

**article:1256****Effects of Behavioral Anchors on Peer Evaluation Reliability**

To ensure that students graduate with an ability to function in multidisciplinary teams, faculty must be able to assess that ability in each student. Recognizing that a significant amount of team interaction occurs beyond the faculty's sight, a student's teammates are a valuable source of information on that student's effectiveness in team situations. Although various methods of peer evaluation in engineering education have been studied in recent years, the instrument on which this study focuses follows from an instrument originally developed by Robert Brown at the Royal Melbourne Institute of Technology, modified for classroom use (rather than research purposes) by Richard Felder, and later studied by Kaufman, Felder, and Fuller. The simple format of the instrument and Felder's reputation in the engineering education community have helped Felder's modified instrument become popular, particularly among members of the Educational Research and Methods Division of the American Society of Engineering Education.

By the time he became acquainted with the work of Brown and Felder, Richard Layton had collected peer evaluations for a few semesters using an instrument that he and colleagues developed at North Carolina A&T. He subsequently collected peer evaluation data using Brown's instrument and Felder's instrument to identify a better approach for classroom use. Even though Layton had not planned a research study, data collected using three different instruments allowed us to answer the research question, "what is the inter-rater reliability of these three instruments and how do they compare?"

In a typical Likert-scale survey, students indicate their level of agreement (i.e., "Strongly agree") with a statement (i.e., "This teammate hasn't shown up for any meetings.") or select from various other short descriptors, such as selecting the frequency (i.e., "Often") with which certain behaviors (i.e., lateness) occur. An alternative is to use behavioral anchors, in which response levels have explicit behavioral descriptions. The first level might be "I have never met this teammate." A higher level might be "this teammate shows up for team meetings reliably." The use of behavioral anchors has the potential to increase inter-rater reliability (the degree to which multiple raters agree in their assessment of a team member's contribution) because it helps students to interpret the scale in a more similar way. Felder's instrument has behavioral anchors for just one question (evaluating team citizenship), and instruments tend to be more reliable as more items are added. Thus the larger number of items would suggest that scores on Layton's 10-item survey would be more reliable, whereas the use of behavioral anchors might lead us to believe that the single-item behaviorally anchored scale would be more reliable. Our research question was to characterize this difference. This is an important issue because in order for peer evaluation in engineering education to be widely used and taken seriously by faculty and students, it is important to be able to measure team effectiveness as accurately as

possible with as short an instrument as is feasible. The brevity of the instrument matters because longer instruments take more time to complete and this reduces the chances that faculty will administer them frequently and also increases the chances that students will take mental shortcuts rather than carefully reading each item before evaluating team members.

Instrument A was developed by Brown at RMIT and used by Felder, and simply uses nine short verbal descriptors such as “Excellent” or “Satisfactory”. Instrument B was developed by Layton and colleagues at NC A&T—students assign a numerical rating from 0-5 to students on categories such as “Listens effectively” and “Completes tasks on time”. Instrument C is an expanded version of Instrument A in which verbal descriptions of “Excellent”, “Satisfactory”, and other levels are provided (e.g., “Satisfactory” is described as “Usually did what he or she was supposed to do, acceptably well prepared and cooperative.” These verbal descriptions make Instrument C a behaviorally anchored rating scale.

The inter-rater reliability of the instrument that collects a numerical rating on 10 items is simple to calculate. The behavioral anchors, however, cannot be considered continuous, so analyzing the data from those instruments requires the use of statistics for categorical variables—in this case, a variable with rank-ordered responses. Furthermore, the statistical approach we selected uses a constant team size, so the various calculations required for the ANOVA (e.g., Mean Square) were computed for each team and then aggregated.

We discovered that the introduction of behavioral anchors, even using a single-item instrument, was enough to improve the inter-rater reliability of the peer evaluation scores. As we expected, the behaviorally anchored scale was a significant improvement over the single-item categorical evaluation that lacked behavioral anchors. Although the inter-rater reliability of the single-item behaviorally anchored scale was not significantly higher than the ten-item unanchored scale, it was no less reliable. This prompted us to focus our efforts on developing a multi-item behaviorally anchored instrument. We expect that a behaviorally anchored instrument with just a few items will prove to be far more reliable and valid while remaining simple for students to complete (especially once they are already familiar with the behavioral anchors).

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.”

Author 1: Matthew Ohland email: [ohland@clmson.edu](mailto:ohland@clmson.edu)

Author 2: Richard Layton email: [layton@rose-hulman.edu](mailto:layton@rose-hulman.edu)

Author 3: Misty Loughry email: [loughry@clmson.edu](mailto:loughry@clmson.edu)

Author 4: Amy Yuhasz email: [ayuhasz@clmson.edu](mailto:ayuhasz@clmson.edu)

[: Back to Summer 2005 Issue Vol. 1, No. 2](#)

[: Back to List of Issues](#)

[: Back to Table of Contents](#)