Reliability, Validity, and Bias in Peer Evaluations of Self-Directed Interdependent Work Teams

Robert S. Thompson
Colorado School of Mines

I. Introduction

Teamwork education has become increasingly important over the last decade. In a recent survey conducted at the Purdue School of Engineering, over 76% of the students responded that they had been involved as members of student work teams (486 out of 1,953 responded) [1]. This emphasis on teamwork skills stems from the widespread use of teams in industry.

Peer evaluations are being used as a source of information for improving team performance [2,3] and accounting for individual contributions to a group project [4]. Peer evaluations as a source of information for small self-directed group work have an appeal because the team members are in the best position to observe the team skills of their fellow team members. Despite this advantage, concerns have been levied against the use of peer evaluations. Abson [5], for example, suggested that peer evaluations can be abused and have undesirable effects on individuals in the group. Mathews [6] studied peer assessment of small group work in a management studies program. He noted patterns of response included giving all group members the same score, collusion between group members, and potential ganging-up on one member. Mathews also noted that perceptions can vary between people accounting for some of the variability. Mathews comments were based on his observations. He did not report any statistical data to support his claims.

Engineering design projects are a common source of teamwork in engineering education. Teamwork in these settings is characterized by three attributes: team members having a common goal, dependence on each other to achieve their goal, and intense work over an extended period of time. Because of the extended nature of the group work and the interdependence among team members, friendships have time to develop over the duration of the project. In many cases, friendships were formed prior to the group work. With this nature of the teamwork and the corresponding use of peer evaluations, there is a need for a better understanding of the reliability, validity, and bias in peer assessments of students working on these interdependent, self-directed, problem-solving teams. This research focuses on the specific problem of the reliability, validity, bias, and user acceptance of peer evaluations in small interdependent self-directed problem-solving teams in an educational setting. Generalizability theory techniques, structured interviews, and survey data were used to answer the following research questions:

1. What is the user reaction to peer evaluations?
2. What is the level of consensus in peer evaluations?
3. What is the level of bias in peer evaluations?

4. What is the reliability (single measure) and stability (repeated measures) of peer evaluations?
5. What is the level of validity in peer evaluations?

II. Methodology

This section provides an overview of the participants in the study, data collection, data analysis, research design, and a theoretical framework and model for the research [7].

A. Participants

Seniors in a capstone design class (Multidisciplinary Petroleum Design, Spring 2000 semester) were selected for the study. The Multidisciplinary Petroleum Design course is a team focused applied problem-solving course. The course is required for all students in the Petroleum Engineering program and was an option for students from the Geology and Geological Engineering program. The course was also open to seniors from the Geophysical Engineering program. However, no students from the Geophysical Engineering program were enrolled in this particular semester. There were 49 students in the course. Table 1 summarizes the demographic characteristics (discipline and gender) of the students enrolled in the course during the Spring 2000 semester. As shown in the table, petroleum engineers and males dominated the composition of the teams.

Table 1
Gender and Discipline of Students Enrolled in
Multidisciplinary Petroleum Design Course, Spring 2000 Semester

	Gender			
Discipline	Male	Female	Count	Percentage
GE	6	1	7	14%
PE	33	9	42	86%
Count	39	10	49	
Percentage	80%	20%		100%

Note: PE = Petroleum Engineering, GE = Geological Engineering

During the semester, student teams worked on two major design projects. Each project was approximately 6 to 7 weeks in duration. Team assignments were random with the constraint that each team should have one non–petroleum engineer. Ten teams were formed for each project. For each project, there were nine teams with five members and one team with four members. Team gender and discipline demographics are summarized in Tables 2 and 3. Team membership changed completely between the first and second projects with the exception of two individuals. In many cases, team members have known each other for over two years. The projects were sufficiently long for friendships to form, regardless of whether or not the team members knew each other at the beginning of the project.

Table 2
Team Discipline and Gender Demographics
First Project

	Discipline		Gender	
Team	PE	GE	MALE	FEMALE
1	5	0	3	2
2	4	1	5	0
3	4	1	5	0
4	4	1	4	1
5	4	1	3	2
6	4	1	4	1
7	3	1	3	1
8	5	0	4	1
9	4	1	4	1
10	5	0	4	1
Total	42	7	39	10

Note: PE = Petroleum Engineers, GE = Geological Engineers

Table 3
Team Discipline and Gender Demographics
Second Project

	Discipline		Gender	
Team	PE	GE	MALE	FEMALE
1	4	1	3	2
2	3	1	2	2
3	4	1	5	0
4	5	0	4	1
5	4	1	5	0
6	5	0	5	0
7	4	1	4	1
8	5	0	4	1
9	4	1	4	1
10	4	1	3	2
Total	42	7	39	10

Note: PE = Petroleum Engineers, GE = Geological Engineers

The teamwork can be described as interdependent self-directed teamwork. The assigned problems were open-ended and required the input from multiple disciplines and data sources. Students define specific objectives, plan, and schedule their work to meet deadlines set by the faculty team. Based on these parameters, the participants were ideal for studying the reliability, validity, user acceptance, and bias in peer evaluations conducted on interdependent self-directed work teams.

B. Research Design

A round-robin research design was used for the analysis of the peer assessment data. A round-robin research design is one in which observations are made on every possible dyad in the team. In this design, each team member observes and rates each other team member. The data layout for this research design is shown in Table 4. The statistical model, referred to as a social relations model (SRM) [8-10] is particularly suited for the stated research objectives.

Table 4
Round Robin Research Design

	Ratee			
Rater	Subject 1	Subject 2	Subject 3	Subject 4
Subject 1		x	x	x
Subject 2	x		x	x
Subject 3	x	x		x
Subject 4	x	x	x	

C. Data Collection

Two peer evaluations were conducted for each project: one near the mid-point of the project and one at or near the end of each project. The peer evaluation instrument is presented in Appendix A. Peer evaluation feedback was kept confidential. Each student was presented a copy of the average peer evaluations given on each of eight identified team skills. Student names were not included on the evaluation feedback. At the beginning of the semester, students were given a unique code. An example of the peer feedback is shown in Figure 1. Self-evaluations were also provided as shown in Figure 2.

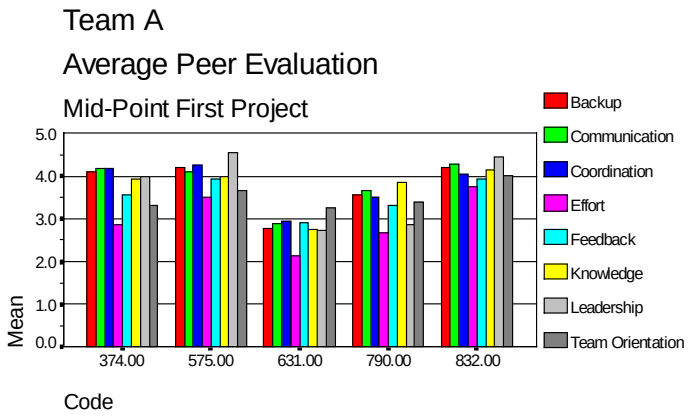


Figure 1: Average peer evaluations for Team A, Mid-point of the first project. Unique (confidential) codes were assigned to each team member

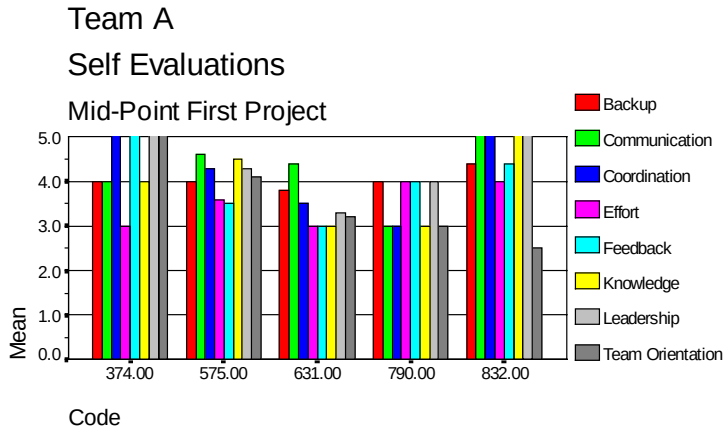


Figure 2: Self-evaluations for Team A, Mid-point first project. Unique (confidential) codes were assigned to each team member

The confidential peer feedback data were given to the students at the project mid-point and end of project review sessions. These sessions follow in the same week the peer data were collected. Concurrent with the review sessions, confidential one-to-one interviews were conducted by a third party (not the author). For each project, eight to ten individuals were interviewed at the project mid-point and at the end of the project. Thirty-five individual interviews were conducted during the semester. The faculty team did not know who (or when) students were interviewed until after course grades were submitted. The interview questions (Appendix B) focused on user acceptance, bias in the ratings, and communication between raters. Finally, the structured confidential interviews were augmented by an end of course survey completed by all the participants. The survey questions paralleled the interview questions (Appendix C).

D. Data Analysis

The data from the four peer evaluations were analyzed using the statistical technique referred to as generalizability theory [11,12]. This is the first application, to my knowledge, of the method to interdependent self-directed work teams working together in an educational setting over an extended period of time (approximately 6 weeks). Reported data on reliability and validity commonly use correlation techniques, percentage of agreement, comparison of average scores (with ANOVA), and rank order comparisons. There are limited examples applying generalizability theory to peer ratings of small groups [13-15]. The group work in the Montgomery research was limited in duration to two 15-minute sessions. Kenny, Lord, and Garg used data from research conducted by Lord, Phillips, and Rush [16]. In the Lord, Phillips and Rush study, the group work was limited to four 15-minute sessions. Hennen [13] was the only example found that applied the generalizability technique to interdependent self-directed group work in an educational setting. In Hennen's research, each team performed simulated self-directed interdependent group work. The groups worked together for three class periods.

Generalizability theory focuses on identifying multiple sources of variation that occur simultaneously in any measure. The focus is on variance and correlations rather than differences in means. The social relations model (SRM) [9] was used in this research. The model uses a two-way random effects ANOVA. The two factors, rater and ratee, are the independent variables. Each factor is one of the roles in the dyadic process. Each level for each factor is one of the team members where the team members are randomly selected. The model partitions the ratings into the relative variance from three sources of variation that are of interest in peer assessments: 1. What is the tendency for raters to give similar ratings to each ratee (rater effect)? 2. What is the tendency among raters to agree (consensus) with other raters (ratee effect)? and 3. What is the variance unaccounted for by the rater and ratee effects (rater by ratee interaction).

E. Theoretical Model

A theoretical model developed by Kenny [17] provides a framework for understanding the interdependency of factors that determine the level of consensus and validity in peer assessment data. The theoretical model was used to understand the factors that influence the partitioning of the variance into rater effects, ratee effects, and rater by rater interaction. This mathematical model is a modified version of Anderson's [18] weighted average model. The model variables are

presented in the following paragraphs. The model, as presented, is a generalization of the Spearman-Brown prophecy formula from classical test theory.

The theoretical model, Figure 3, includes three components that influence rater perception and inter-rater consensus: meaning attached to observed behavioral acts, meaning attached to stereotypes, and meaning not attached to observed behavioral acts or stereotypes (unique component). The schematic shown in Figure 3 is for two raters interacting with one ratee. The following terminology is used in the model:

A_n, the n^{th} behavior act
S_{jn}, meaning (scale value) given by rater "j", act "n"
I_j, the impression formed by j^{th} rater
S_{ju}, unique meaning (scale value) not attached to observed behavioral acts or stereotypes by rater "j"
k, the weighting factor for the unique meaning
S_{jp}, stereotype meaning (scale value) by rater "j"
w, weighting factor for stereotype meaning
a, the degree to which raters influence one another

The variables that determine the level of consensus are defined as follows:

1. Acquaintance (n). Acquaintance is the amount of information to which the rater is exposed.
2. Overlap (q). Overlap is the extent that two raters observe the ratee at the same time.
3. Consistency within a rater across acts (r_1). Within rater consistency, correlation between S_{11} and $S_{12,}$ as shown in Figure 3. This can also reflect the consistency of the ratee's acts.
4. Shared meaning systems (r_2). The extent to which an act is given the same meaning by two raters, correlation between S_{12} and $S_{22,}$ as shown in Figure 3.
5. Consistency between-raters across acts (r_3). The model assumes the between-rater consistency correlation equals $r_1 \times r_{2.}$
6. Agreement between raters about stereotypes (r_4). To what extent do the raters agree with each other about stereotypes. It is the correlation between S_{1P} and S_{2P}, the meaning (scale value) attached to stereotypes by rater 1 and 2.
7. Consistency within a rater between stereotypes and an act (r_5). It is the correlation between S_{1P} and $S_{11,}$ the meaning (scale value) attached to stereotypes and behavior act 1 by rater 1.
8. Consistency between a rater's evaluation of a stereotype and another rater's evaluation of an act (r_6). This parameter can be viewed as a "kernel of truth" since it represents the correlation between "truth" (the ratee's behavior) and the stereotype that the rater has about the ratee's behavior. The "kernel of truth" correlation equals $r_4 \times r_{5.}$
9. Communication between raters (a). The degree that raters influence one another.

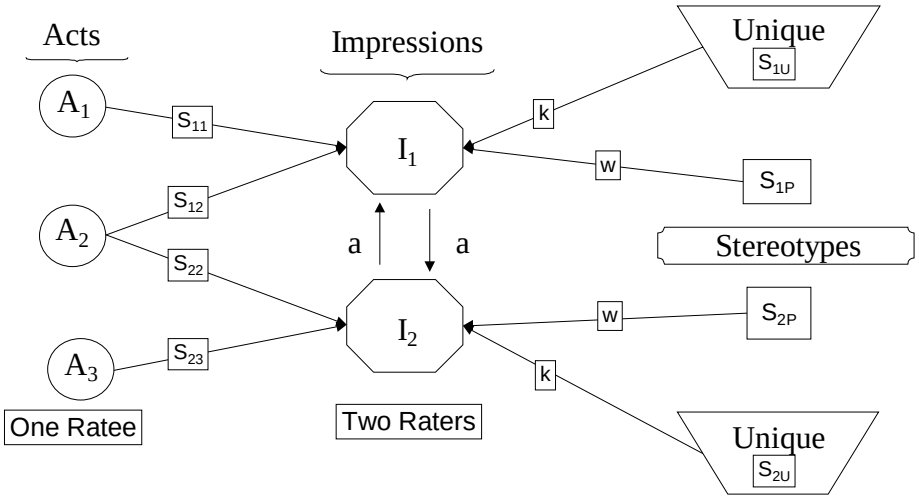


Figure 3: Conceptual model for inter-rater consensus after Kenny [17]

The rater impression is the weighted average of each of these components. This relationship is shown in the following equation for two raters observing one ratee over "n" behavioral acts:

$$I_1 = [(wS_{1p} + kS_{1u} + \sum_{j=1}^{n} S_{1j}/(w + k + n)] + aI_2 \qquad \text{Equation 1}$$

where:

I_1, impression formed by rater 1
w, the weighting factor for stereotype impressions
S_{1p}, the stereotype meaning attached to ratee by rater 1
k, the weighting factor for unique impressions (impressions not attached to observed behavior acts or stereotypes)
S_{1u}, the unique meaning attached to ratee by rater 1 that is not based on observed behavioral acts or stereotypes
$\sum_{j=1}^{n} S_{1j}$, the summation of the meaning attached to "n" behavioral acts by rater 1
a, is the degree that the two raters influence each other
I_2, the impression formed by rater 2

As shown, the behavioral acts are weighted equally while individual weighting factors are applied to the stereotype (w) and unique (k) impression variables. The degree of consensus (c) between a pair of raters evaluating a common set of ratees with no communication effect (a = 0) is defined by the following equation [8].

$$c = \frac{w^2 r_4 + 2wnr_6 + qnr_2(1 - r_1) + n^2 r_1 r_2}{k^2 + w^2 + n^2 r_1 + n(1 - r_1) + 2wnr_5} \qquad \text{Equation 2}$$

where:
w, the weighting factor for stereotype impressions
k, the weighting factor for unique impressions (impressions not attached to observed behavior acts or stereotypes)
n, the number of behavioral acts observed by each rater
q, the fraction of observations that the raters have in common
r_1, consistency within a rater across acts. Can also reflect the consistency of the ratee's behavior.
r_2, the extent that an act is given the same meaning by two raters
r_4, agreement between two raters about stereotypes
r_5, consistency within a rater between a stereotype and an act
r_6, consistency between a rater's evaluation of a stereotype and another rater's evaluation of an act

F. Model Implications

The theoretical model is used to demonstrate several important determinants of consensus and accuracy in peer evaluations. The limiting factor (maximum value) for consensus is r_2, the extent to which an act is given the same meaning by two raters. This statement assumes that there is no communication between raters. Consensus increases rapidly for cases where overlap is high. Thus, if communication between raters is zero, overlap and similar meaning systems control the level of consensus. *Figure 4 demonstrates the importance of overlap in the observations, and the maximum value of consensus being equal to the assumed value of 0.50 for r_2 (similar meaning systems).* In Figure 4, stereotypes and unique impression are given weighting factors of zero

In Figure 4 the level of within rater consistency (r_1) was assumed to be 0.10, a low value. Keeping the value for r_2 (similar meaning systems) at 0.50 and increasing r_1 from 0.10 to 0.50, the level of consensus can reach the maximum value of 0.50 at relatively low levels of acquaintance. *Figure 5 demonstrates the concept that acquaintance may not be a significant variable in reaching the maximum value for consensus in peer ratings when there are moderate levels of within rater consistency.* Within rater consistency also reflects the consistency of the ratee's behavior.

Finally, communication between raters can mask all the other parameters. This is demonstrated in Figure 6. In this figure, the assumptions are the same as in Figure 4. Again, the limiting value for consensus is the correlation for similar meaning systems ($r_2 = 0.50$) when there is no communication between raters. Under the same assumptions but with the communication between raters factor ($a = 0.50$) added, the level of consensus increases rapidly to 0.90. *Figure 6 demonstrates the importance of understanding the level of communication between raters that takes place in peer evaluations.*

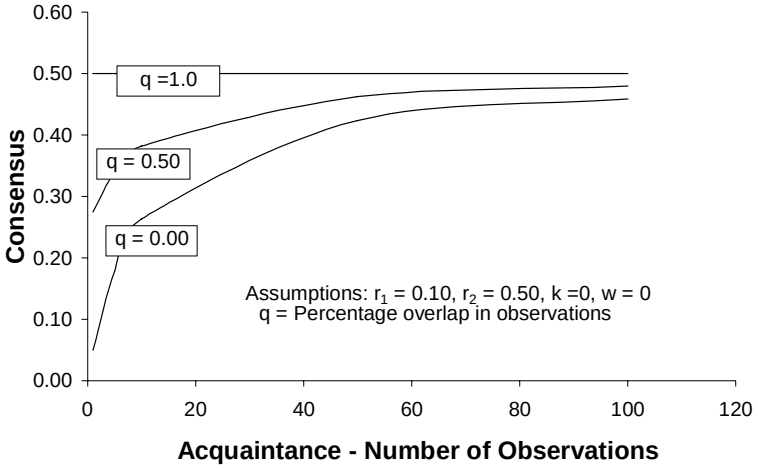


Figure 4: Consensus as a function of acquaintance and overlap. The weighting factors for unique impression (k) and for stereotype (w) are 0.0. Similar meaning systems (r_2) is assumed to be 0.50 and limits the level of consensus. Within rater consistency (r_1) is assumed to be 0.10.

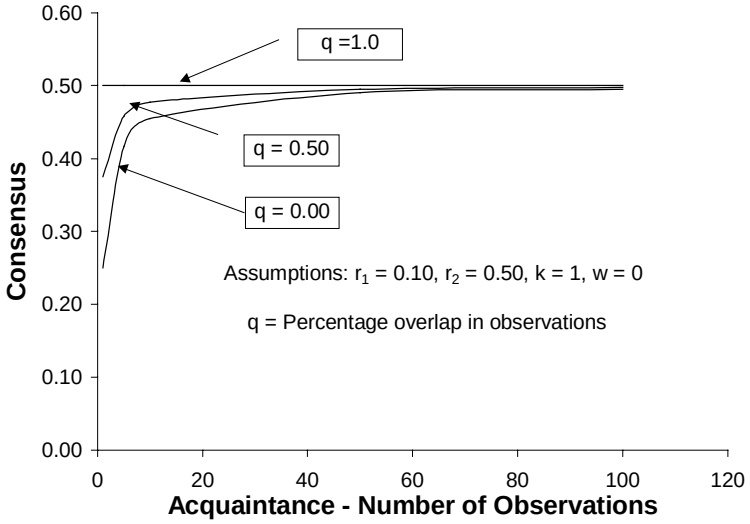


Figure 5: Consensus as a function of acquaintance and overlap. The value for within rater consistency (r_1) is increased to 0.50 compared to 0.10 in Figure 4. All other parameters are the same as Figure 4. Similar meaning systems ($r_2 = 0.50$) limits the level of consensus.

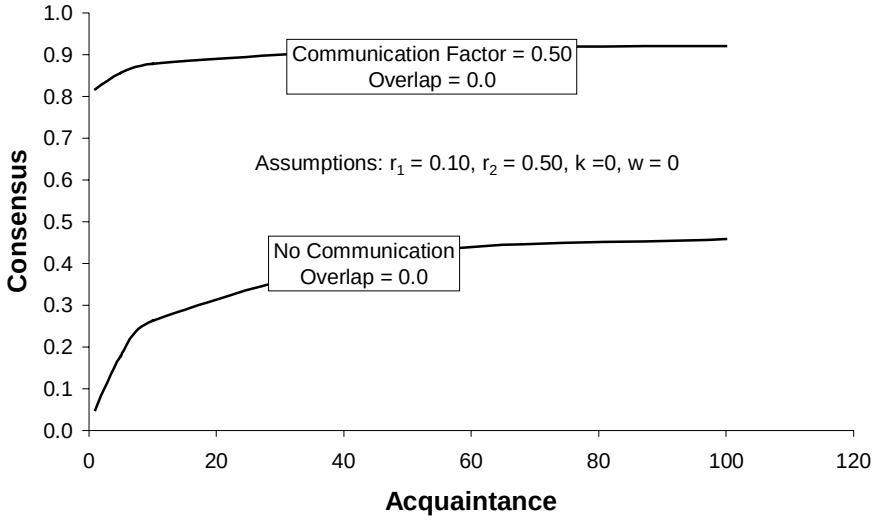


Figure 6: Consensus as a function of acquaintance and communication at zero overlap. The parameter values are the same as Figure 4. Similar meaning systems ($r_2 = 0.50$) limits the level of consensus for the no communication case.

Some researchers assume that consensus implies accuracy [17]. As Kenny points out, this is true if the raters do not communicate, there is no overlap in the observation of the acts (q = 0), and there are no stereotype (w = 0) or unique impression (k = 0) effects. In this situation, the square root of the consensus correlation can be used to determine the maximum level of accuracy. This fits classical test theory, where for two equally valid measures of a construct, the validity coefficient equals the square root of the correlation between the two indicators. For interpersonal perception, accuracy is the square root of consensus with the overlap term dropped as shown below:

$$v = \sqrt{\frac{w^2 r_4 + 2wnr_6 + n^2 r_1 r_2}{k^2 + w^2 + n^2 r_1 + n(1 - r_1) + 2wnr_5}} \qquad \text{Equation 3}$$

Both consensus and accuracy are limited by similar meaning systems (r_2). The correlation for similar meaning systems is the limiting value for consensus. The square root of the correlation for similar meaning systems is the limiting value for accuracy. *Figure 7 demonstrates that consensus is not always a proxy for accuracy.* In this example, overlap is high (q = 1.0), communication between raters is low (a = 0), similar meaning systems is high ($r_2 = 0.5$), and within rater consistency is low ($r_1 = 0.05$). For these assumptions, consensus is greater than accuracy at low levels of acquaintance. Accuracy is greater than consensus at higher values of acquaintance and continues to increase with acquaintance until the maximum value of approximately 0.71 is reached.

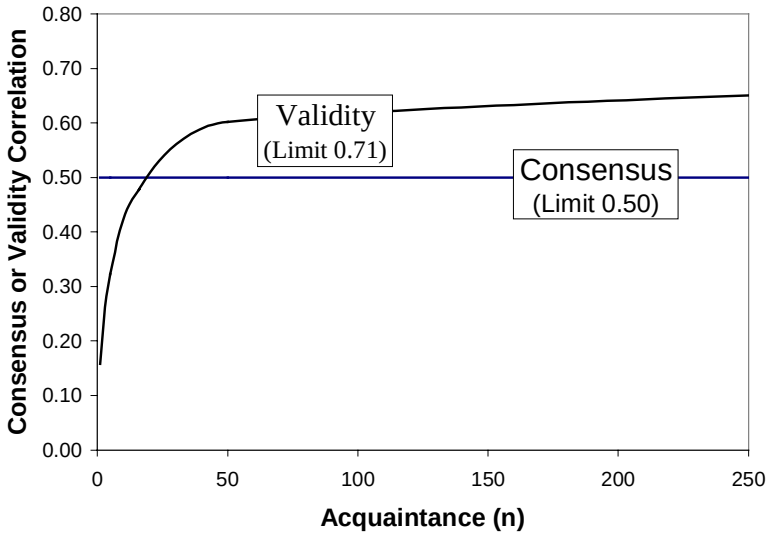


Figure 7: Consensus and validity as a function of acquaintance. For consensus, no communication between raters (a = 0) and 100% overlap (q = 1.0). Similar meaning systems (r_2) equals 0.50 and limits the level of consensus. The square root of r_2 limits the level of accuracy. The within rater consistency factor (r_1) equals 0.05. All other model parameters are assumed to have a value of zero.

The theoretical model has implications for the variance partitioning that is performed in the statistical model (SRM). Stereotypes that are unique to a rater but apply to all ratees are reflected in SRM's rater effect. The level of consensus is reflected in SRM's ratee effect. As discussed in the previous sections, consensus (and therefore ratee effect) is attributable to overlap in the observations, similar meaning systems between raters, consistency within a rater (includes consistency between acts), agreement about stereotypes, and communication between raters. Finally, the rater-ratee interaction (referred to as relationship effect in this research) is attributable to unique impressions, lack of similar meaning systems, and the lack of overlap in observations. The rater by ratee interaction also captures ratings that are unique because of friendships that may exist. The SRM model estimates the correlation between pairs of raters. For example, what is the relationship between how rater A rates B and rater B rates A for a given variable. This correlation is used as an indication of the level of friendship bias in the ratings [15].

III. Results

The results are organized by each of the research questions: user reaction to peer evaluations, consensus, bias, reliability and stability, and validity.

A. User Reaction

Peer evaluations in self-directed groups are generally accepted as fair (Figure 8) and a majority of students individually benefited from the process (Figure 9) or felt the feedback made a difference (Figure 10). From the confidential interviews, there was a clear pattern of concern over the use of peer evaluations to determine a part of the course grade. Most of the reported use

of peer evaluations in small interdependent group work was for evaluative purposes. The peer evaluations were not a part of the course grade in this research. In a significant study, Farh, Cannella, and Bedeian [19] reported that purpose of the evaluation had an impact on user acceptance and peer ratings. Ratings for developmental purposes had greater user acceptance and greater reliability and validity. This fact seems to have gone unnoticed by educators using peer evaluations to determine grades for work in small interdependent group work. This research reiterates the concern in using peer evaluations for evaluative purposes.

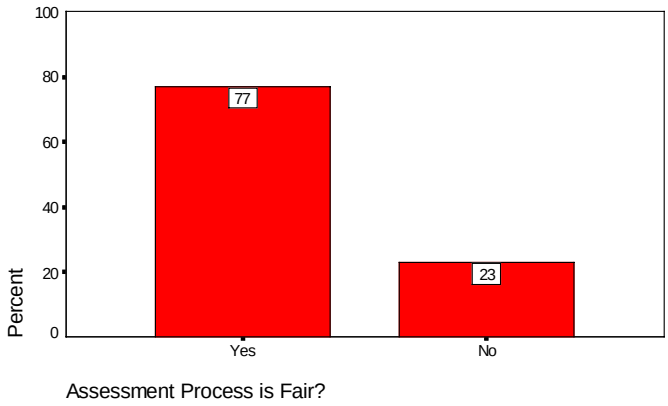


Figure 8: Survey Question 1, is the peer assessment process fair? 48/49 Responded.

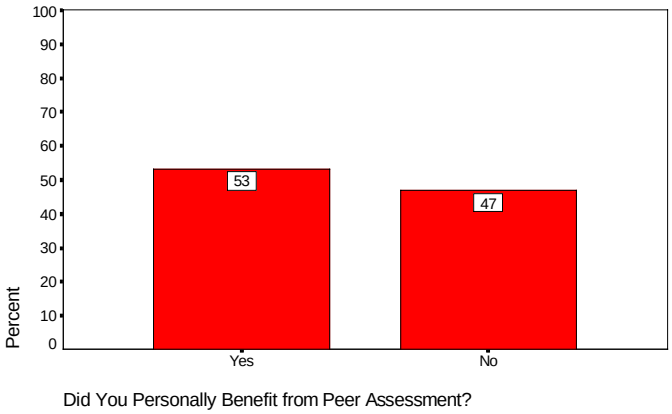


Figure 9: Survey Question 2a, did you personally benefit from process? 47/49 responded

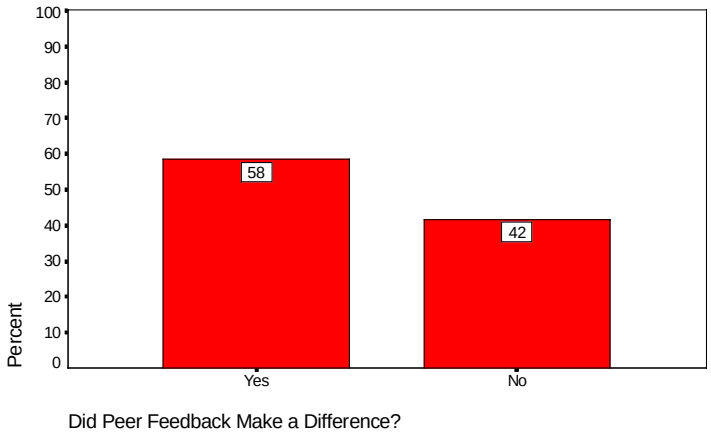


Figure 10: Survey Question 2b, did the feedback make any difference? 48/49 responded.

B. Consensus

The level of consensus in the peer evaluations is indicated by the proportion of variance that is due to the ratee effect. The relative ratee effects were generally small and ranged from 12% to 36% for the four peer evaluations collected during the semester. The relative ratee effect averaged over the four evaluations was 23%. Hennen [13] investigated peer appraisals in simulated self-managing work groups using the SRM model. Undergraduate students rated each other four times. One evaluation took place at the beginning of class before group work commenced. The other evaluations were completed at the end of three working sessions. The group task was to "manufacture" words and package them into sentences. A "raw material" word or phrase was provided from which new words and phrases were developed. The group task required cooperation among the group members. The average ratee effect derived from data reported in the Hennen study was 18%. This average is based on the three evaluations conducted at the end of each working session. Sullivan [20] used the SRM for small cooperative groups working on course-oriented assignments. The average relative ratee effect for two task related behaviors derived from the reported results was 25%. Thus, based on these studies, the level of consensus in peer evaluations of small self-directed group work is relatively low. The Hennen research is the only study found that provides a direct basis for comparison to the research in this study.

Another comparison can be made using the interpretation that the relative ratee effect is the "predicted correlation between pairs of different raters rating the same ratees" [8, p. 31]. If the same raters rate all of the ratees, the correlation is estimated by the ratio of the ratee variance to the sum of ratee variance plus the relationship variance [8, Footnote p. 34]. Using these assumptions, average inter-rater reliability for this study would be 0.34. This compares to 0.34 and 0.41 reported by Ohland and Layton [21]. Ohland and Layton applied a nested single-facet generalizability study (G-study) design [22] to estimate inter-rater reliability for two peer evaluation instruments. The G-study results are an "estimate of how well a single rater's score approximates the true score that would be obtained if enough raters evaluated each student" [21, p. 3].

Using these two related comparisons, the ratee effects reported in this study are consistent with findings in the literature. *These findings indicate that peer evaluations have relatively low levels of consensus.* I believe that the parameter similar meaning systems across raters is one of the factors that limits the level of consensus for the self-directed work teams in this research. Acquaintance (number of observations) and overlap (percentage of observations in common) are probably not significant in limiting the level of consensus. Stable levels of consensus were reported by Hennen [13] after one hour of group work. Longitudinal results reviewed by Kenny [8] also suggest that only low levels of acquaintance were required to reach the limiting level of consensus. The theoretical model indicates that overlap is not a significant factor when acquaintance levels are high which is the case in this research. Lack of overlap in observations on some teams is a possible explanation for the reported high values of relationship variance. Overlap may have been low on some teams because of the tendency to delegate sequential tasks and to not integrate the results with interpretations derived from other sources. An example would be accepting a geologic interpretation (map) based on static information without the benefit of dynamic production information. Written comments after the first project provided evidence that task delegation without follow-up comparisons of information from alternative sources was high on some teams.

The tendency for raters to give similar ratings to each ratee (rater effect) is another reason for the low levels of consensus. The average rater effect over the four evaluations was 38%. A component of the rater effect captures individual rater differences in the expected performance of the group. If raters have an expectation for the group and the expectations differ across raters, the impact is an increase in rater effects. Correlational bias (halo) is another reason for high rater effects. I believe that "expectations" were a significant factor in the rater effects for this research [7]. Further research is needed to confirm this hypothesis.

Unique friendships are another possible explanation for reduced levels of consensus. Unique friendships have the impact of increasing the proportion of variance due to relationship variance. The average relationship effect was 39% for this research. Friendship bias exists in the ratings and is explored in the following section.

C. Bias

The results of this research support the argument that a bias in the peer ratings exists. The confidential individual interviews revealed a clear pattern of friendship as a bias in the ratings. The end of course survey confirmed that a small fraction (16% of the 49 students responding) included some form of bias in their ratings. The strength of the friendship bias, to the extent it is unique between pairs of raters, can be measured by the dyadic (friendship) correlation between relationship effects. The average friendship correlations are relatively low (0.06 to 0.15) with the exception of the end of project evaluations for the first project in which the friendship correlation was 0.41. There was a consistent trend of increasing friendship bias correlations from the mid-point to the end of project evaluation. Also, the relatively low average masks a few high values for some specific team skills. Further research is needed in this area.

There are only a few empirical studies investigating the bias in peer evaluations in small self-directed problem solving teams in education. Layton and Ohland [23] used peer evaluations in

project teams where the majority of the students were African-American. Layton and Ohland did not find a gender bias in the ratings. The finding was consistent with Kaufman, Felder, and Fuller [24]. Kaufman, Felder, and Fuller's study was based on groups working on homework problems. Layton and Ohland, however, reported bias in the ratings based on race. Layton and Ohland hypothesized that "students seem to base ratings on perceived abilities instead of real contributions" [23, p. 6]. This effect would be reflected as a component of the rater effect if the expectation of performance differs across raters. Expected performance, as noted earlier, is believed to be a significant factor of the rater effect in this research. Abson [5] reported a definite bias in the peer ratings for the single group he interviewed. In this case, one peer lowered the ratings for one peer. Falchikov [25] reported that the students felt that the calculation of a final mark was fair and accurately reflected the group work. Mathews [6] in his qualitative study reported some targeting of individuals in the peer rating process. Finally, Sullivan [20] in his cooperative learning groups found a lack of a correlation (ranged from -0.14 to 0.13) between "liking" ratings of individual group members and task related functions. Sullivan used this result to infer that "liking" did not interfere with how individuals assessed skill or attributed responsibility.

In summary, in the literature there was a clear pattern of concern for bias in peer evaluations and evidence that biases exist in small group peer evaluations. These patterns are supported by the results in this study. Friendship was the dominant form of bias in this research.

D. Reliability and Stability of Measures

The SRM statistical model calculates an estimate of the reliability of the mean rater and ratee effects. The reliability represents the percentage of the variation that is attributable to variation in true scores. The reliability coefficient is calculated by taking the ratio of the obtained variance for the effect (rater or ratee) and the expected variance for the effect (rater or ratee). The reliability provides a sense as to whether one can meaningfully interpret the rater and ratee effects for a given variable [15]. The reliability estimates generated by the SRM program are summarized in Tables 5 to 8. The average reliability of the rater and ratee effects ranges from a low of 0.51 to a high of 0.90 indicating a moderate to high level of reliability in the rater and ratee effect estimates.

Table 5
Reliability of Rater and Ratee Effects
Mid-Point First Project

Variable	Rater	Ratee
Back-up	0.56	0.68
Communication	0.72	0.60
Coordination	0.71	0.81
Feedback	0.85	0.84
Leadership	0.70	0.81
Team Orientation	0.74	0.74
Effort	0.60	0.79
Knowledge	0.75	0.83
Average	0.70	0.76

Table 6
Reliability of Rater and Ratee Effects
End First Project

Variable	Rater	Ratee
Back-up	0.54	0.56
Communication	0.62	0.55
Coordination	0.60	0.57
Feedback	0.67	0.66
Leadership	0.50	0.68
Team Orientation	0.74	0.66
Effort	0.15	0.69
Knowledge	0.50	0.74
Average	0.54	0.64

Table 7
Reliability of Rater and Ratee Effects
Mid-Point Second Project

Variable	Rater	Ratee
Back-up	0.90	0.71
Communication	0.91	0.57
Coordination	0.96	0.71
Feedback	0.90	0.47
Leadership	0.87	0.67
Team Orientation	0.93	0.48
Effort	0.85	0.59
Knowledge	0.91	0.72
Average	0.90	0.62

Table 8
Reliability of Rater and Ratee Effects
End Second Project

Variable	Rater	Ratee
Back-up	0.80	0.40
Communication	0.76	0.50
Coordination	0.83	0.54
Feedback	0.82	0.47
Leadership	0.69	0.55
Team Orientation	0.82	0.17
Effort	0.22	0.69
Knowledge	0.85	0.78
Average	0.72	0.51

The SRM model also estimates stable and unstable variance for the rater, ratee, and relationship effects when multiple measures of the same construct are made. For each project, variance stability was determined for the mid-point and end of project evaluations. Stable variance is the variance that replicates across measures. The unstable variance is unique to each measure. In most situations the measures (e.g. communication) provide relatively stable rater and ratee effects [8]. In these cases, the unstable variances for each effect (rater, ratee, and relationship) are added together and commonly referred to as error. The investigator believes that stability as a measure of reliability for rater, ratee, and relationship effects may not be appropriate in the case of self-directed teamwork. This argument is more compelling when there is a large time lapse between ratings as is the case in this research. Teams are reported to go through developmental phases over time [26,27]. Considering the dynamics that occur in self-directed teamwork, I believe that unstable rater, ratee, and relationship variance comprises the contextual nature of teamwork that can take place over time. For this reason, the common practice of classifying the unstable variance as error is not adopted for the reported results in this research.

The results presented in Table 9 demonstrate that the rater, ratee, and relationship effects in this study have a significant component of unstable variance. Hennen [13] provides the only direct comparison for the level of stability found in repeated measures of peer evaluations on self-directed work teams. Hennen reported that ratings recorded at a particular time were "somewhat independent from other time periods and are more directly tied to the performance just observed than to evaluations previously recorded" (p. 73). Later Hennen states that "over time results have been surprising in their small relationship with one another" (p. 87).

E. Validity

The convergent validity of two constructs was measured by comparing criterion values for "knowledge applied to the task" and "effort applied to the task" to ratee effects for each construct. The criterion values were determined by the faculty and TA team using the same instrument used in the peer evaluations. The correlation coefficients between faculty team set criterion and the ratee effect (level of consensus) for "technical knowledge applied to the task" were 0.76 and 0.68 for the end of the first and second projects respectively. The corresponding correlation coefficients for the team skill "effort applied to the task" were 0.71 and 0.61. These values indicate a high level of validity.

Validity is limited by the square root of the correlation for similar meaning systems (r_2). Park, Dekay, and Kraus [28] estimated the value for similar meaning systems to be about 0.40. This would indicate an upper limit for validity of about 0.65. The maximum validity of 0.76 found in this research would imply a correlation for similar meaning systems of about 0.60. Training in the use of the instrument has the effect of increasing similar meaning among raters. Team skills training was conducted and linked to the team skills instrument.

Table 9
Proportion of Variance that is Stable and Unstable for First and Second Project

First Project

	Rater Effect	Ratee Effect	Relationship Effect	Total
Stable Variance	0.10	0.22	0.15	0.47
Unstable Variance	0.13	0.09	0.30	0.53
Total	0.24	0.31	0.45	1.00

Second Project

	Rater Effect	Ratee Effect	Relationship Effect	Total
Stable Variance	0.17	0.09	0.06	0.32
Unstable Variance	0.32	0.05	0.30	0.68
Total	0.50	0.15	0.36	1.00

IV. Limitations

Three potentially significant limitations are discussed in the following sections. The first limitation is based on the specifications of the statistical model, the second is based on the peer evaluation instrument and training, and the third is based on the level of student motivation for completing the peer evaluations.

A. Statistical Model.

The SRM statistical model assumes that only dyads or pairs are involved in forming perceptions for ratings. This assumption is violated if there is communication between raters on how they would rate other team members, on how they planned to rate each other, or on how they planned to rate a specific team member without the knowledge of that team member. The individual interviews indicated a relatively low level of communication between raters. This conclusion was further supported by the final course survey. I believe that there was a high level of independence between raters. There was minimal indication that individuals were targeted for a specific evaluation. This form of communication can mask the level of consensus and invalidate the results.

B. Peer Evaluation Instrument and Training

The instrument used behaviorally anchored scales (BARS) for the peer ratings. BAR scales are reported to be appropriate for peer ratings [29]. Landy and Farr [30] reported reduced leniency, lower response set, and greater reliability when BAR scales were used compared to using adjective anchors. Generalizability of anchors and scales from one setting to another may be a limitation of the BAR scales. Landy and Farr [30] also reported that response set was not changed when a rater rates all ratees on one trait at a time. Kenny [8] reported that there was some "evidence that rating each ratee on all the traits before moving on to the next ratee" (p. 39) may actually be a better approach. A trait format in which raters rated each ratee on one trait at a time was used in this study.

Ambiguity in the anchors is another source of error in the ratings. One student during the confidential interviews stated that "some of the anchors are too general, not enough detail, too vague" (mid project, first project interview).

The lack of anonymity impacted the results in at least one case. This was discovered in the individual interviews conducted for the end of second project. One student stated, "Yeah I wish it were more anonymous. I was too close to my peers this time when I was filling out the assessment, as in sitting right next to each other. I felt really self conscious about being totally honest with my ratings. The other times it was no problem because I was not right next to my team members." During the last peer evaluation, the students were not split into two different rooms as was done in the previous rating sessions.

Finally, training on the use of the instrument is reported to have an impact on the level of consensus and response set in the ratings. Borman [31] reported that short training (approximately five minutes) reduced response set but did not impact the level of validity. Borman also reported somewhat lower reliability after the training. The type of training is also important. Latham, Wexley, and Pursell [32] found that group discussion was effective in reducing response set but that lectures were not effective. The training in the use of the instrument in this class lasted approximately two hours and included a discussion of the constructs. Before each rating session, the students were reminded that the average rating for the skill "effort applied to the task" should be 3.0. No other comments were made. Training has the effect of increasing similar meaning systems (r_2) as defined in this research.

C. Motivation

Response set is likely to be greater when the motivation to respond carefully is low. The only motivation for completing the ratings carefully was the emphasis placed on the value of the information to team members to improve team skills. There was an increase in response set noted in the first ratings for the second project. These ratings were made two weeks after spring break when the student motivation was perceived by the faculty and TA team to be low. Most students are graduating seniors and for them, the end is in sight. In this course, most students are relieved that 50% of the course work is completed and that graduation is only a few weeks away.

V. Conclusion

The level of consensus of peer evaluations in interdependent self-directed group work was relatively low. The average proportion of the variance due to consensus (ratee effects) was 23%. The rater's expectation of group performance and response set (rater effects) accounted for approximately 38% of the total variance. I believe that the rater's expectation of group performance accounted for a significant proportion of the rater effects but further research is needed to confirm this hypothesis. Relationship effects (rater by ratee interaction) accounted for 39% of the total variance. This component of the variance captures the unique relationships that develop on the interdependent teams including biases that may exist at the dyad level. There was a clear pattern of friendship bias in the peer evaluations. The average bias correlation was relatively small with the exception of one rating period.

The timing of the peer evaluations is believed to impact the level of consensus and the proportion of the variance that is due to rater effects and relationship effects. This conclusion highlights the importance of using peer evaluations for developmental purposes rather than evaluative purposes. Group membership changed from the first project to the second project with the exception of two individuals. The variance partitioning did not generalize from the first to the second projects. This offers further support for limiting peer evaluations to developmental applications.

The convergent validity for "effort applied to the task" and "technical knowledge applied to the task" was high. The correlation coefficients between faculty team set criterion and the level of consensus (ratee effect) for "technical knowledge applied to the task" were 0.76 and 0.68 for the end of the first and second projects respectively. The corresponding correlation coefficients for the team skill "effort applied to the task" were 0.71 and 0.61. Thus, even though the level of consensus was relatively low, the validity of the ratings was high.

Finally, user reaction to the peer evaluations was generally positive. A majority of the students indicated that they benefited from the process. Targeting of specific individuals by team members was limited. Friendship bias was a factor in the peer evaluations but was not a significant factor in reducing the reliability or validity of the measurements. Peer evaluations in small self-directed interdependent teams are a good practice in higher education. Emphasis should be placed on developmental uses of the ratings rather than evaluative. Unfortunately, the use of peer evaluation for evaluative purposes is common practice in higher education.

VI. Implications for Practice

The use of peer evaluations for developmental purposes on self-directed interdependent work teams is good practice. The reported high validity, positive benefits, and low agreement between measures of the same characteristic made at different points in time support this argument. High validity suggests that peer evaluations are a good source of feedback information for improving a student's team skills. The individual interview results suggest that the benefits of the peer assessment and feedback process outweigh any potential negative aspects. Generally, students benefited or were neutral to the process. Finally, the relatively low consistency in the measures over time offers additional support for limiting the use of peer evaluations to developmental

purposes. I hypothesize that the low consistency in the measures reported over time is the result of the contextual nature of the teamwork ratings. Ratings at any given point in time more closely reflect the current stage of team development and the current activities of the team.

The usage of peer evaluations for evaluative purposes is not recommended. Although the peer evaluations were not used as a formal part of an individual's grade in this research, confidential interviews confirmed a clear pattern of concern over the use of peer evaluations to determine a component of an individual's grade. This is especially important considering the variability over time found in the peer evaluations. Peer evaluations are believed to be a function of the stage of team development and directly tied to the performance just observed. Thus a student's grade would be dependent on the timing of the peer evaluation and may not reflect overall team performance. Finally, confidential interviews revealed a clear concern over friendship as a bias in the peer evaluations. Biases have the impact of reducing the reliability of the evaluations. There was evidence that biases existed in the peer evaluations with friendship being the dominant form of bias.

VII. Further Research

The following have been identified as topics for further research:

1. What is the impact of input medium (paper and pencil versus computerized entry) on peer evaluation reliability and validity?

2. How much of the rater effect is due to expectations of performance versus correlational bias?

3. What is the relationship between self-evaluations and peer evaluations? For example, are individuals who see themselves as team leaders (self-evaluation) seen as team leaders (peer evaluations)?

4. Do students that "contribute more than their fair share" give lower ratings?

5. Are certain team skills more prone to bias than others?

6. What is the relative validity of peer and self-evaluations?

VIII Acknowledgement

Thanks to the facultly team (John B. Curtis, Tom Davis, Max Peeters), and teaching assistants (Jennifer Miskimins, Mike Sherwood) for the Multidisciplinary Petroleum Design Course. Also thanks to my thesis advisor (Laura Goodwin), faculty at the University of Colorado at Denver (Brent Wilson and Rodney Muth), and Ruth Streveler (Director for the Center for Engineering Education at the Colorado School of Mines) for their help in completing this reaserch. Finally, thanks to Carole Edwards-Knight for conducting the confidential interviews and helping with the peer evaluation and survey instruments.

APPENDIX A

TEAM SKILLS INSTRUMENT

The instrument for the following team skills are presented:

- Back-up Behavior
- Communication
- Coordination
- Feedback
- Team Leadership
- Team Orientation
- Effort Applied to Task
- Technical Knowledge Applied to Task

The team skills instrument was set-up on mail merge so that the evaluator name, team number, and evaluatee names were automatically printed on the form. A single page was used for each team skill with the four or five evaluatees listed below the anchors on each page. The order of the evaluatees on each page was the same for a given team. Thus, each participant was provided a package of eight pages printed on 8 ½ by 14 inch paper.

Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Backup Behavior

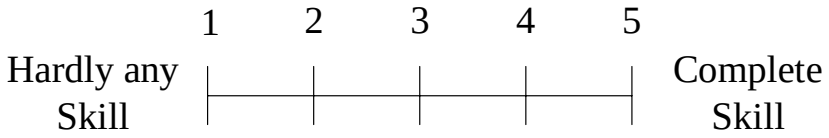
Backup Behavior involves assisting the performance of other team members. This implies that members have an understanding of other members' tasks. It also implies that members are willing and able to provide and seek assistance.

The reference to subject member in the following refers to the team member being evaluated.

Complete Skill	5. Subject **always** provides assistance without neglecting own tasks. Subject always seeks assistance rather than struggle.
High Skill	4. Subject **almost always** provides assistance and rarely neglects own tasks. Subject almost always seeks assistance rather than struggle.
Adequate Skill	3. Subject **usually** provides assistance but waits until asked. Subject usually struggles before seeking help.
Some Skill	2. Subject member **sometimes** provides assistance, but tends to neglect own tasks when helping others. Sometimes seeks assistance after struggling and making mistakes.
Hardly Any Skill	1. Subject member **consistently fails** to provide assistance or if he/she does provide assistance, tends to neglect own tasks. Subject is unwilling to ask for help even when available.

Place an X on the following scale:

Evaluatee:
«evaluatee»



Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Communication

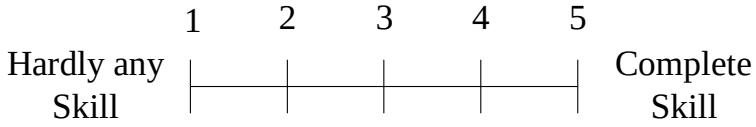
Communication involves the exchange of information between two or more team members using proper terminology. The purpose of communication is to clarify or acknowledge the receipt of information, to initiate and elicit a response, and to engage in dialogue in an attempt to understand alternative perspectives, to propose critical information and data needs, to propose decision criteria, and to make recommendations.

The reference to subject member in the following refers to the team member being evaluated.

Complete Skill	5. Subject member **always** passes on important information, clarifies intentions and plans, and engages in dialogue on team tasks.
High Skill	4. Subject member **almost always** passes on important information, clarifies intentions and plans, and engages in dialogue on team tasks.
Adequate Skill	3. Subject member **usually** passes on important information, clarifies intentions and plans, and engages in dialogue on team tasks.
Some Skill	2. Subject member **sometimes** passes on important information, clarifies intentions and plans, and engages in dialogue on team tasks.
Hardly Any Skill	1. Subject member **consistently fails** to pass on important information, clarify intentions and plans, and engage in dialogue on team tasks.

Place an X on the following scale:

Evaluatee:
«evaluatee»



Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Coordination

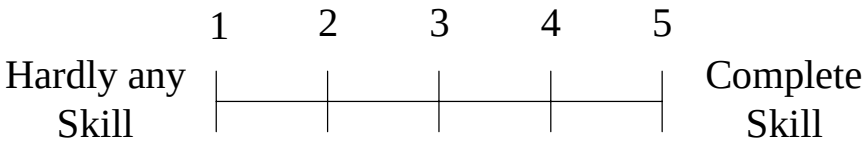
Coordination refers to team members' executing their activities in a timely and integrated manner. It implies that the performance of some team members influences the performance of other team members. This may involve an exchange of information that subsequently influences another members' performance.

The reference to subject member in the following refers to the team member being evaluated.

Complete Skill	5. Subject member **always** carries out tasks in a timely manner, enabling team to accomplish tasks.
High Skill	4. Subject member **almost always** carries out tasks in a timely manner, enabling team to accomplish tasks.
Adequate Skill	3. Subject member **usually** works in synchrony and usually carries out tasks in timely manner.
Some Skill	2. Subject member **sometimes** misses deadlines or carries out task ineffectively leading to delay.
Hardly Any Skill	1. Subject member **consistently** carries out tasks ineffectively or unpredictably, leading to the delay or failure of other team members in executing their own tasks.

Place an X on the following scale:

Evaluatee: «evaluatee»



Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Feedback

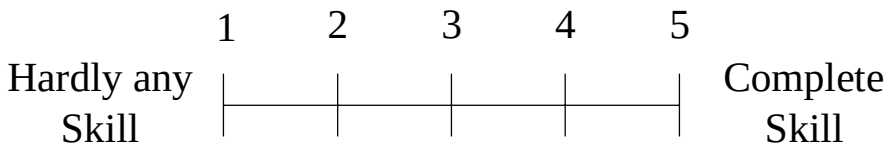
Feedback could relate to group process or to task-related issues. Feedback involves the giving and receiving of information among members. Giving feedback refers to providing information regarding other members performance. Receiving feedback refers to accepting positive and negative information regarding performance.

The reference to subject member in the following refers to the team member being evaluated.

Complete Skill	5. Subject member **consistently** provides information regarding other members performance and identifies mistakes and how to correct them. Subject member listens and incorporates suggestions for own improvement.
High Skill	4. Subject member **almost always** provides specific information regarding other members performance. Subject member almost always listens and incorporates suggestions for own improvement.
Adequate Skill	3. Subject member **usually** gives other members general comments rather than specific advice on performance. When given suggestions from other members, subject usually listens patiently and generally takes advice.
Some Skill	2. Subject member **sometimes** makes sarcastic comments when tasks do not go as planned, tends to resist asking for advice. Subject member usually rejects suggestions offered by other team members.
Hardly Any Skill	1. Subject member **makes sarcastic comments** when tasks do not go as planned, resists asking for advice. Subject member **consistently rejects suggestions** offered by other team members.

Place an X on the following scale:

Evaluatee:
«evaluatee»



Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Team Leadership

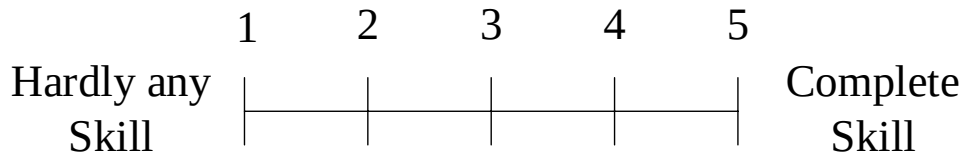
Team Leadership involves providing direction, structure, and support for other team members. It does not necessarily refer to a single individual with formal authority over others. Team Leadership can be demonstrated by several team members

The reference to subject member in the following refers to the team member being evaluated.

Complete Skill	5. Subject member **always** reviews the situation and facilitates dialogue when other members are not contributing, helps resolve conflicts and keeps team focused.
High Skill	4. Subject member **almost always** reviews the situation, almost always facilitates the dialogue, and almost always keeps the team focused on the task.
Adequate Skill	3. Subject member **sometimes** reviews the situation, sometimes facilitates the dialogue, and sometimes keeps the team focused on the task.
Some Skill	2. Subject member **rarely** reviews the situation, rarely facilitates the dialogue, and rarely keeps the team focused on the task.
Hardly Any Skill	1. Subject member **does not** help facilitate the team process, does not help resolve conflict, does not help the team focus.

Place an X on the following scale:

Evaluatee: «evaluatee»



Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Team Orientation

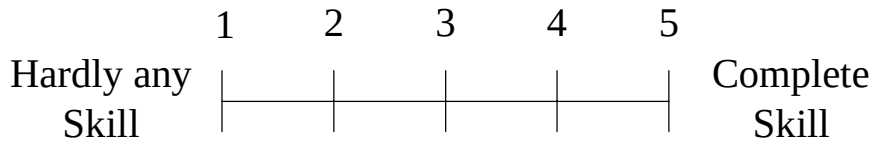
Team Orientation refers to the attitudes that team members have towards one another and the team task. It reflects acceptance of team norms, level of group cohesiveness, and importance of team membership.

The reference to subject member in the following refers to the team member being evaluated.

Complete Skill	5. Subject member **always** pulls together for a common team goal, displays pride in performing tasks and trust toward other members.
High Skill	4. Subject member **almost always** pulls together for a common team goal, almost always displays pride in performing tasks and trust toward other members.
Adequate Skill	3. Subject treats team performance issues as important; when necessary, member **usually** pulls together and places team goals ahead of personal interests.
Some Skill	2. Subject member **rarely places** personal goals and interest ahead of those of the team; some lack of trust and cooperation evident.
Hardly Any Skill	1. Subject member **places** personal goals and interest ahead of those of the team; lack of trust and cooperation evident.

Place an X on the following scale:

Evaluatee:
«evaluatee»



Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Effort Applied to Task

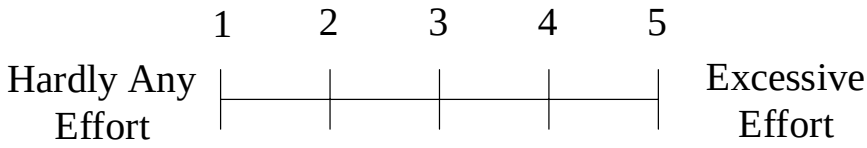
Effort Applied to Task refers to the effort team members are willing to commit to complete individual and assigned team tasks.

The reference to subject member in the following refers to the team member being evaluated.

Excessive Effort	5. Subject member did most of work that could have been done by other team members because one or more members were unwilling to do their fair share.
	4. Subject member did some work that could have been done by other team members because one or more members were unwilling to do their fair share.
Balanced Effort	3. Subject member **did his/her fair share** of the work.
	2. Subject member did less than his/her fair share of the work. The subject member let others do some of the work necessary to complete the team project.
Hardly Any Effort	1. Subject member did hardly any work. The subject member let others do the work necessary to complete the team project.

Place an X on the following scale:

Evaluatee:
«evaluatee1»



Team Skill Measures
Peer and Self Evaluation

Evaluator: ***«evaluator»*** *Team:* ***«Team»*** *Date:*

Technical Knowledge Applied to Task

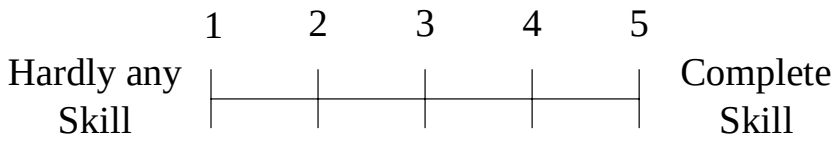
Technical Knowledge Applied to Task refers to the technical knowledge that the team member is capable of **and** brings to solving the problem.

The reference to subject member in the following refers to the team member being evaluated.

Complete Skill	5. Subject member brings a **high** level of technical skill to solving the task.
High Skill	4. Subject member brings a **more than adequate** level of technical skills to solving the task.
Adequate Skill	3. Subject member brings an **adequate** level of technical skills to solving the task.
Some Skill	2. Subject member brings a **low** level of technical skills to solving the task.
Hardly Any Skill	1. Subject member **does not bring** technical skills to solving the task.

Place an X on the following scale:

Evaluatee:
«evaluatee»



APPENDIX B

STUDENT INTERVIEW QUESTIONS

Mid-Point First Project

1. Do you think the peer assessment process was fair? Why or why not? Please give examples if applicable.

2. From your perspective, have you benefited from the peer assessment process? If so, how? If not, how could it be changed to be more beneficial?

3. When you rated your team members on team skills, do you think that factors such as friendship, physical appearance, gender, country of origin, academic discipline, group membership or other factors not listed in the team skills scales influenced any of your ratings of team members on the team skills scales? If so, please explain which factors influenced you and in what direction (raised or lowered ratings).

4. Did you talk to other team members about how you rated or planned to rate each other? Did you consult with any other team members on how to rate a specific team member without their knowledge?

5. Do you think that the task performance strategy was helpful to you or your team in any way? If so, how? If not, how could it be improved?

6. Did you learn anything from the integration of information within your discipline and/or across disciplines? If so, please give a brief summary.

7. Do you think this course will help you become better prepared for professional practice? If so, how? If not, how could it be improved?

STUDENT INTERVIEW

End First Project

1. Do you think the peer assessment process is fair? Why or why not? Please give examples if applicable.

2. From your perspective, have you benefited from the peer assessment process? If so, how? Did the feedback make any difference? If not, how could the process be changed to be more beneficial?

3. When you rated your team members on team skills, do you think that factors such as physical appearance, gender, country of origin, academic discipline, friendship, group membership or other factors not listed in the team skills scales influenced any of your ratings of team members on the team skills scales? If so, please explain which factors influenced you and in what direction (raised or lowered ratings).

4. Did you talk to other team members about how you rated or planned to rate each other? Did you consult with any other team members on how to rate a specific team member without their knowledge?

5. Do you think that the task performance strategy was helpful to you or your team in any way? If so, how? If not, how could it be improved?

6. Did you learn anything from the integration of information across disciplines? If so, please give a brief summary.

7. Do you think this course will help you become better prepared for professional practice? If so, how? If not, how could it be improved?

8. How would you rate the overall performance of your team? What could your team have done differently to improve its performance?

STUDENT INTERVIEW

Mid-Point Second Project

1. Do you think the peer assessment process is fair? Why or why not? Please give examples if applicable. Was it premature in the project?

2. From your perspective, have you benefited from the peer assessment process? If so, how? Did the feedback make any difference? If not, how could the process be changed to be more beneficial?

3. When you rated your team members on team skills, do you think that factors such as physical appearance, gender, country of origin, academic discipline, friendship, group membership or other factors not listed in the team skills scales influenced any of your ratings of team members on the team skills scales? If so, please explain which factors influenced you and in what direction (raised or lowered ratings).

4. Did you talk to other team members about how you rated or planned to rate each other? Did you consult with any other team members on how to rate a specific team member without their knowledge?

5. Do you think that the task performance strategy was helpful to you or your team in any way? If so, how? If not, how could it be improved?

6. Did you learn anything from the integration of information across disciplines? If so, please give a brief summary.

7. Do you think this course will help you become better prepared for professional practice? If so, how? If not, how could it be improved?

8. How would you rate the overall performance of your team? What could your team have done differently to improve its performance?

9. Is this team functioning better than the last one?

STUDENT INTERVIEW

End Second Project

1. Do you think the peer assessment process is fair? Why or why not? Please give examples if applicable.

2. From your perspective, have you benefited from the peer assessment process? If so, how? Did the feedback make any difference? If not, how could the process be changed to be more beneficial?

3. When you rated your team members on team skills, do you think that factors such as physical appearance, gender, country of origin, academic discipline, friendship, group membership or other factors not listed in the team skills scales influenced any of your ratings of team members on the team skills scales? If so, please explain which factors influenced you and in what direction (raised or lowered ratings).

4. Did you talk to other team members about how you rated or planned to rate each other? Did you consult with any other team members on how to rate a specific team member without their knowledge?

5. Do you think that the task performance strategy was helpful to you or your team in any way? If so, how? If not, how could it be improved?

6. Did you learn anything from the integration of information across disciplines? If so, please give a brief summary.

7. Do you think this course will help you become better prepared for professional practice? If so, how? If not, how could it be improved?

8. How would you rate the overall performance of your first team? Poor, good or excellent?

9. Is this team functioning better than the last one?

10 How would you rate the overall performance of your first team? Poor, good or excellent? What could your team have done differently to improve its performance?

11. Did you like having multiple projects?

APPENDIX C

END OF COURSE SURVEY QUESTIONS

Please circle the answer or number that most closely represents your response or opinion.

1. Overall, do you think the peer assessment process is fair? **YES NO**

2 a) From your perspective, have you benefited from the peer assessment process? **YES NO**

2 b) Did the feedback make any difference? **YES NO**

3. For any of the rating sessions, when you rated your team members on team skills, do you think that factors such as physical appearance, gender, country of origin, academic discipline, friendship, group membership or other factors not listed in the team skills scales influenced any of your ratings of team members on the team skills scales? **YES NO**

4 a) For any of the rating sessions, did you talk to other team members about how you rated or planned to rate each other? **YES NO**

4 b) For any of the rating sessions, did you consult with any other team members on how to rate a specific team member without their knowledge? **YES NO**

5. Overall, do you think that the task performance strategy was helpful to you or your teams in any way? **YES NO**

6. Overall, did you learn anything from the integration of information across disciplines? **YES NO**

7. Do you think this course will help you become better prepared for professional practice? **YES NO**

8. How would you rate the overall performance of your first team?

POOR		**GOOD**		**EXCELLENT**
1	**2**	**3**	**4**	**5**

9. Did your second team function better than your first one? **YES NO**

10. How would you rate the overall performance of your second team?

POOR		**GOOD**		**EXCELLENT**
1	**2**	**3**	**4**	**5**

Bibliography

[1]Cliff Goodwin and Rob Wolter, "Student work group/teams: Current practices in an engineering and technology curriculum compared to models found in team development literature," presented at the ASEE 1998 Annual Technical Conference, Seattle, WA, 1998 (unpublished).

[2]Jack McGourty, Peter Dominick, and Richard R. Reily, "Incorporating student peer review and feedback into the assessment process," presented at the Frontiers in Education 1998, Tempe, AZ, 1998 (unpublished).

[3]Robert S. Thompson, "Repeated measures design for assessment of critical team skills in multidisciplinary teams," presented at the 2000 ASEE Conference and Exposition, St. Louis, MO, 2000 (unpublished).

[4]J. Arvid Andersen, "Assessment techniques used in multidisciplinary and cross-cultural student teamwork, Session 1360," presented at the 2000 ASEE Annual Conference and Exposition, St. Louis, 2000 (unpublished).

[5]D. Abson, "The effects of peer evaluation on the behavior of undergraduate students working on tutorless groups," in *Group and Interactive Learning*, edited by H. C. Foot, C. J. Howe, A. Anderson *et al*. (Computational Mechanics, South Hampton, England, 1994), Vol. 1, pp. 153-158.

[6]Brian Mathews, "Assessing individual contributions: Experience of peer evaluation in major group project," British Journal of Educational Technology **25** (1), 19-28 (1994).

[7]Robert S. Thompson, "Reliability, validity, user acceptance, and bias in peer evaluations on self-directed interdependent work teams," Doctoral Dissertation, University of Colorado at Denver, 2000.

[8]David A. Kenny, *Interpersonal perception: A social relations analysis* (The Guilford Press, New York, 1994).

[9]D. A. Kenny and L. La Voie, "The social relations model," in *Advances in experimental social psychology*, edited by L. Berkowitz (Academic Press, Orlando, FL, 1984), Vol. 18, pp. 142-182.

[10]R. M. Warner, D. A. Kenny, and M. Stoto, "A new round robin analysis of variance for social interaction data," Journal of Personality and Social Psychology **37**, 1742-1757 (1979).

[11]Laura D. Goodwin and William L. Goodwin, "Research notes: Using generalizability theory in early childhood special education," Journal of Early Intervention **15** (2), 193-204 (1991).

[12]Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda *et al*., *The dependability of behavioral measurements: Theory of generalizability for scores and profiles* (John Wiley & Sons, New York, 1972).

[13]Martha E. Hennen, *Consensus and meta-accuracy in self-managing work groups: A social relations analysis* (University of Connecticut, Storrs, CT, 1996).

[14]Barbara M. Montgomery, "An interactionist analysis of small group peer assessment," Small Group Behavior **17** (1), 19-37 (1986).

[15]D.A. Kenny, R. Lord, and S. Garg, *A social relations analysis of peer ratings* (Unpublished paper, University of Connecticut, 1983).

[16]R. G. Lord, J. S. Phillips, and M. C. Rush, "Effects of sex on perceptions of emergent leadership, influence, and social power," Journal of Applied Psychology **65**, 176-182 (1980).

[17]David Kenny, "A general model of consensus and accuracy in interpersonal perception," Psychological Review **98**, 155-163 (1991).

[18]N. Anderson, *Foundations of information integration theory* (Academic Press, New York, 1981).

[19]Jiing-Lih Farh, Albert A. Jr. Cannella, and Arthur G. Bedeian, "Peer ratings: The impact of purpose on rating quality and user acceptance," Group &Organization Studies **16** (4), 367-386 (1991).

[20]Michael P. Sullivan and Raymond R. Reno, "Perceiving groups accurately," Group Dynamics: Theory, Research, and Practice **3** (3), 196-205 (1999).

[21]Matthew W. Ohland and Richard A. Layton, "Comparing the reliability of two peer evaluation instruments," presented at the ASEE 2000 Annual Technical Conference, St. Louis, MO, 2000 (unpublished).

[22]Linda Crocker and James Algina, *Introduction to classical and modern test theory* (Harcourt Brace Jovanovich, Fort Worth, 1986).

[23]Richard A. Layton, "Peer evaluations in teams of predominantly minority students," presented at the ASEE 2000 Annual Technical Conference, St. Louis, MO, 2000 (unpublished).

[24]Deborah B. Kaufman, Richard M. Felder, and Hugh Fuller, "Peer ratings in cooperative learning groups," presented at the 1999 ASEE Annual Conference and Exposition, Charlotte, NC, 1999 (unpublished).

[25]Nancy Falchikov, "Self and peer assessment of a group project designed to promote the skills of capability," Programmed Learning & Educational Technology **25** (4), 327-339 (1988).

[26]C.J.G. Gersick and J.R. Hackman, "Habitual routines in task-performing groups," Organizational Behavior and Human Decision Processes **47**, 65-97 (1990).

[27]J. Richard Hackman, "Groups that work (and those that don't)," in *The Jossey-Bass Management Series* (Jossey-Bass, San Francisco, 1990), pp. 504.

[28]B. Park, M. L. DeKay, and S. Kraus, "Aggregating social behavior into person models: Perceiver-induced consistency," Journal of Personality and Social Psychology **66**, 437-459 (1994).

[29]J. S. Kane and E.E. Lawler, "Method of peer assessment," Psychological Bulletin **85**, 555-586 (1978).

[30]F.J. Landy and J.L. Farr, "Performance rating," Psychological Bulletin **87**, 72-107 (1980).

[31]Walter C. Borman, "Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings," Journal of Applied Psychology **60** (5), 556-560 (1975).

[32]G.P. Latham, K.N. Wexley, and E.D. Pursell, "Training managers to minimize rating errors in the observations of behavior," Journal of Applied Psychology **60**, 550-555 (1975).

ROBERT S. THOMPSON

Robert Thompson is an Associate Professor in Petroleum Engineering at the Colorado School of Mines (CSM). Robert received his professional degree in petroleum engineering from CSM, his MBA from the University of Houston, and his Ph.D. in Educational Leadership from the University of Colorado at Denver.