Comparing the Reliability of Two Peer Evaluation Instruments

Matthew W. Ohland, Richard A. Layton
University of Florida / North Carolina A&T State University

Abstract

This paper presents an analysis of student peer evaluations in project teams to compare the reliability of two different evaluation procedures. The project teams consist of junior-level students in a mechanical engineering design course taught by Layton for five semesters in 1997, 1998, and 1999.

The peer-evaluation instruments were used by students to evaluate their teammates’ contributions to the team’s deliverables—oral and written presentations of their solution to a technical design problem. The first instrument is an adaptation of the one advocated by Brown, in which students use a prescribed list of terms such as “excellent,” “very good,” “satisfactory,” and so forth. The second form, by Layton, asked students to assign a numerical rating (from 0 to 5) to 10 different aspects of contribution to the team.

Analysis of variance was used to study the reliability of each of the instruments, using a special form by Crocker and Algina to study inter-rater reliability. The similarity of the reliability coefficients of the two instruments ($\rho_i = 0.34$ for Brown’s instrument and $\rho_i = 0.41$ for Layton’s instrument) strengthens the assumption made in the first study—that data from the two instruments are similar enough to be normalized for comparison. At the same time, the higher reliability of Layton’s instrument lends credence to Layton and Ohland’s conclusion that focusing on identified behavioral characteristics of good teamwork (as Layton’s instrument does) can improve peer evaluation. Layton’s instrument accomplishes this to an extent, yielding a modest improvement in reliability. More focused attempts to define teamwork success behaviorally, such as the modification of Brown’s instrument by Kaufman et al., may yield further improvements in reliability. The overall reliability of the two instruments validates such instruments as repeated measures of a consistent trait.

I. Introduction

In order to satisfy ABET EC 2000’s charge for outcomes assessment, evaluation techniques that are largely new in engineering academe are coming into use. One such technique is the peer evaluation instrument. Recent papers by Brown,1 Kaufman et al.,2 and Layton and Ohland3 have described peer evaluation instruments and their use in measuring students’ ability to function in teams. It is important in this context to assess the reliability of peer evaluation instruments.

In a mechanical engineering design course at North Carolina A&T State University, students were assigned to groups for the purpose of completing term projects in design. The project teams
consisted of junior-level students in a mechanical engineering design course taught by Layton for five semesters in 1997, 1998, and 1999. Peer-evaluation instruments were used by students to evaluate their teammates’ contributions to the team’s deliverables—oral and written presentations of their solution to a technical design problem. Each student is rated by his or her teammates, producing multiple ratings of a single measure—the student’s contribution to the team effort. Having multiple ratings of a single measure permits the reliability of the peer evaluation instrument itself to be evaluated. This process is similar to correlating different raters’ evaluations of each team member.

Over the course of the five terms, two different instruments were used. The first instrument is an adaptation of the one advocated by Brown, in which students use a prescribed list of terms such as “excellent,” “very good,” “satisfactory,” and so forth. The other form, by Layton, asked students to assign a numerical rating (from 0 to 5) to 10 characteristics of good teamwork. We compare the two instruments to determine which instrument is more reliable, and draw conclusions on how the instrument might be further improved.

II. Class and team demographics

Peer evaluation data from five offerings of a single course taught by Layton in 1997, 1998, and 1999 are included in this study. This study includes more semesters than the related study by Layton and Ohland, because two semesters accepted evaluations anonymously, precluding their use in studying gender and minority effects in ratings, whereas anonymity does not affect the use of evaluations for the present reliability study. The course is MEEN 440 Mechanism Design and Analysis, a required junior-level design course in Mechanical Engineering at North Carolina A&T State University. Student demographics are shown in Table 1.

Table 1: Demographic data.

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women</th>
<th>Non-minorities</th>
<th>Minorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>118</td>
<td>75%</td>
<td>25%</td>
<td>14%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>86%</td>
</tr>
</tbody>
</table>

Here, N is the number of students receiving final course grades. “Minorities” includes African-American, Hispanic, and Native American students (<1% Hispanic and Native American), and “non-minorities” includes students of all other ethnic backgrounds. More details of the class and group structure can be found in Layton and Ohland.

III. Peer rating procedures

Peer evaluations are performed at the conclusion of each of two design projects assigned during the semester. The evaluations at the end of the first project, usually due by the fifth week of the semester, accomplish several goals:

- Students learn about the peer evaluation procedure.
- Students reflect on the evaluation criteria.
- Students get feedback on how the group assesses each member’s work to date.
- The instructor is alerted if groups are not functioning.

This first evaluation allows the groups to identify both “hitchhikers” and “overachievers,” that is, group members that are contributing either too little or too much to the group effort. The instructor meets with such groups outside of class to help them find ways to more evenly
distribute the work load and to help resolve interpersonal difficulties and time conflicts. Also, because the first project is weighted less than the second project in computing the final course grade, a poor peer evaluation after the first project has only a modest impact on a students’ course grade. Students are encouraged to view this first evaluation as a chance to identify areas of improvement.

Two different peer evaluation instruments were used throughout the period of this study. The first peer evaluation instrument uses a system adapted from Brown, in which students use a prescribed list of terms such as “excellent,” “very good,” “satisfactory,” and so forth to evaluate one another’s contributions to the team’s deliverables. The verbal ratings are converted to a numerical equivalent and an individual’s weighting factor is the individual’s average rating divided by the group average. An individual student’s grade is the group grade multiplied by this weighting factor. A maximum factor of 1.07 was imposed to prevent students from receiving artificially high grades due to having a teammate with very low ratings.

The second instrument solicits ratings of unsatisfactory (1), marginal (2), ordinary (3), satisfactory (4), or excellent (5) on ten modes of team contribution: attends meetings regularly; contributes to decisions; has good communication skills; is committed to group goals; listens effectively; takes responsibilities seriously; accepts criticism gracefully; performs significant tasks; tasks have technical content; and completes tasks on time. The results from the different instruments are normalized to a common 0-100 scale for comparison in the study by Layton and Ohland examining gender and ethnicity effects. Sample peer evaluation instruments are given in the appendix.

IV. Data analysis

Instruments that collect multiple observations for each subject using the same measures interpreted by different individuals require a special form to estimate their reliability, measuring the consistency of the observations from one rater to another. The present peer evaluation instruments are of this type, and inter-rater reliability was studied using analysis of variance in a special form by Crocker and Algina.4

Using the terminology of Crocker and Algina, the investigation capable of estimating how well the sample of measurements can be generalized to all possible measurements is a nested single-facet G-study design. A design is considered “nested” if the subjects are evaluated under different conditions. The term “single-facet” indicates that only one factor, or facet, is changing between multiple measurements. A “G-study” refers to one designed to determine the potential of an instrument to be generalized. This potential is quantified by a “generalizability coefficient,” \( \rho_i \), where \( 0 \leq \rho_i \leq 1 \) 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.

We apply this design to the problem of quantifying the reliability of a peer evaluation instrument as follows. Different raters (one’s teammates) rate an individual on his or her contribution to the team work (nested). The only changing factor among these ratings is who does the rating (single-facet). A generalizability coefficient is computed for each instrument using all ratings made using that instrument (G-study). Layton’s adaptation of Brown’s instrument yielded a generalizability coefficient \( \rho_i = 0.34 \), while Layton’s own instrument yielded a coefficient of \( \rho_i = 0.41 \).
V. Conclusions and recommendations

The similarity of the reliability coefficients of the two instruments (\( \rho = 0.34 \) for Brown’s instrument and \( \rho = 0.41 \) for Layton’s instrument) strengthens the assumption made in the first study—that data from the two instruments are similar enough to be normalized for comparison. The similarity indicates that each instrument is approximately as good as the other at obtaining a consistent measurement of the trait being measured.

At the same time, the higher reliability of Layton’s instrument lends credence to Layton and Ohland’s conclusion that focusing on identified behavioral characteristics of good teamwork (as Layton’s instrument does) can improve peer evaluation. Layton’s instrument accomplishes this to an extent, yielding a modest improvement in generalizability. More focused attempts to define teamwork success behaviorally, such as the modification of Brown’s instrument by Kaufman et al., may yield further improvements in generalizability.

Another issue that deserves further consideration is that of validity. Reliability (or, in this case, generalizability) is a measure of how well a set of measurements can be generalized to the entire population of possible measurements of a trait. An instrument with high reliability is measuring some characteristic in a consistent and repeatable way, but reliability estimates cannot tell us if we are measuring what we intend to measure. In two independent studies, Kaufman et al. and Layton and Ohland conclude that the correlation of course grades with teamwork skill ratings indicates that the student raters may be evaluating perceived academic ability rather than teamwork skills. Validity investigates that question—whether we are measuring that which we seek to measure, teamwork skills in this case. One approach to establishing validity, called concurrent validity, is to compare the results of the instrument in question to a known good measure. Barring the availability of such a measure, the best approach is to seek alternative measures of the same trait that are not likely to be biased in the same fashion. In order to assure validity, we recommend that the measurement of teamwork skills through peer evaluation be supplemented by other measurements such as instructor observation.

VI. Acknowledgements

The second peer evaluation instrument used in this study was developed jointly by Sam Ofori, Devdas Pai, and Richard Layton for use in a three-course sequence of design courses in mechanical engineering at NC A&T. The authors gratefully acknowledge our colleagues’ contribution.

Bibliography

MATTHEW W. OHLAND
Matt received his Ph.D. in Civil Engineering with a minor in Education from the University of Florida in 1996, and since then has been the Assistant Director of the NSF-sponsored SUCCEED Engineering Education Coalition. In 1998, he was awarded an NSF Postdoctoral Fellowship for Science, Mathematics, Engineering, and Technology Education, under which he is studying the impact of a wide variety of student programs on student success.

RICHARD A. LAYTON
Richard A. Layton received his Ph.D. from the University of Washington in 1995 and he is currently an Assistant Professor of Mechanical Engineering at North Carolina A&T State University. Prior to his academic career, Dr. Layton worked 12 years in consulting engineering, with the final five years as a group head and project manager. His technical research is in the area of dynamic systems and control. He is a registered Professional Engineer in NC.

Appendix

Samples of the two peer evaluation instruments used in this study are given below. The first is Layton’s adaptation of Brown’s instrument. In sample shown, the instrument is anonymous. The second is Layton’s instrument assessing the ten modes of team contribution. The sample shown asked students not to rate themselves. Anonymity and self-ratings are two features that sometimes vary from semester to semester. Although space for 8 team members are shown on the second form, no team has been larger than five members.

---

**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.

- excellent
- very good
- satisfactory
- ordinary
- marginal
- deficient
- unsatisfactory
- superficial
- no show

<table>
<thead>
<tr>
<th>Name</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Date:__________

Project no.__________

Group no.__________
# 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.

<table>
<thead>
<tr>
<th>Team member</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
</tr>
</thead>
</table>
| 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 |

Rating scale:
- 5 excellent
- 4 satisfactory
- 3 ordinary
- 2 marginal
- 1 unsatisfactory