design
R. Layton

Project no. _____

Date._____

Instructions:

1. For each attribute listed below, assign a number using the 0-5 scale to describe each person's contribution to group work to date.
2. These evaluations are used to assign individual grades from group grades.
3. Your responses are confidential.

Rating scale:

- 5 excellent
- 4 satisfactory
- 3 ordinary
- 2 marginal
- 1 unsatisfactory

Team member —>	#1	#2	#3	#4	#5	#6	#7	#8
In alphabetical order, write team members' names, including your own. —> Include your name, but do not rate yourself.								
Attends meetings regularly								
Contributes to discussions								
Has good communication skills								
Committed to group goals								
Listens effectively								
Takes responsibilities seriously								
Accepts criticism gracefully								
Performs significant tasks								
Tasks have technical content								
Completes tasks on time								