Research Brief

Effects of Behavioral Anchors on Peer Evaluation Reliability

MATTHEW W. OHLAND
General Engineering
Clemson University

RICHARD A. LAYTON
Mechanical Engineering
Rose-Hulman Institute of Technology

MISTY L. LOUGHRY
Management
Clemson University

AMY G. YUHASZ
General Engineering
Clemson University

ABSTRACT

This paper presents comparisons of three peer evaluation instruments tested among students in undergraduate engineering classes: a single-item instrument without behavioral anchors, a ten-item instrument, and a single-item behaviorally anchored instrument. Studies using the instruments in undergraduate engineering classes over four years show that the use of behavioral anchors significantly improves the inter-rater reliability of the single-item instrument. The inter-rater reliability (based on four raters) of the behaviorally anchored instrument was 0.78, which was not significantly higher than that of the ten-item instrument (0.74), but it was substantially more parsimonious. The results of this study add to the body of knowledge on evaluating students' performance in teams. This is critical since the ability to function in multidisciplinary teams is a required student learning outcome of engineering programs.

Keywords: peer evaluation, assessment, behaviorally anchored rating scale

I. INTRODUCTION

In this paper, we compare the inter-rater reliability of three peer-evaluation instruments when the instruments are used to adjust team members' grades based on the ratings of their contributions to the team. The research setting involves project teams comprised of junior-level engineering students. Our results show that adding behavioral anchors and descriptive instructions to a one-item instrument significantly increases instrument reliability and that a one-item behaviorally anchored instrument has inter-rater reliability as high as that of a ten-item unanchored instrument.

II. LITERATURE REVIEW

1) Teamwork in engineering courses: In recent years, there has been a great deal of engineering education research aimed at evaluating teamwork. This is driven both by engineering's industrial stakeholders and accreditation standards. ABET's EC2000 Criterion 3, outcome (d) is "an ability to function on multi-disciplinary teams" [1]. Although there has been debate about how to apply the term "multi-disciplinary," the ability to function on a team is central to this outcome.

Many engineering professors incorporate teamwork into their courses not only because employers and accrediting bodies look for these skills, but also because they value team-based educational methods. Advocates of cooperative learning methods believe that the best way for students to achieve the learning objectives in their courses is to work in learning teams. Many studies have shown that when correctly implemented, cooperative learning improves information acquisition and retention, and enhances higher-level thinking skills, interpersonal and communication skills, and self-confidence [2]. Cooperative learning is an instructional paradigm wherein teams of students work on structured tasks (e.g., homework assignments, laboratory experiments, or design projects) under conditions that meet five criteria: positive interdependence, individual accountability, face-to-face interaction, appropriate use of collaborative skills, and regular self-assessment of team functioning [3]. In addition to creating individual accountability, the average of team members' peer ratings can be used as a self-assessment of team functioning. The instruments presented here do not measure the other criteria for successful cooperative learning.

2) Formative vs. summative assessment: Peer evaluations have been administered to engineering student teams in one of two ways: formative assessment [4] or summative assessment [5]. Formative assessments (such as *Team Developer*™) [4] are used to provide feedback to students in order to help them improve their teamwork skills. Therefore, they should provide specific information about which student behaviors are effective and ineffective. Longer, more detailed evaluation instruments can be appropriate for formative assessment because the targeted feedback helps students to understand what they are doing well and in what areas they need to improve. Research has found that peer raters often provide better, more accurate feedback when the peer reviews do not affect the ratees' rewards, suggesting that instructors carefully evaluate the benefits and drawbacks of using peer evaluations to adjust students' grades if the main objective is formative assessment [6].

The research reported here uses peer ratings for summative assessment and adjusts students' grades based on their peer ratings. Summative assessments are used to describe a person's past performance to a team. When summative assessments are tied to a reward system, such as students' grades, they have the potential to motivate team members to behave in ways that earn them higher ratings. Many instructors use summative peer evaluations to adjust students' team grades in order to achieve the individual accountability that is necessary for successful cooperative learning environments. Peer evaluations have the potential to make individuals' contributions to the team's work more identifiable, which reduces the tendency of people to contribute less effort to group tasks than they do to individual tasks [7]. Although individual contributions to the team could be measured in other ways (such as by instructor observation), peer evaluation is often the most appropriate method. Another benefit of using peer evaluations is that the process of completing them helps students understand the performance criteria and how everyone will be evaluated [7]. Millis and Cottell argue strongly that instructors should use peer evaluations to adjust students' team grades, again emphasizing the importance of individual accountability [3].

3) Peer evaluation validity issues: Several meta-analytic studies in the human resource management literature have examined the inter-rater reliability of peer ratings and their correlations with other rater sources such as supervisors, self-ratings, and ratings by subordinates [8–11]. These studies have found that peer ratings are positively correlated with other rating sources and have good predictive validity for various performance criteria. Peer and self ratings had a correlation, corrected for measurement unreliability, of 0.36 in the Harris and Schaubroeck meta-analysis [10], and 0.31 in the Conway and Huffcutt meta-analysis [8]. Although Kaufman, Felder, and Fuller did not report the correlation coefficient of self and peer ratings, they observed little difference between the average peer rating and the average self rating, and they propose that the self ratings thus did not affect substantially the grade adjustment. Brown makes a case for including self ratings in the rating pool, observing that students are surprisingly honest when rating themselves [12]. Typical point-distribution peer evaluation methods implicitly include a self rating [13].

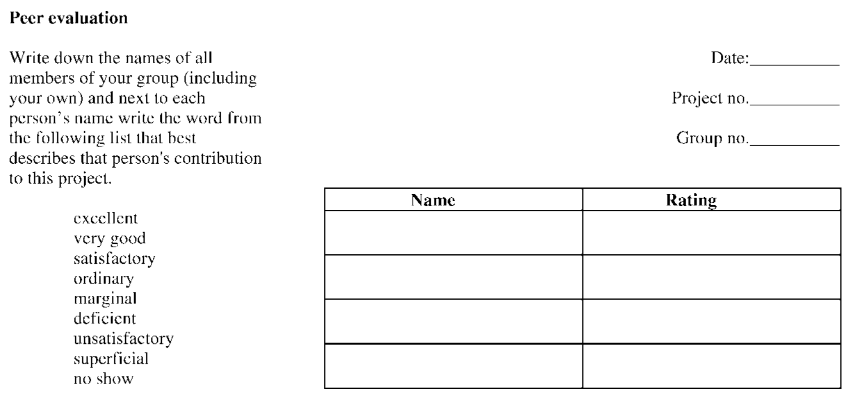
Most peer evaluations systems in both industry and academic settings keep the source of peer ratings confidential and provide rated individuals with only summary feedback, so that peers are more likely to provide critical feedback [13]. Attitudes toward peer evaluations are mixed, with many students resenting the systems, but others welcoming the opportunity to punish lazy or low performing teammates [13]. Many students are concerned that peer ratings will be biased by friendships, popularity, jealousy, or revenge [13].

Although peer evaluations are widely used in educational and industry settings, there are no generally accepted peer evaluation instruments. Some that have been proposed are so lengthy that they may be impractical for summative assessment in many classroom settings. For example, the Team Developer has 50 items [4]. Van Duzer and McMartin's instrument [14], which was developed for peer evaluation with group projects in engineering education, has 11 items plus a space to nominate the member who provided the most leadership, and asks the rater to distribute points among all team members in accordance with their performance. Although these long instruments may be appropriate for formative assessment, often instructors will want shorter instruments for summative assessment, particularly when the teamwork being evaluated is only a small part of the course. The study that follows compares the inter-rater reliability of three short peer evaluation instruments.

III. Study Design

1) Instruments compared: Form A (Figure 1) is based on the "autorating" instrument developed at the Royal Melbourne Institute of Technology (RMIT) by Robert Brown [12], and was adapted by Richard Layton for use at North Carolina A&T State University. Form A asks raters to write each team member's name and choose the one word "that best describes that person's contribution to this project" using nine words ranging from "no show" to "excellent". The instructor assigns numerical values to each rating ("Excellent" = 100, "Very Good" = 87.5, "Satisfactory" = 75,…,"No show" = 0). An average rating is computed for each student and for the team as a whole. Each student's project grade is then adjusted by a grading factor based on how that student's average rating compares to the team's average rating. More detail on strategies for adjusting student grades based on peer evaluations may be found in Kaufman, et al. [5].

Form B (Figure 2) was developed by Sam Ofori, Devdas Pai, and Richard Layton for use in a three-course sequence of design courses in mechanical engineering at North Carolina A&T State University. Form B provides more information than Form A about how the ratings will be used. It asks students to write the names of all team members and rate them on each of ten characteristics of good teamwork, using a five-point Likert scale ranging from "unsatisfactory" to "excellent".



Peer evaluation

Write down the names of all members of your group (including your own) and next to each person's name write the word from the following list that best describes that person's contribution to this project.

Date:__________

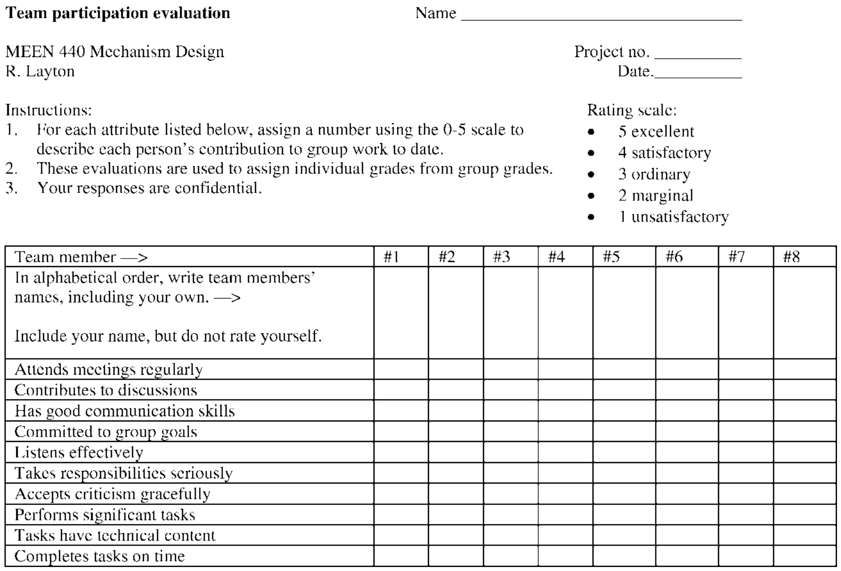
Project no.__________

Group no.__________

excellent
very good
satisfactory
ordinary
marginal
deficient
unsatisfactory
superficial
no show

Name	Rating

Figure 1. Peer evaluation instrument, Form A.



Team participation evaluation Name ______________________________

MEEN 440 Mechanism Design Project no. __________
R. Layton Date.__________

Instructions:

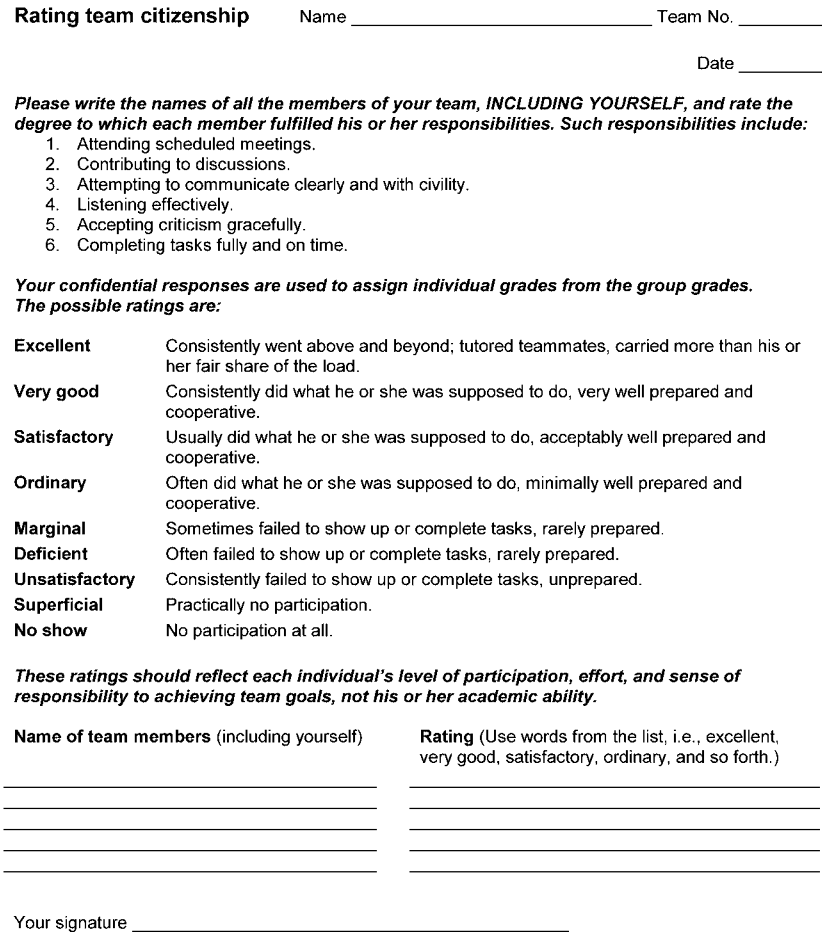
1. For each attribute listed below, assign a number using the 0-5 scale to describe each person's contribution to group work to date.
2. These evaluations are used to assign individual grades from group grades.
3. Your responses are confidential.

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								

Figure 2. Peer evaluation instrument, Form B.



Rating team citizenship Name ______________________________ Team No. ________

Date ________

Please write the names of all the members of your team, INCLUDING YOURSELF, and rate the degree to which each member fulfilled his or her responsibilities. Such responsibilities include:

1. Attending scheduled meetings.
2. Contributing to discussions.
3. Attempting to communicate clearly and with civility.
4. Listening effectively.
5. Accepting criticism gracefully.
6. Completing tasks fully and on time.

Your confidential responses are used to assign individual grades from the group grades. The possible ratings are:

Excellent	Consistently went above and beyond; tutored teammates, carried more than his or her fair share of the load.
Very good	Consistently did what he or she was supposed to do, very well prepared and cooperative.
Satisfactory	Usually did what he or she was supposed to do, acceptably well prepared and cooperative.
Ordinary	Often did what he or she was supposed to do, minimally well prepared and cooperative.
Marginal	Sometimes failed to show up or complete tasks, rarely prepared.
Deficient	Often failed to show up or complete tasks, rarely prepared.
Unsatisfactory	Consistently failed to show up or complete tasks, unprepared.
Superficial	Practically no participation.
No show	No participation at all.

These ratings should reflect each individual's level of participation, effort, and sense of responsibility to achieving team goals, not his or her academic ability.

Name of team members (including yourself) **Rating** (Use words from the list, i.e., excellent, very good, satisfactory, ordinary, and so forth.)

______________________________ ______________________________
______________________________ ______________________________
______________________________ ______________________________
______________________________ ______________________________
______________________________ ______________________________

Your signature ____

Figure 3. Peer evaluation instrument, Form C.

Form C (Figure 3), a one-item behaviorally anchored instrument, is similar to the adaptation of Brown's instrument [12] reported in Kaufman, et al. [5]. Form C provides more detailed instructions and information about how the ratings will be used. Forms similar to Form C have been used by a cadre of researchers in engineering education [15–20]. The instrument asks students to rate "team citizenship" and provides a list of different characteristics of good team citizenship, such as "attending scheduled meetings" and "attempting to communicate clearly and with civility." In addition, it gives behavioral anchor terms for each of the nine possible ratings using the same anchors as Form A.

2) Team assignment: On the first day of class in all semesters under study, students completed a questionnaire indicating their GPA, gender, course grade in a prerequisite course, and whether they were repeating the course. A seven-day scheduling table was included on which students indicate times that they cannot meet for group work. All information is voluntary except a signature verifying that prerequisites have been satisfied. The instructor used this information to form teams according to the following guidelines, based on Felder et al. [4]: (1) groups of three or four, selected by the instructor; (2) women and minorities are not outnumbered in a group; (3) heterogeneous ability level using GPA and grade in prerequisite course; (4) heterogeneity of major discipline, i.e., mechanical, electrical, civil, and so forth; and (5) times available to meet for group work.

3) Courses in which peer evaluation instruments were tested: Either Form A or B was administered to each section of Layton's required junior-level mechanical engineering course at North Carolina A&T State University during five semesters over three years. Form A was used anonymously by classes in the fall of 1996 and the spring of 1997 and confidentially (but not anonymously) in the fall of 1997. Form B was confidentially (but not anonymously) used in the fall of 1998 and the spring of 1999.

Seventy students participated (and received final course grades). Seventy-three percent were male, and 87 percent were minorities, and nearly all were African American (which is typical for the North Carolina A&T population). There were 21 teams. The team gender composition was as follows: one all female team, 11 all male teams, and 9 mixed gender teams. The team ethnicity composition was 12 all minority teams and 9 mixed ethnicity teams.

Form C was administered in two sections of Layton's junior-level Dynamics class in the winter of 2000. This is a core course for most engineering majors at North Carolina A&T State University. Seventy students participated. Eight-five percent were male. Ninety percent were minorities. There were 17 teams. The gender composition was 11 all-male teams and 6 mixed-gender teams. The team ethnicity was 11 all minority teams and 6 mixed-ethnicity teams.

4) Team projects and use of the peer ratings: Students completed peer evaluations in conjunction with group-based term projects for the course. Students were assigned to teams that worked all semester on two technical design problems that were collectively worth 20 percent of the students' final course grades (the remainder of the coursework did not involve formal cooperative learning). The teams completed oral presentations and written reports of their solution for each of the two projects and received team grades. Students were required to complete peer evaluations at the conclusion of each project, which the professor used, in conjunction with his own judgment, to adjust team members' individual scores so that students who contributed more (or less) than their teammates could receive higher (or lower) scores than the team grade. It is important to note that the professor did need to use discretion in adjusting grades based on peer evaluation scores because some students attempted to skew the ratings in their favor by rating themselves high and their teammates low. For example, one student rated himself "excellent" and everyone else on his team a "no show," whereas the rest of the team members assigned him ratings as low as "ordinary". These ratings were not deleted from the study, because the incidence of this was not widespread, and such incidences do ultimately affect the reliability of the instrument. The same approach was used to generate grade adjustment factors regardless of which form was used, so this is not a factor in the study. Self ratings were included in the study, as discussed earlier, based on the rationale of Kaufman, Felder, and Fuller [5] and Brown [12], with the exception of data from the spring of 1999, when Layton did not collect self ratings.

The peer evaluations were administered at the end of the first project, usually due by the fifth week of the semester. Students were encouraged to view the first evaluation as a chance to identify areas of improvement. The first administration of the peer evaluation instrument accomplished several goals: (1) teaching students about the peer evaluation procedure; (2) calibrating students' perceptions of the evaluation criteria; (3) giving students feedback on how their teammates perceive their work; and (4) alerting the instructor to groups or individuals needing attention. Specifically, the first peer evaluation allowed teams to identify "hitchhikers" and "overachievers," that is, group members that were either contributing too little or dominating the team effort. The instructor met outside of class with teams whose peer evaluations showed evidence of either of these problems. The purpose of these meetings was to help the teams to find ways to more evenly distribute the workload and to help resolve interpersonal difficulties and time conflicts. Therefore, the first evaluation was formative and is not included in the study. The second evaluation, which was administered at the end of the project, is summative.

5) Analytical procedure: Peer evaluation within a team yields multiple measures (one measure from each teammate) of the same traits (aspects of team contribution). The consistency of ratings between raters, i.e., how well students agree in their ratings of a particular teammate, is estimated by inter-rater reliability statistics. The technique used to measure inter-rater reliability in this study is a special form of analysis of variance described by Crocker and Algina [21].

Using Crocker and Algina's terminology, this study involves a nested, single-facet G-study design, which investigates how well the sample of measurements can be generalized to all possible measurements. A design is considered "nested" if the subjects are evaluated under different conditions (in this case by different raters). Because one of the subjects is self rating, but not always the same subject, the design is nested. The term "single-facet" indicates that only one factor, or facet, is changing between multiple measurements. Because all raters are rating the same person for the same period of behavior, it is only the rater that is changing between measurements, making this a single-factor design. A "G-study" refers to a study designed to determine the potential of an instrument to be generalized. This potential is quantified by a "generalizability coefficient," ρi, where $0 \leqslant \rho i \leqslant 1$ is an estimate of how well a single rater's score approximates the true score that would be obtained if enough raters evaluated each student. The special conditions of this type of design require a change in the standard formulas used to compute the terms of the ANOVA table, so the computational formulas used for a nested single-facet G-study design are shown in Tables 1 (a) and (b).

Since the design is nested within teams, these calculations are performed on a team-by-team basis and then accumulated to form the sums shown in Table 1 (a). Those accustomed to ANOVA notation will notice that MS_p, the mean square for examinees, does not estimate σ_p^2, as is usually the case—the residual term is part of the estimate because of the nesting. The residual term must be estimated and subtracted in order to estimate σ_p^2, which must be isolated in order to compute the generalizability coefficient. Further details are found in Crocker and Algina [21].

As described earlier, the results from Forms A and C are computed as a percentage, and are thus used directly. Form B is scored from 0–50, so Form B scores were scaled by a factor of two to put them on the same scale as the other forms.

SV	SS	df	MS	EMS
Examinees	$n_i \sum_p (X_{pI} - X_{PI})^2$	n_p–1	$SS_p/(n_p\text{-}1)$	$\sigma_e^2 + \sigma_i^2 + n_i\sigma_p^2$
Residual	$\sum_p \sum_i (X_{pi} - X_{pI})^2$	$n_p(n_i–1)$	$SS_r/[n_p(n_i\text{-}1)]$	$\sigma_e^2 + \sigma_i^2$
	$X_{pI} = \sum_i X_{pI} / n_i$		$X_{PI} = \sum_p \sum_i X_{pi} / n_p n_i$	

Table 1 (a). Computational formulas and expected mean squares for a one-way ANOVA in a nested single-facet G-study design.

SV	source of variance
SS	sum of squares
Df	degrees of freedom
MS	mean square (SS/df)
EMS	expected mean square in terms of population parameters
n_p	total number of examinees
n_i	the number of raters in a G-study
X_{pi}	the rating of examinee *p* by rater *i*
X_{pI}	the average rating of examinee *p* by all raters
X_{PI}	the average rating of all examinees by all raters
σ_p^2	the variance of the examinees' universe scores
σ_i^2	the variance of the rater means
σ_e^2	the average of the residual variance over all raters

Table 1 (b). Definition of terminology used in the ANOVA formulas.

IV. Results

Using the approach of Crocker and Algina's, we calculated the generalizability coefficient, which estimates how well the average of four raters' scores approximates the true score that would be obtained if a large number of raters evaluated the same student. Using four raters for each administration, the generalizability coefficient was 0.67 for Form A, which used one-word descriptors and a single item, and 0.74 for Form B, a ten-item instrument measuring different aspects of team contribution. This improvement in reliability is not statistically significant. Form B, with multiple items, would be expected to have a higher reliability than Form A, because reliability improves with increasing numbers of items [22]. The generalizability coefficient for Form C, the one-item instrument with behavioral anchors, was 0.78, which was a statistically significant improvement over Form A ($p < 0.10$), but was not significantly different from Form B. However, because Form C is one-item instrument and Form B has 10 items, Form C has the advantage of being more parsimonious with a comparable or slightly better level of inter-rater reliability.

Just as additional items improve the internal consistency reliability of instruments, the inter-rater reliability increases as the number of raters increases, according to the formula ,

$$\hat{p}^2_\iota = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + (\hat{\sigma}^2_i + \hat{\sigma}^2_e)/n'}$$, described in Crocker and Algina,

where is equal to ($MSp- MS_r)/n_i$. This is relevant because certain engineering projects feature larger teams, improving the estimated reliability of each of the instruments (for example, for a five-person team, estimated inter-rater reliabilities increase to 0.72, 0.78, and 0.82 for Forms A, B, and C, respectively). Professors considering using one of the peer evaluation instruments described in this study could use this equation to estimate how many raters they would need to have to achieve what they considered to be an acceptable level of inter-rater reliability for any form (A, B, or C) they adopt. The results of the formula are shown graphically in Figure 4. The plot is an extrapolation, so the data are computed rather than measured, and integer values of raters are indicated by data markers. These curves are most valid for a four-person team, since that was the most common grouping in the data studied. Note that the improvement in reliability gained by adding each additional rater (team member) diminishes as raters are added.

V. Limitations and Future Research

Form C achieved an inter-rater reliability of 0.78 based on four raters, which, although promising, leaves room for improvement. More importantly, a one-item survey, even though it shows better results psychometrically in this case, is risky because there is no measure of internal consistency. A higher level of reliability will be especially important for classes that use small teams where four raters are not available, and therefore the estimated reliability is

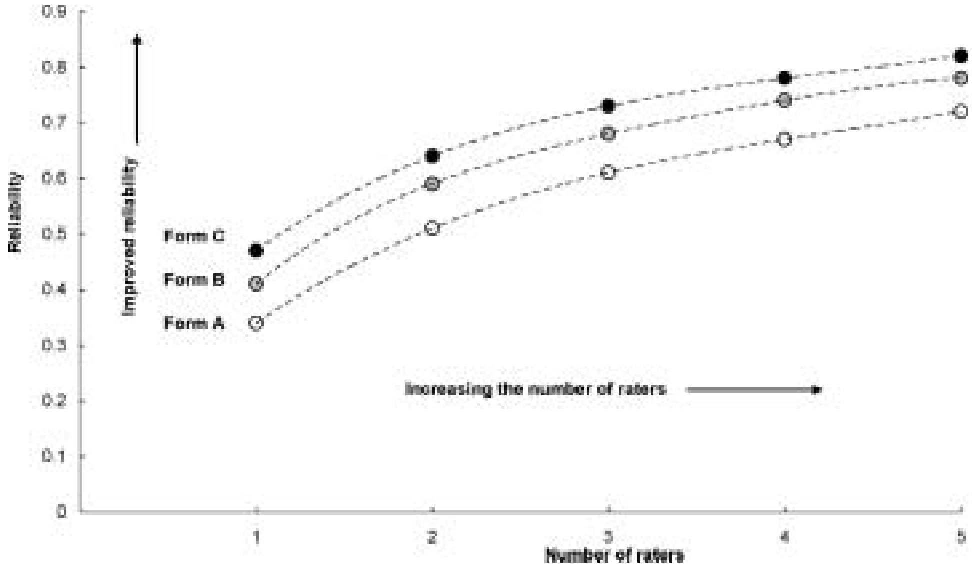


Figure 4. Computed inter-rater reliability of Forms A, B, and C for different numbers of raters.

lower than the results reported here. For these reasons, further research is being conducted on peer evaluation design and administration in hopes of developing more useful peer evaluation instruments with even higher levels of reliability. The authors and others are presently validating a five-item behaviorally anchored rating scale that extends the work described in this paper [23]. The new instrument was developed based on an extensive examination of the literature on teamwork. The five areas of team-member contribution that are assessed in the new instrument were determined based on an empirical analysis of data from two large surveys of students. Data from validity studies at five universities are being collected and analyzed, and preliminary results should be available by the summer of 2005.

Future studies could examine how including or excluding self ratings would affect instrument reliabilities. Although we explained our rationale for including self ratings earlier, some studies suggest they are problematic. Lejk and Wyvill found that strong students tend to underrate themselves and weaker students tend to overrate themselves [24]. Thompson found similar results for formative peer evaluations and showed that the validity of peer assessments was high relative to self assessments [25]. If self assessments are less reliable, our reliability has been lowered by including them. The estimates reported here, therefore, are conservative.

The present study involved a population of students that had more minorities than is typical for engineering education as a whole. This is good because it is important to test the robustness of this class of instruments in different populations. However, additional studies using behaviorally anchored rating scales such as Form C in other settings are recommended to confirm the generalizability of the inter-rater reliability estimates computed here.

It is important to be aware that peer evaluations can, in certain cases, be self-serving and dishonest. When professors are familiar with teams and their members, as Layton was in this study, they can use their professional judgment about how best to deal with individual circumstances. For cases in which the instructor does not know the students well, it may be helpful to use formulas to identify situations that raise concerns, such as when there is low agreement among peer raters.

VI. Conclusions

Form C is a simple peer evaluation instrument with the best reliability of the three forms examined. The inter-rater reliability of Form C, a single-item behaviorally anchored peer evaluation instrument, exceeded that of an instrument with 10 times as many items due to the addition of behavioral anchors. It is known that behaviorally anchored rating scales improve instrument reliability (and therefore validity) [26], but it is notable that this improvement was sufficient to offset the negative effect on reliability of having only a single item. Because students are likely to complete shorter instruments more conscientiously than longer ones, parsimony is a desirable characteristic of a peer ratings instrument, along with high inter-rater reliability, and, for multi-item instruments, adequate internal consistency.

It is also noteworthy that Form C included more statements and instructions on the evaluation instrument, such as listing teamwork behaviors that reminded students of various important aspects of teamwork. Perhaps these additional instructions encouraged students to think about various ways in which team members contributed to the team effort before they made their ratings of each team member. Although parsimony is important, it is also important that peer evaluations give credit to peers for various ways in which they contribute to their teams.

Acknowledgments

Our thanks to Rich Felder and Deborah Follman (formerly Kaufman) for their suggestions that led to the improvements in Form C. We also acknowledge our many colleagues who have joined us in conversation on this subject at workshops and conferences. This material is precedent to work supported by NSF DUE-ASA, Award Number 0243254, "Designing a Peer Evaluation Instrument that is Simple, Reliable, and Valid."

REFERENCES

[1] *Criteria for Accrediting Engineering Programs*, published by ABET, Inc., Baltimore, Maryland, www.abet.org/images/Criteria/E001 04-05 EAC Criteria 11-20-03.pdf, 2004.

[2] Johnson, D.W., Johnson, R.T., and Smith, K.A., *Active Learning: Cooperation in the College Classroom*, Edina, Minn.: Interaction Book Co., 1998.

[3] Millis, B.J., and Cottell, Jr., P.G., *Cooperative Learning for Higher Education Faculty*, Phoenix, Ariz.: American Council on Education/Oryx Press, 1998, p. 194.

[4] McGourty, J., and De Meuse, K., *The Team Developer: An Assessment and Skill Building Program*, New York, New York: J. Wiley and Company, 2000.

[5] Kaufman, D.B., Felder, R.M., and Fuller, H., "Accounting for Individual Effort in Cooperative Learning Teams," *Journal of Engineering Education* Vol. 89, No. 2, 2000, pp. 133–140.

[6] Moran, L., Musselwhite, E., Zenger, J.H., *Keeping Teams on Track: What to Do When the Going Gets Rough*, Chicago, Ill.: Irwin Professional Publishing, 1996, pp. 271–274 and 280–283.

[7] Williams, K., Harkins, S., and Latané, B., "Identifiability as a Deterrent to Social Loafing: Two Cheering Experiments," *Journal of Personality and Social Psychology*, Vol. 40, 1981, pp. 303–311.

[8] Conway, J.M., and Huffcutt, A.I., "Psychometric Properties of Multisource Performance Ratings: A Meta-analysis of Subordinate, Supervisory, Peer, and Self-ratings," *Human Performance*, Vol. 10, 1997, pp. 331–360.

[9] Viswesvaran, C., Ones, D.S., and Schmidt, F.L., "Comparative Analysis of the Reliability of Job Performance Ratings," *Journal of Applied Psychology*, Vol. 81, 1996, pp. 557–574.

[10] Harris, M.M., and Schaubroeck, J., "A Meta-analysis of Self-supervisor, Self-peer, and Peer-supervisor Ratings," *Personnel Psychology*, Vol. 41, 1988, pp. 43–62.

[11] Schmitt, N., Gooding, R.Z., Noe, R.A., and Kirsch, M., "Meta-analyses of Validity Studies Published between 1964 and 1982 and the Investigation of Study Characteristics," *Personnel Psychology*, Vol. 37, 1984, pp. 407–422.

[12] Brown, R.W., "Autorating: Getting Individual Marks from Team Marks and Enhancing Teamwork," *Proceedings, 1995 Frontiers in Education Conference*, IEEE/ASEE, Pittsburgh, November, 1995.

[13] Sheppard, S., Chen, H.L., Schaeffer, E., Steinbeck, R., Neumann, H., and Ko, P., *Peer Assessment of Student Collaborative Processes in Undergraduate Engineering Education*, Final Report to the National Science Foundation, Award Number 0206820, NSF Program 7431 CCLI-ASA, 2004.

[14] Van Duzer, E., and McMartin, F., "Methods to Improve the Validity and Sensitivity of a Self/Peer Assessment Instrument," *IEEE Transactions on Education*, Vol. 43, No. 2, May 2000, pp. 153–158.

[15] Finelli, C.J., "Assessing Improvement in Students' Team Skills and Using a Learning Style Inventory to Increase It," *Proceedings, 2001 Frontiers in Education Conference*, IEEE/ASEE, Reno, Nevada, October, 2001.

[16] Ohland, M.W., and Finelli, C.J., "Peer Evaluation in a Mandatory Cooperative Education Environment," *Proceedings, 2001 American Society of Engineering Education Conference and Exposition*, Albuquerque, New Mexico, June, 2001.

[17] Layton, R.A., and Ohland, M.W., "Peer Evaluations in Teams of Predominantly Minority Students," *Proceedings, 2000 American Society of Engineering Education Conference and Exposition*, Washington, D.C., 2000.

[18] Ohland, M.W., and Layton, R.A., "Comparing the Reliability of Two Peer Evaluation Instruments," *Proceedings, 2000 American Society of Engineering Education Conference and Exposition*, Washington, D.C., 2000.

[19] Layton, R.A., and Ohland, M.W., "Peer Evaluations Revisited: Focus on Teamwork, Not Ability," *Proceedings, 2001 American Society of Engineering Education Conference and Exposition*, Albuquerque, New Mexico, June, 2001.

[20] Ohland, M.W., Loughry, M.L., Carter, R.L., Bullard, L.F., Felder, R.M., Finelli, C.J., Layton, R.A., and Schmucker, D.G., "Developing a Peer Evaluation Instrument that is Simple, Reliable, and Valid," in press, *Proceedings, 2005 American Society of Engineering Education Conference and Exposition*, Portland, Oregon, June, 2005.

[21] Crocker, L., and Algina, J., *Introduction to Classical and Modern Test Theory*, Chicago, Ill.: Holt, Rinehart and Winston, Inc., 1986, p. 143, 157ff.

[22] Kubiszyn, T., and Borich, G., *Educational Testing and Measurement: Classroom Application and Practice*, 4th Ed., Harper Collins, New York, 1993, Chapter 15.

[23] Ohland, M.W., Loughry, M.L., Carter, R.L., and Yuhasz, A.G., "Designing a Peer Evaluation Instrument that is Simple, Reliable, and Valid" *Proceedings, 2004 American Society of Engineering Education Conference and Exposition*, Salt Lake City, Utah, June 2004.

[24] Lejk, M., and Wyvill, M., "The Effect of the Inclusion of Self-assessment with Peer Assessment of Contributions to a Group Project: A Quantitative Study of Secret and Agreed Assessments," *Assessment & Evaluation in Higher Education*, Vol. 26, 2001, pp. 551–561.

[25] Thompson, R.S., "Relative Validity of Peer and Self-evaluations in Self-directed Interdependent Work teams," *Proceedings, 2001 Frontiers in Education Conference*, IEEE/ASEE, pp. T4A/9-T4A/14.

[26] Smith, P.C., and Kendall, L.M., "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales," *Journal of Applied Psychology*, Vol. 47, No. 2, 1963, pp. 149–155.

AUTHORS' BIOGRAPHIES

Matthew W. Ohland is an assistant professor in Clemson University's General Engineering Program. He served as the assistant director of the NSF-sponsored SUCCEED engineering education coalition and as an NSF postdoctoral fellow. His studies are in engineering education and he has spoken and conducted workshops nationally and internationally. He is the 2002–2006 president of Tau Beta Pi, the national engineering honor society. Ohland received a B.S. in engineering and a B.A. in religion in 1989 from Swarthmore College. He earned M.S. degrees from Rensselaer Polytechnic Institute in mechanical engineering in 1991 and in materials engineering in 1992. He received his Ph.D. in civil engineering with a graduate minor in education from the University of Florida in 1996.

Address: 104 Holtzendorff Hall, Clemson, South Carolina 29634-0902; telephone: (864) 656-2542; fax: (864) 656-1327; e-mail: ohland@clemson.edu.

Richard A. Layton is an associate professor of Mechanical Engineering at Rose-Hulman Institute of Technology. Prior to his academic career, Dr. Layton worked for twelve years in consulting engineering, culminating as a group head and a project manager. His interest in the quality of the student teaming experience and the technical merit of student team deliverables is based on this background in project management. He is an active member of the Educational Research and Methods (ERM) Division of ASEE and has given workshops on building student teams for the ERM's Regional Effective Teaching Institute. Layton's professional interests include modeling and simulation of dynamic systems as well as curriculum and lab development in mechanical engineering. He earned a B.S. in engineering (1991) from California State University, Northridge. He received his M.S. (1993) and Ph.D. (1995), both in mechanical engineering, from the University of Washington.

Address: Rose-Hulman Institute of Technology, 5500 Wabash Avenue, CM191, Terre Haute, Indiana 47803; telephone: (812) 877-8905; fax: (812) 877-8025; e-mail: layton@rose-hulman.edu.

Misty L. Loughry is an assistant professor of Management at Clemson University. She holds a Ph.D. in management from the University of Florida, an M.B.A. from Loyola College in Maryland, and a B.A. from Towson State University. Dr. Loughry worked in banking for ten years prior to beginning her academic career. She is published in *Research in Personnel and Human Resource Management*, *Business Horizons*, and *Journal of Information Technology Management*. Her research interests center on control in organizations, particularly peer control, teamwork, and supervision.

Address: Department of Management, 123-D Sirrine Hall, Box 341305, Clemson, South Carolina 29634-1305; telephone: (864) 656-3763; fax: (864) 656-2015; e-mail: loughry@clemson.edu.

Amy G. Yuhasz contributed to this work as a Postdoctoral Fellow in Clemson University's General Engineering Program. Dr. Yuhasz received her B.A. in Mathematics from Huntingdon College (1997), M.S.I.E. (1999) from Clemson University and Ph.D. in Industrial Engineering from Clemson University (2003). Her postdoctoral work in General Engineering at Clemson University (2003) focused on statistical analysis of freshman retention programs. Prior to joining the General Engineering Program, she was a visiting professor of Industrial Engineering at The University of Alabama specializing in statistics and engineering economics (2001).

Address: 104 Holtzendorff Hall, Clemson, South Carolina 29634-0902; telephone: (864) 656-0329; fax: (864) 656-1327; e-mail: ayuhasz@clemson.edu.