

Developing a Peer Evaluation Instrument that is Simple, Reliable, and Valid

Matthew W. Ohland, Misty L. Loughry, Rufus L. Carter, Lisa G. Bullard, Richard M. Felder, Cynthia J. Finelli, Richard A. Layton, and Douglas G. Schmucker

General Engineering, Management, Clemson University / Institutional Research and Assessment, Marymount University / Chemical and Biomolecular Engineering, North Carolina State University / Center for Research on Learning and Teaching-North, University of Michigan / Mechanical Engineering, Rose-Hulman Institute of Technology / Civil Engineering, Western Kentucky University

Abstract

A multi-university research team is working to design a peer evaluation instrument for cooperative learning teams that is simple, reliable, and valid. In this work, an overview of the process of developing behaviorally anchored rating scales (BARS) will be presented, including the establishment of a theoretical basis for the instrument and a description of the extensive classroom testing of the draft instrument conducted during fall 2004.

Introducing the draft instrument to the engineering education community through exposure in the NSF grantees' poster session is expected both to improve the validity of the scale itself through the feedback we receive and to accelerate the dissemination of the instrument.

Introduction

This project and its goals were introduced in earlier work.¹ ABET's requirement that engineering graduates have an ability to function on multi-disciplinary teams² has driven an expanded use of cooperative learning in engineering curricula.³ A fundamental tenet of cooperative learning is holding individualThis will be achieved by improving individual accountability by adjusting team members accountable for fulfilling their responsibilities to the team. An effective and increasingly common way of addressing this tenet is to have team members rate one another's performance and to use the ratings to adjust the team assignment grades for individual performance. The challenge is to devise a rating system that is fair, simple to administer, reliable, and valid.

Our prior experience was based on a peer rating system developed by Robert Brown of the Royal Melbourne Institute of Technology.^{4,5} Brown's system is a single-item form of behaviorally anchored rating scale (BARS), an instrument that aims to improve validity by reducing the subjectivity of ratings by providing verbal descriptions to anchor the points of the scale.⁶ The BARS was one of a variety of rating scales studied between 1960 and 1980. As rating-scale researchers became convinced that performance ratings were robust to changes in rating scale format,⁷ research into new rating formats waned, and has remained slow in the past 15 years.⁸

Our earlier work also described the theoretical underpinnings of the development of a new rating scale drawing upon the expertise of the broader community of researchers studying teamwork. Development of a new BARS instrument began by studying items using a Likert scale to consolidate an exhaustive list of team measures into a smaller number of factors that have theoretical underpinnings.¹

The status of the development of a new BARS peer evaluation instrument

Our work continued with extensive data collection to assure that the instrument would be both theoretically sound and backed by solid empirical research. This process will be described in detail in a future paper. After selecting the factors to be included in the development of a BARS scale, we adapted the “critical incident methodology” to identify behaviors to anchor the scale, attempting to ensure that a representative distribution of performance behavior was discussed.⁹ To calibrate our own thinking, we tried to categorize behaviors at three levels—“does not meet expectations,” “meets expectations,” and “exceeds expectations.” The result was a new peer evaluation instrument that collected student ratings on five factors using a BARS scale for each:

1. Contributing to the Team’s Work
2. Interacting with Teammates
3. Keeping the Team on Track
4. Expecting Quality
5. Having Task-Related Knowledge, Skills, and Abilities

Some on the research team felt that the ability-related fifth factor would inappropriately focus the students’ attention on the ability of their teammates rather than on their citizenship. Others on the team thought it was inappropriate to ignore the special contributions that a team member might make because of his or her unique knowledge, skills, or abilities. As a result, two instruments are being tested—one with the fifth factor, and one without.

Fall 2004 data collection

Establishing concurrent validity with the Team Developer: The *Team Developer*TM is a computer-based survey that provides students with multi-source assessment and feedback.¹⁰⁻¹² A web-based implementation of this instrument was used to establish concurrent validity. Students in Clemson University’s first engineering class, *Introduction to Engineering Disciplines and Skills*, alternated between using the Team Developer and either the 4-factor or 5-factor version of the BARS instrument. An experimental design was devised to evaluate concurrent validity without confounding the outcomes by the order in which the instruments were administered. Student teams were assigned randomly into one of four protocols, with the instrument specified below administered at the end of each of four phases of a semester project. All members of a team used the same instrument. Each protocol was executed by approximately 50 four-person teams, comprising the 787 students enrolled after the drop-add period. As part of studying the instrument’s validity, the results will be used to investigate how the presence of the fifth factor affected student rating practices.

| Protocol used | Phase 1 instrument | Phase 2 instrument | Phase 3 instrument | Phase 4 instrument |
|----------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Protocol A | Team Developer | 4-factor BARS | Team Developer | 4-factor BARS |
| Protocol B | 4-factor BARS | Team Developer | 4-factor BARS | Team Developer |
| Protocol C | Team Developer | 5-factor BARS | Team Developer | 5-factor BARS |
| Protocol D | 5-factor BARS | Team Developer | 5-factor BARS | Team Developer |

Establishing concurrent validity with the Likert-scale version of the instrument: We compared the BARS instrument with the Likert-scale instrument from which it was derived, since the latter instrument had a strong theoretical base. The outcomes would provide a sense of how well the theoretical constructs established in developing the Likert instrument held up through the application of critical incident methodology to generate behavioral descriptions. This study was conducted at North Carolina State University using a reduced version of the Clemson protocol. Student teams in a chemical engineering course were assigned randomly into one of two protocols, with all members of a team using the same instrument for any given administration.

| Protocol used | Checkpoint 1 instrument | Checkpoint 2 instrument |
|----------------------|--------------------------------|--------------------------------|
| Protocol A | Likert instrument | 5-factor BARS |
| Protocol B | 5-factor BARS | Likert instrument |

Establishing validity using verbal protocol analysis and expert observation: At Western Kentucky University, graduate students (from Psychology or a similar field) will review videotaped team meetings from several Civil Engineering courses and rate team members independently of a peer evaluation. This protocol will test the newly developed peer evaluation instrument's concurrent validity with behavioral observation of the psychology students. All team members will review their ratings with a graduate assistant and explain their choices. Content validity will then be established using a verbal protocol analysis. Students will have the option of reviewing the tape and reconsidering their ratings.

Establishing the influence of a training protocol using a time series: Concerned about the unwillingness of students to evaluate their teammates candidly and the robustness of performance evaluation to different rating methods, the research team has anticipated that the only way to reduce the subjectivity of a peer evaluation instrument will be to train students how to use the instrument, thereby calibrating their responses in much the same way as is used in Calibrated Peer Review.¹³ At Rose-Hulman, Mechanical Engineering students used the 5-factor BARS instrument multiple times in a quarter. Between two of the administrations, students received training in how to use the instrument. Preliminary results indicate that the training is very valuable in helping students understand the instrument dimensions, increasing both the variability and reliability of student ratings.

Establishing concurrent validity with the Van Duzer-McMartin instrument: Concurrent validity with an instrument by Van Duzer and McMartin will be investigated at the University of Michigan.¹⁴ This study will be conducted using a protocol similar to the one used at Clemson.

Student teams will be assigned randomly into one of two protocols, with all members of a team using the same instrument at each of four checkpoints.

| Protocol used | Checkpoint 1 instrument | Checkpoint 2 instrument | Checkpoint 3 instrument | Checkpoint 4 instrument |
|---------------|-------------------------|-------------------------|-------------------------|-------------------------|
| Protocol A | VanDuzer and McMartin | 5-factor BARS | VanDuzer and McMartin | 5-factor BARS |
| Protocol B | 5-factor BARS | VanDuzer and McMartin | 5-factor BARS | VanDuzer and McMartin |

Plans for analysis of data collected in Fall 2004 and next steps

Results from Fall 2004 are being compiled and analyzed by Rufus Carter. There is some concern that there was not enough variation in the collected data, and measures are being taken to improve future results—guaranteeing confidentiality, creating an online version of the instrument, and the expanded use of the training process are all expected to improve variability. We hope that the results of the Western Kentucky verbal protocol analysis and the Rose-Hulman time series should identify how students are interpreting the instrument scale and how training affects the instrument’s administration.

More detailed descriptions of each institution’s protocol will be published as results from that institution become available.

Acknowledgments

This material is based upon work supported by NSF DUE-ASA Award Number 0243254, “Designing a Peer Evaluation Instrument that is Simple, Reliable, and Valid.”

References

1. Ohland, M.W., M.L. Loughry, R.L. Carter, and A.G. Yuhasz, “Designing a Peer Evaluation Instrument that is Simple, Reliable, and Valid” *Proc. Amer. Soc. Eng. Ed.*, Salt Lake City, Utah, June 2004.
2. Criteria for Accrediting Engineering Programs. Published by The Accreditation Board for Engineering and Technology (ABET), Baltimore, Maryland. Last accessed on January 5, 2005; http://www.abet.org/images/Criteria/E001_05-06_EAC_Criteria_11-17-04.pdf (criteria approved November 17, 2004).
3. Johnson, D.W., R.T. Johnson, and K.A. Smith, *Active learning: Cooperation in the college classroom*, Edina, MN: Interaction Book Co., (1998).
4. Brown, R.W., “Autorating: Getting individual marks from team marks and enhancing teamwork,” *Proc. Frontiers in Education Conference*. IEEE/ASEE, Pittsburgh, November (1995).
5. Kaufman, D.B., R.M. Felder, and H. Fuller, “Accounting for Individual Effort in Cooperative Learning Teams,” *J. Engr. Education*, **89**(2), 133–140 (2000).
6. Smith, P.C. & Kendall, L.M., Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, **44**: 149-155 (1963).
7. Landy, F.J. & Farr, J.L. “Performance rating,” *Psychological Bulletin*, **87**: 72-107 (1980).
8. Woehr, D.J., and M.J. Miller, “Distributional ratings of performance: more evidence for a new rating format,” *Journal of Management*, Sept-Oct (1997).

9. Hedge, J. W., Bruskiwicz, K. T., Logan, K. K., Hanson, M. A., and Buck, D. *Crew resource management team and individual job analysis and rating scale development for air force tanker crews* (Technical Report No. 336). Minneapolis, MN: Personnel Decisions Research Institutes, Inc. (1999).
10. McGourty, J. and K. De Meuse, *The Team Developer: An assessment and skill building program*, J. Wiley and Company, New York (2000).
11. McGourty, J., C. Sebastian, and W. Swart, "Development of a comprehensive assessment program in engineering education." *J. Engineering Education* **87**(4), 355-361 (1998).
12. McGourty, J., "Using multisource feedback in the classroom: A computer-based approach," *IEEE Trans. Ed.*, **43**(2), 120-124, (2000).
13. Chapman, O. L. and M. A. Fiore. "Calibrated Peer Review™," *Journal of Interactive Instruction Development*. Winter, pp. 11-15 (2000).
14. Van Duzer, E. and F. McMartin, "Methods to improve the validity and sensitivity of a self/peer assessment instrument," *IEEE Trans. Ed.*, **43**(2), 153-158, May (2000).

Author biographies

MATTHEW W. OHLAND

is an Assistant Professor in Clemson University's General Engineering program and is the President of Tau Beta Pi, the national engineering honor society. He received his Ph.D. in Civil Engineering with a minor in Education from the University of Florida in 1996. Previously, he served as Assistant Director of the NSF-sponsored SUCCEED Engineering Education Coalition. His research is primarily in freshman programs and educational assessment.

MISTY L. LOUGHRY

Misty L. Loughry is an Assistant Professor in Clemson University's Management Department. She received her Ph.D. in Management from the University of Florida in 2001. Her research focuses on control in organizations, especially peer monitoring. Prior to her academic career, Dr. Loughry worked in banking for ten years, holding the position of Assistant Vice President of Small Business Lending at the time she left to begin her graduate studies.

RUFUS L. CARTER

Rufus Carter is Coordinator of Institutional Assessment in the Office of Institutional Research and Assessment of Marymount University. He provides consulting services to this project as a measurement and assessment specialist. His research interests include: institutional assessment, student retention, scale development and test validity.

RICHARD M. FELDER

is Hoechst Celanese Professor Emeritus of Chemical Engineering at North Carolina State University. He is coauthor of *Elementary Principles of Chemical Processes* (3rd Edition, John Wiley & Sons, 2000), has authored or coauthored over 200 papers on chemical process engineering and engineering education, and has presented hundreds of seminars, workshops, and short courses in both categories to industrial and research institutions and universities throughout the United States and abroad.

LISA G. BULLARD

Lisa G. Bullard received her BS in ChE from NC State and her Ph.D. in ChE from Carnegie Mellon. She served in engineering and management positions within Eastman Chemical Company from 1991-2000. At N.C. State, she is currently the Director of Undergraduate Studies in Chemical Engineering. Her research interests include the integration of teaming, writing, and speaking into the undergraduate curriculum.