© *Academy of Management Learning & Education*, 2012, Vol. 11, No. 4, 609–630. http://dx.doi.org/10.5465/amle.2010.0177

The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation

MATTHEW W. OHLAND
Purdue University

MISTY L. LOUGHRY
Georgia Southern University

DAVID J. WOEHR
University of North Carolina at Charlotte

LISA G. BULLARD
RICHARD M. FELDER
North Carolina State University

CYNTHIA J. FINELLI
University of Michigan

RICHARD A. LAYTON
Rose-Hulman Institute of Technology

HAL R. POMERANZ
Deer Run Associates

DOUGLAS G. SCHMUCKER
Consultant

Instructors often incorporate self- and peer evaluations when they use teamwork in their classes, which is common in management education. However, the process is often time consuming and frequently does not match well with guidance provided by the literature. We describe the development of a web-based instrument that efficiently collects and analyzes self- and peer-evaluation data. The instrument uses a behaviorally anchored rating scale to measure team-member contributions in five areas based on the team effectiveness literature. Three studies provide evidence for the validity of the new instrument. Implications for management education and areas for future research are discussed.

Peer evaluations are widely used in management education, where teamwork is frequently required, even though it often creates challenges for students and faculty. Teams often have problems, such as team members who work independently rather than collaboratively, poor communication, conflict, differences in team-members' skills, motivation, and goal levels, and free riding or social loafing (Bacon, Stewart, & Silver, 1999; Burdett, 2003; Felder & Brent, 2007; McCorkle, Reardon, Alexander, Kling, Harris, & Iyer, 1999; Verzat, Byrne, & Fayolle, 2009). Students become dissatisfied when they perceive that members of their team do not contribute sufficiently to the team project, forcing them to work harder or get a lower grade than they want (Burdett & Hastie, 2009; Jassawalla, Sashittal, & Malshe, 2009; Oakley, Felder, Brent, & Elhajj, 2004). Instructors often use peer evaluations to deter or remediate these problems, especially free riding, and to assign grades fairly based upon students' contributions (Chen & Lou, 2004; Felder & Brent, 2007; Fellenz, 2006; Kaufman, Felder, & Fuller, 2000; Michaelsen, Knight, & Fink, 2004).

Peer evaluations and self-evaluations can also be used to develop students' team skills. They teach students about teamwork and what is expected of team members; encourage students to reflect on team processes, their own team contributions, and their teammates' contributions; and provide students with developmental feedback (Dominick, Reilly, & McGourty, 1997; Gueldenzoph & May, 2002; Mayo, Kakarika, Pastor, & Brutus, this issue). For these reasons, using peer evaluations appropriately can help students learn to be more effective team members (Brutus & Donia, 2010; Oakley, Felder, Brent, & Elhajj, 2004).

Although research supports using peer evaluations and they are widely used by faculty, there is no consensus about what instrument should be used. A more consistently used system for self- and peer evaluation could reduce the time required for instructors to implement an effective system and allow students to benefit from using a common system across courses.

In the work reported here, we describe the development and testing of a behaviorally anchored rating scale for self- and peer evaluation that is reasonably brief and based upon the research about team-member behaviors that are necessary for effective teamwork. It measures the same categories as the "Comprehensive Assessment of Team Member Effectiveness (CATME)" Likert-style instrument, shown in Appendix A, developed by Loughry, Ohland, and Moore (2007). Our goal is to create a practical tool for self- and peer evaluation that makes it easier for faculty members to effectively manage the teamwork experiences of their students. The instrument, shown in Appendix B, describes team-member behaviors that are characteristic of high, medium, and low performance in each of five categories. The system uses a web-based interface to confidentially collect self- and peer-evaluation data and includes tools for using those data to provide student feedback, adjust grades, and quickly identify rating patterns that warrant the instructor's attention. Three studies provide support for the instrument.

Literature Review

Teams are often used in management education to achieve a number of pedagogical goals and make students more active participants in their education than in traditional coursework (Kolb & Kolb, 2005; Loyd, Kern, & Thompson, 2005; Raelin, 2006; Zantow, Knowlton, & Sharp, 2005). Instructors also use team-based learning methods to develop the interpersonal and teamwork skills that are often lacking in students, yet top the list of skills that recruiters of business students desire (Alsop, 2002; Boni, Weingart, & Evenson, 2009; Calloway School, 2004; The Wall Street Journal/Harris Interactive Survey, 2002; Verzat et al., 2009).

Using self- and peer evaluations is one way instructors can attempt to manage teamwork in their classes to create better teamwork experiences for their students (Gueldenzoph & May, 2002; Hansen, 2006). Students, as members of a team, are better able to observe and evaluate members' contributions than are instructors, who are outsiders to the team (Millis & Cottell, 1998). Instructors use peer evaluations to increase students' accountability to their teammates, motivate students to expend effort at teamwork, reduce free riding, and increase the degree to which students' grades reflect their contributions (Millis & Cottell, 1998). Having teammates who free-ride is a major cause of student dissatisfaction with teamwork (Oakley, Felder, & Brent, 2004; Pfaff & Huddleston, 2003). Instructors often base a portion of students' grades on their peer-evaluation scores or use these scores to adjust students' team grades to reflect their team contributions (Fellenz, 2006). The use of peer eval-

An earlier version of this paper was presented at the 2010 Annual Meeting of the Academy of Management in Montreal, Canada. This material is based upon work supported by the National Science Foundation under Grant nos. 0243254 and 0817403. The authors would like to thank Drs. Lorelle Meadows and Pauline Khan at University of Michigan for allowing their classes to participate in this research. We also want to thank Editor Kenneth Brown and the anonymous reviewers for their helpful feedback.

uations among business students is associated with less social loafing, greater satisfaction with team members' contributions, higher perceived grade fairness, and more positive attitudes toward teamwork (Aggarwal & O'Brien, 2008; Chapman & van Auken, 2001; Erez, LePine, & Elms, 2002; Pfaff & Huddleston, 2003). Self-appraisals are often used with peer evaluations because ratees want to have input in their evaluations and because the information they provide can facilitate a discussion about their performance (Inderrieden, Allen, & Keaveny, 2004).

In addition to motivating students to contribute to their teams, using self- and peer evaluations shows students what is expected of them and how their team contributions will be evaluated (Sheppard, Chen, Schaeffer, Steinbeck, Neumann, & Ko, 2004). Self- and peer evaluations may also be used to provide feedback to improve students' team skills and develop reflective skills and self-management skills that enable students to become lifelong learners (Chen, Donahue, & Klimoski, 2004; Dochy, Segers, & Sluijsmans, 1999; Felder & Brent, 2007; Young & Henquinet, 2000).

Because self- and peer evaluations are often used in work organizations, completing them as part of college classes prepares students for the workplace (Druskat & Wolff, 1999). Self- and peer ratings are used in 360-degree performance appraisal systems and assessment center selection techniques, in addition to executive education and traditional educational settings (Hooijberg & Lane, 2009; London, Smither, & Adsit, 1997; Saavedra & Kwun, 1993; Shore, Shore, & Thornton, 1992). Several meta-analytic studies have found that peer ratings are positively correlated with other rating sources and have good predictive validity for various performance criteria (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Schmitt, Gooding, Noe, & Kirsch, 1984; Viswesvaran, Ones, & Schmidt, 1996; Viswesvaran, Schmidt, & Ones, 2005). Our research adds to the body of education literature supporting the use of peer ratings as a valid part of the evaluation process.

In spite of their benefits, a number of problems are associated with self- and peer ratings. Self-appraisals are vulnerable to leniency errors (Inderrieden et al., 2004). Furthermore, people who are unskilled in an area are often unable to recognize deficiencies in their own skills or performance (Kruger & Dunning, 1999); therefore, people with poor teamwork skills tend to overestimate their abilities and contributions to the team and are less able to accurately appraise their teammates' team skills (Jassawalla et al., 2009). Research shows that many raters, particularly average and below-average performers, do not differentiate in their ratings of team members when it is warranted, sometimes because they worry that providing accurate ratings would damage social relations in the team (Saavedra & Kwun, 1993). Raters also tend to use a social comparison framework and rate members relative to one another rather than using objective, independent criteria (Saavedra & Kwun, 1993). Performing self- and peer ratings can also be stressful for some students (Pope, 2005). Even when students have favorable attitudes toward group work and the use of peer evaluations, they may worry that they lack enough training to rate their peers and that peer ratings may be biased (Walker, 2001).

Peer-Evaluation Systems

Having a well-designed peer-evaluation instrument is useful so that the peer-evaluation system can teach students which teamwork behaviors are important and how to evaluate team-member contributions (Young & Henquinet, 2000). Furthermore, if the peer-evaluation instrument is well-designed, any feedback that students receive from it is likely to have more value for student learning. Various peer- and self-evaluation instruments and approaches have been described in the management education literature and in the pedagogical literature from other disciplines, yet none has gained widespread acceptance.

Some systems ask students to divide a certain number of points among team members (sometimes with restrictions, such as forbidding students from distributing the points equally; e.g., Erez et al., 2002; Michelsen et al., 2004; Saavedra & Kwun, 1993). Point distribution systems are common because they are simple and yield a score that can be used for grading; however, point distributions do not give students information about what teamwork behaviors are important, and so they do not teach team skills or provide specific feedback for improvement. Peer rankings or peer nominations can also be used, although they are uncommon in business education. Any method that forces students to differentiate ratings or rankings of team members can be criticized as threatening team cohesiveness or risking team members' colluding so that no team member gets a better score (Baker, 2008).

Some instructors allow students to create their own criteria by which they will evaluate one another, so that students will feel more ownership of the evaluation criteria and hopefully work harder to meet the standards that they developed for themselves (Thomas, Martin, & Pleasants, 2011).

Other instructors develop evaluation instruments that relate closely to the particular group assignments in their classes.

Evaluation systems that use rating scales have been developed for more general use (see Baker, 2008, for a review). Examples include an 87-item scale (Rosenstein & Dickinson, 1996); a 7-factor, 52-item scale (Harris & Barnes-Farrell, 1997); a 4-factor (communication, decision making, collaboration, self-management), 50-item measure (McGourty & De Meuse, 2001); a 14-factor, 46-item measure based on Stevens and Campion's (1999, 1994) selection research (Taggar & Brown, 2001); and a 35-item scale measuring competence (but not quality of work contributed), task and maintenance orientation, domineering behavior, dependability (attending meetings), and free-riding behavior (Paswan & Gollakota, 2004).

Some evaluation systems require combinations of point distributions, rating scales, or open comments (e.g., Brooks & Ammons, 2003; Brutus & Donia, 2010; Davis et al., 2010; Fellenz, 2006; Gatfield, 1999; Willcoxson, 2006). Drawbacks of these systems are their length and complexity and the amount of time required for students and faculty to use them. For example, the Van Duzer and McMartin (2000) instrument asks for Likert-scale responses on three items about the team as a whole, self- and peer ratings on 11 items for all team members, plus written open-ended comments, and asks students to label each team-member's role on the team, nominate the member who provided the most leadership on the team, describe what they learned about teamwork, and distribute a fixed number of points among team members.

CATME Likert-Scale Instrument

To meet the need for a research-based peer-evaluation instrument for use in college classes, Loughry et al. (2007) developed the CATME. The researchers searched the teamwork literature to identify the ways by which team members can help their teams to be effective. Based upon the literature, they created a large pool of potential items to evaluate for their peer-evaluation instrument. They then tested these items using two large surveys of college students. They used both exploratory and confirmatory factor analysis to select the items to retain for the final instrument and group them into factors that reflect college students' perceptions of team-member contributions.

The researchers found 29 specific types of team-member contributions that clustered into five broad categories (Contributing to the Team's Work, Interacting with Teammates, Keeping the Team on Track, Expecting Quality, and Having Relevant Knowledge, Skills, and Abilities). The full (87-item) version of the instrument uses three items to measure each of the 29 types of team-member contributions with high internal consistency. The short (33-item) version uses a subset of the items to measure the five broad categories (these items are shown in Appendix A). Raters rate each teammate on each item using Likert scales (*strongly disagree–strongly agree*).

The CATME instrument reflects one research-based model of team-member contributions. There are other respected models of teamwork, many of which have considerable conceptual overlap with CATME, yet have different foci and different categories. One model that is highly influential is Salas, Sims, and Burke's (2005) "big five in teamwork." It has a high degree of overlap with CATME; however, it applies to highly interdependent work teams in which shared mental models, closed-loop communication, and mutual trust exist. The "big five" model, therefore, assumes that team members will have the skills and motivation to contribute effectively to the team, yet these are frequently key deficiencies in student teams.

Although the CATME instrument is solidly rooted in the literature on team effectiveness and was created for use with college students, many instructors who wish to use peer evaluations may need a more pragmatic, easier to administer instrument. Even the short version of the CATME requires that students read 33 items and make judgments about each item for each of their teammates. If there are 4-person teams and the instructor requires a self-evaluation, each student must make 132 independent ratings to complete the evaluation. To consider all of these decisions carefully may require more effort than students are willing to put forth and may generate more data than instructors have time to review carefully. Although the Loughry et al. (2007) paper has been cited in other research about peer evaluation of teamwork in high school, college, and the workplace (e.g., Hughes & Jones, 2011; Wang, MacCann, Zhuang, Liu, & Roberts, 2009; Davis et al., 2010; Zhang & Ohland, 2009; Zhu, Chen, & Lu, 2010), the full instrument has not been used in other published research, and there is no evidence that it is widely used by college faculty. The number of ratings required may be a key reason why the CATME instrument, like other instruments cited earlier, has not been widely adopted.

Another concern about the CATME instrument is that students using the Likert-scale format to make their evaluations may have different perceptions about which rating a teammate deserves. This is

because the response choices (*strongly disagree–strongly agree*) do not describe the behaviors that are associated with the various rating levels.

Behaviorally Anchored Ratings Scale Instrument

Although the CATME Likert-type instrument has a number of strengths, there are benefits of developing a behaviorally anchored rating scale (BARS) instrument that measures the five categories of team-member contributions identified by the CATME research. In a 4-person team, this would reduce the number of rating decisions from 132 decisions per rater with the short form of the CATME instrument to 20 with a BARS instrument. Furthermore, by providing descriptions of the behaviors that a team member would display to warrant a particular rating, a BARS instrument could teach students what constitutes good performance and poor performance, building students' knowledge about teamwork. If the students, in an attempt to earn a good score on the evaluation, try to display more team-member behaviors associated with high ratings and refrain from behaviors at the low end of the scales, using the BARS instrument could result in students contributing more effectively to their teams.

Behaviorally anchored rating scales provide a way to measure how an individual's behavior in various performance categories contributes to achieving the goals of the team or organization of which the individual is a member (Campbell, Dunnette, Arvey, & Hellervik, 1973). The procedure for creating BARS instruments was developed in the early 1960s and became more popular in the 1970s (Smith & Kendall, 1963). Subject-matter experts (SMEs), who fully understand the job for which the instrument is being developed, provide input for its creation. SMEs provide specific examples of actual performance behaviors, called "critical incidents," and classify whether the examples represent high, medium, or low performance in the category in question. The scales provide descriptions of specific behaviors that people at various levels of performance would typically display.

Research on the psychometric advantages of BARS scales has been mixed (MacDonald & Sulsky, 2009); however, research suggests that BARS scales have a number of advantages over Likert ratings scales, such as greater interrater reliability and less leniency error (Campbell et al., 1973; Ohland, Layton, Loughry, & Yuhasz, 2005). In addition, instruments with descriptive anchors may generate more positive rater and ratee reactions, have more face validity, and offer advantages for raters from collectivist cultures (MacDonald & Sulsky, 2009). Developing behavioral anchors also facilitates frame-of-reference training, which has been shown to be the best approach to rater training (Woehr & Huffcutt, 1994). In addition, providing training with the BARS scale prior to using it to evaluate performance may make the rating process easier and more comfortable for raters and ratees (MacDonald & Sulsky, 2009).

DEVELOPMENT OF THE CATME-B INSTRUMENT

A cross-disciplinary team of nine university faculty members with expertise in management, education, education assessment, and engineering education collaborated to develop a BARS version of the CATME instrument, which we will refer to as "CATME-B." We used the "critical incident methodology" described in Hedge, Bruskiewicz, Logan, Hanson, and Buck (1999) to develop behavioral anchors for the five broad categories of team-member contributions measured by the original CATME instrument. This method is useful for developing behavioral descriptions to anchor a rating scale and requires the identification of examples of a wide range of behaviors from poor to excellent (rather than focusing only on extreme cases). Critical incidents include observable behaviors related to what is being assessed, context, and the result of the behavior, and must be drawn from the experience of subject-matter experts, who also categorize and translate the list of critical incidents. All members of the research team can be considered as subject-matter experts on team learning and team-member contributions. All have published research relating to student learning teams and all have used student teams in their classrooms. Collectively, they have approximately 90 years of teaching experience.

We began by working individually to generate examples of behaviors that we have observed or have heard described by members of teams. These examples came from our students' teams, other teams that we have encountered in our research and experience, and teams (such as faculty committees) of which we have been members. We noted which of the five categories of team contributions we thought that the observation represented. We exchanged our lists of examples and then tried to generate more examples after seeing each other's ideas.

We then held several rounds of discussions to review our examples and arrive at a consensus about which behaviors described in the critical incidents best represented each of the five categories in the CATME Likert instrument. We came to an agreement that we wanted 5-point response

scales with anchors for high, medium and low team-member performance in each category. We felt that having more than five possible ratings per category would force raters to make fine-grained distinctions in their ratings that would increase the cognitive effort required to perform the ratings conscientiously. This would have been in conflict with our goal of creating an instrument that was easier and less time consuming to use than the original CATME.

We also agreed that the instrument's medium level of performance (3 on the 5-point scale) should describe satisfactory team-member performance; therefore, the behaviors that anchored the "3" rating would describe typical or average team-member contributions. The behavioral descriptions that would anchor the "5" rating would describe excellent team contributions, or things that might be considered as going well above the requirements of being a fully satisfactory team member. The "1" anchors would describe behaviors that are commonly displayed by poor or unsatisfactory team members and, therefore, are frequently heard complaints about team members. The "2" and "4" ratings do not have behavioral anchors, but provide an option for students to assign a rating between the levels described by the anchors.

Having agreed on the types of behaviors that belonged in each of the five categories and on our goal to describe outstanding, satisfactory, and poor performance in each, we then worked to reach consensus about the descriptions that we would use to anchor each category. We first made long lists of descriptive statements that everyone agreed represented fairly the level of performance for the category. Then we developed more precise wording for the descriptions and prioritized the descriptions to determine which behaviors were most typical of excellent, satisfactory, and poor performance in each category.

We engaged in considerable discussion about how detailed the descriptions for each anchor should be. Longer, more-detailed descriptions that listed more aspects of behavior would make it easier for raters using the instrument to recognize that the ratee's performance was described by the particular anchor. Thus, raters could be more confident that they were rating accurately; however, longer descriptions require more time to read and, thus, are at odds with our goal of designing an instrument that is simple and quick to complete. Our students have often told us that the longer an instrument is, the less likely they are to read it and conscientiously answer each question. We, therefore, decided to create an instrument that could fit on one page. We agreed that three bulleted lines of description would be appropriate for each level of behavior. The instrument we developed after many rounds of exchanging drafts among the team members and gathering feedback from students and colleagues is shown in Appendix B.

PILOT TESTING AND DEVELOPMENT OF WEB INTERFACE

We pilot tested the CATME-B instrument in several settings to improve the instrument and our understanding of the factors important for a smooth administration. Our final pilot test was in a junior-level (3rd year) introduction to management course at a large university in South Carolina in spring 2005. This course had 30 sections taught by five graduate assistants. Because the BARS format is less familiar to students than is the Likert format, we explained how to use the BARS scale, which took about 10 minutes. It then took about 10 minutes for students in 3- to 4-member teams to conscientiously complete the self- and peer evaluations. To lessen peer pressure, students were seated away from their teammates.

After our pilot testing, we decided it would be beneficial to develop a web-based administration of the instrument. A web administration provides confidentiality for the students, causes only one factor to be displayed on the screen at a time (reinforcing the independence of the factors), makes it easier for instructors to gather the self- and peer-evaluation data, eliminates the need for instructors to type the data into a spreadsheet, and makes it easier for instructors to interpret the results. After extensive testing to ensure that the web interface worked properly, we conducted three studies of the web-based instrument.

STUDY 1: EXAMINING THE PSYCHOMETRIC PROPERTIES OF CATME-B

The primary goal of Study 1 was to evaluate the psychometric characteristics of the web-based CATME-B instrument relative to the paper-based CATME instrument. We, therefore, administered both measures to a sample of undergraduate students engaged in team-based course-related activity. We then used an application of generalizability (G) theory (Brennan, 1994; Shavelson & Webb, 1991) to evaluate consistency and agreement for each of the two rating forms. G theory is a statistical theory developed by Cronbach, Glesser, Nanda, and Rajaratnam (1972) about the consistency of behavioral measures. It uses the logic of analysis of variance (ANOVA) to differentiate mul-

tiple systemic sources of variance (as well as random measurement error) in a set of scores. In addition, it allows for the calculation of a summary coefficient (i.e., a generalizability coefficient) that reflects the dependability of measurement. This generalizability coefficient is analogous to the reliability coefficient in classical test theory. A second summary coefficient (i.e., dependability coefficient) provides a measure of absolute agreement across various facets of measurement. In essence, generalizability (as reflected in these indices) provides a direct assessment of measurement invariance across multiple facets of measurement. As such, G theory may be used to establish measurement equivalence (e.g., Sharma & Weathers, 2003). Thus, G theory provides the most appropriate approach for examining the psychometric characteristics of the CATME instruments as well as the degree of equivalence across the two measures.

In this study, each participant rated each of his or her teammates using each of the two CATME instruments (at two different times). Three potential sources of variance in ratings were examined: person effects, rater effects, and scale effects. In the G theory framework, person effects represent "true scores," while rater and scale effects represent potential sources of systematic measurement error. In addition to these main effects, variance estimates were also obtained for Person × Rater, Person × Scale, and Scale × Rater 2-way interactions, and the 3-way Person × Scale × Rater Interaction (it is important to note that in G theory, the highest level interaction is confounded with random measurement error). All interaction terms are potential sources of measurement error. This analysis allowed for a direct examination of the impact of rating scale as well as a comparison of the measurement characteristics of each scale.

Method

Participants and Procedure

Participants were 86 students in two sections of a sophomore-level (2nd year) course (each taught by a different instructor) at a large university in North Carolina in fall of 2005. Participants were assigned to one of 17 teams, consisting of 4–5 team members each. The teams were responsible for completing a series of nine group homework assignments worth 20% of students' final grades.

Participants were randomly assigned to one of two measurement protocols. Participants in protocol A (n = 48) completed the CATME-B instrument at Time 1 (about 4 weeks after the teams were formed) and the CATME instrument at Time 2 (approximately 6 weeks later during the last week of class). Participants were told that they would be using two different survey formats for the mid-semester and end-of-semester surveys to help determine which format was more effective. Participants in protocol B (n = 38) completed the CATME instrument at Time 1 and the CATME-B at Time 2.

All participants rated each of their teammates at both measurement occasions; however, 13 participants did not turn in ratings for their teammates. Thus, ratings were obtained from 73 raters, resulting in a response rate of 85% (56 participants received ratings from 4 team members, 16 participants received ratings from 3 team members, and 1 participant received ratings from only 1 team member). Therefore, the total dataset consisted of a set of 273 ratings ([56 × 4] + [16 × 3] + 1). This set of ratings served as input into the G theory analyses.

Measures

All participants completed both the CATME and web-based CATME-B. Ratings on the Likert-type scale were made on a 5-point scale (i.e., 1 = *Strongly Disagree*, 5 = *Strongly Agree*). Likert dimension scales were scored as the mean item response across items corresponding to the dimension.

Results

Descriptive statistics for each dimension on each scale are provided in Table 1. Reliabilities (Cronbach's alpha) for each Likert dimension are also provided in Table 1. This index of reliability could not be calculated for the CATME-B instrument because each team member was only rated on one item per dimension, per rater; however, G theory analysis does provide a form of reliability (discussed in more detail later) that can be calculated for both the Likert-type and BARS scales. Correlations among the BARS dimensions and Likert-type dimensions are provided in Table 2. Cross-scale, within-dimension correlations were modest (i.e., .41–.59), with the exception of those for the dimensions of Interacting with Teammates (0.74) and Keeping the Team on Track (0.74), which were adequate. It is important, however, to note that the two scales were administered at two different times (separated by 6 weeks) and, thus, the ratings likely represent different samples of behavior (i.e., performance may have changed over time). Also of interest from the correlation matrix is the fact that correlations between the dimensions for the BARS scale are lower than correlations be-

TABLE 1
Descriptive Statistics for CATME BARS and Likert-Type Scale Dimensions in Study 1

	N	*Min*	*Max*	*M*	*SD*	*α*
Bars scale						
Contributing to the team's work	260	2.00	5.00	4.59	0.62	—
Interacting with teammates	264	1.00	5.00	4.33	0.75	—
Keeping the team on track	264	1.00	5.00	4.20	0.78	—
Expecting quality	264	1.00	5.00	4.18	0.75	—
Having relevant KSAs	264	2.00	5.00	4.24	0.71	—
Likert-type scales[α]						
Contributing to the team's work	273	1.88	5.00	4.31	0.56	.90
Interacting with teammates	273	2.20	5.00	4.32	0.54	.91
Keeping the team on track	273	2.14	5.00	4.19	0.59	.87
Expecting quality	273	2.50	5.00	4.60	0.49	.81
Having relevant KSAs	273	2.25	5.00	4.44	0.54	.78

[α] Likert-type dimension scales were scored as the mean item response across items corresponding to the dimension.

tween dimensions for the Likert scale. This may indicate that the BARS scale offers better discrimination between dimensions than the Likert scale.

G theory analyses were conducted using the MIVQUE0 method by way of the SAS VARCOMP procedure. This method makes no assumptions regarding the normality of the data, can be used to analyze unbalanced designs, and is efficient (Brennan, 2000). Results of the G theory analysis are provided in Table 3. These results support the convergence of the BARS and Likert-type scales. Specifically, the type of scale used explained relatively little variance in responses (e.g., 0% for Interacting with Teammates and Keeping the Team on Track, and 5, 10, and 18% for Having Relevant Knowledge, Skills and Abilities (KSAs), Contributing to the Team's Work, and Expecting Quality, respectively). Also, ratee effects explained a reasonable proportion of the variance in responses (e.g., 31% for Contributing to the team's work). Generalizability coefficients calculated for each dimension were adequate (Table 3). Note that Table 3 presents two coefficients: ρ (rho), which is also called the *generalizability coefficient* (the ratio of universe score variance to itself plus relative error variance) and is analogous to a reliability coefficient in classical test theory (Brennan, 2000); and φ (phi), also called the *dependability coefficient* (the ratio of universe score variance to itself plus absolute error variance), which provides an index of absolute agreement.

It is important to note that, in this study, rating scale format is confounded with time of measurement. Thus, the variance associated with scale (i.e., the 2-way interactions as well as the main effect) may reflect true performance differences

TABLE 2
Correlations Among CATME BARS and Likert-Type Scale Dimensions in Study 1

		Bars scale					Likert-type scale				
		1	2	3	4	5	6	7	8	9	10
	Bars scale										
1	Contributing to the team's work	—									
2	Interacting with teammates	.37	—								
3	Keeping the team on track	.26	.44	—							
4	Expecting quality	.35	.51	.38	—						
5	Having relevant KSAs	.35	.47	.50	.52	—					
	Likert-type scale										
6	Contributing to the team's work	**.59**	.63	.55	.51	.65	—				
7	Interacting with teammates	.40	**.74**	.66	.54	.58	.71	—			
8	Keeping the team on track	.47	.66	**.74**	.71	.78	.80	.81	—		
9	Expecting quality	.43	.50	.52	**.41**	.50	.68	.63	.65	—	
10	Having relevant KSAs	.73	.48	.38	.44	**.45**	.73	.57	.64	.59	—

Notes. $N = 260–273$.
Same dimension cross-scale correlations are in bold.
All correlations are significant ($p < .01$).

TABLE 3
Variance Estimates and Generalizability Coefficients for Each CATME Dimension in Study 1

Component variance	Contributing to the team's work		Interacting with teammates		Keeping the team on track		Expecting quality		Having relevant KSAs	
	Estimate	%	Estimate	%	Estimate	%	Estimate	%	Estimate	%
Ratee	.12	31.45	.09	20.51	.07	15.19	.03	6.94	.06	15.65
Rater	.04	11.02	.14	33.10	.19	39.45	.06	12.24	.05	11.98
Scale	.04	9.95	.00	0.00	.00	0.00	.09	18.37	.02	5.13
Ratee × Rater	.04	10.48	.06	14.92	.07	15.61	.05	10.20	.06	14.67
Ratee × Scale	.02	6.18	.03	7.69	.02	4.43	.01	1.22	.02	4.40
Rater × Scale	.05	13.44	.04	9.32	.05	10.55	.17	34.69	.12	30.07
Error	.07	17.47	.06	14.45	.07	14.77	.08	16.33	.07	18.09
Total variance	.37		.43		.47		.49		.41	
ρ (rho)	.85		.76		.75		.70		.75	
φ (phi)	.71		.64		.59		.31		.58	

Notes. Scale = BARS vs. Likert-type.

(i.e., differences between the two measurement occasions). As a result, the previous analyses may underestimate the consistency of measurement associated with the CATME. Consequently, we conducted a second set of G theory analyses in which we analyzed the data separately for each scale. The results from the second G theory analysis are presented in Table 4. These results were fairly consistent with the variance estimates provided by the first set of analyses. Generalizability coefficients (similar to estimates of reliability) calculated for each dimension on each scale are also provided in Table 4. These estimates indicate similar characteristics for both rating scale formats.

Discussion of Study 1

The primary goal of Study 1 was to examine the psychometric characteristics of the CATME-B relative to that of the previously developed CATME instrument. Results indicate that both scales provide relatively consistent measures of the teamwork dimensions assessed.

The results of Study 1 do not, however, address the extent to which the rating scale correlates with

TABLE 4
Variance Estimates and Generalizability Coefficients by Scale Type for Each CATME Dimension in Study 1

Component variance	Contributing to the team's work		Interacting with teammates		Keeping the team on track		Expecting quality		Having relevant KSAs	
	Estimate	%	Estimate	%	Estimate	%	Estimate	%	Estimate	%
Bars scale										
Ratee	.14	36.86	.17	29.31	.08	12.77	.05	8.93	.05	10.52
Rater	.11	30.89	.20	35.00	.31	50.90	.35	60.42	.27	52.58
Ratee × Rater	.12	32.25	.21	35.69	.22	36.33	.18	30.65	.19	36.90
Total variance	.37		.58		.61		.57		.50	
ρ (rho)	.90		.87		.74		.70		.70	
φ (phi)	.82		.77		.54		.44		.48	
Likert-type scale										
Ratee	.15	47.60	.06	20.43	.09	26.63	.03	10.64	.10	37.82
Rater	.08	25.88	.17	60.57	.17	51.18	.13	53.19	.09	31.64
Ratee × Rater	.08	26.52	.05	19.00	.08	22.19	.09	36.17	.08	30.55
Total variance	.31		.28		.34		.24		.28	
ρ (rho)	.93		.90		.91		.70		.91	
φ (phi)	.88		.67		.74		.49		.83	

external variables of interest. Therefore, we conducted a second study.

STUDY 2: EXAMINING THE VALIDITY OF THE CATME-B MEASURE

The objective of Study 2 was to examine the extent to which the CATME-B ratings were related to course grades and the peer-evaluation scale created by Van Duzer and McMartin (2000). We chose this instrument as a comparison because it was well designed and a group of scholars at several prominent universities (Arizona State, University of Massachusetts, Dartmouth, University of Alabama, Rose-Hulman Institute of Technology, University of Wisconsin, and Texas A&M) were advocating for more consistency in the assessment of team skills and felt that the Van Duzer and McMartin instrument was suitable. These were scholars in the engineering education community who were trying to find a solution for accreditation requirements that programs demonstrate that students can work effectively in teams (ABET, 2000). The Van Duzer and McMartin instrument has been cited in a number of published studies, but none used the instrument in the research.

In this study, we expect a high degree of convergence between the two peer-evaluation measures, both in terms of overall ratings and individual items. In addition, it is important to note that the Van Duzer and McMartin (2000) instrument was developed as a general measure of "teamwork skills" and is scored as a single overall scale; whereas, the CATME-B specifies five different dimensions of teamwork. Thus, it is difficult to posit specific hypotheses with respect to the individual items comprising the Van Duzer and McMartin instrument. Nonetheless, we expect: (a) significant correlations between the majority of Van Duzer and McMartin items and the CATME-B dimensions, and (b) a differential pattern of relationships across the five CATME-B dimensions.

We further expect that both measures will be significantly related to final course grades in courses requiring team activity. The reason is that individuals' performance on teamwork and overall performance tend to be correlated because general mental ability, conscientiousness, and interpersonal skills facilitate both types of work (Neuman & Wright, 1999; Offerman, Bailey, Vasilopoulos, Seal, & Sass, 2004; Stevens & Campion, 1999). In addition, course grades are influenced by team performance in courses requiring teamwork, and so, if the peer-evaluation measures more effective team contributions, the teams should perform better and earn higher grades on the teamwork portion of the grade. Furthermore, one dimension of the CATME-B scale measures whether team members have the knowledge, skills, and abilities (KSAs) to contribute to the team, which should be similar to the KSAs needed to do well in the course.

Method

Participants and Procedure

Participants in this study were 104 students enrolled in four discussion sections of a freshman-level (1st year) course at a large university in Michigan during the fall 2005 semester. We chose this course because 40% of each student's grade was based on team-based projects. All participants were assigned to 5-member teams based upon a number of criteria, including grouping students who lived near one another and matching team-member schedules. Six students were excluded from the final data analysis due to missing data; thus, the effective sample size was 98, which resulted in a response rate of 94%.

As part of their coursework, students worked on a series of team projects over the course of the semester. All participants were required to provide peer-evaluation ratings of each of their teammates at four times, which occurred during Weeks 5, 8, 11, and 13 of the semester. Teams were randomly assigned to one of two groups, which determined the order in which they received the instruments. All participants completed two administrations of the CATME-B (CATME-B 1 and CATME-B 2) and two administrations of the Van Duzer and McMartin instrument (VM 1 and VM 2). Group A participants used the Van Duzer and McMartin measure to provide ratings at Times 1 and 3 and the CATME-B measure at Times 2 and 4. Group B participants used the CATME-B measure to provide ratings at Times 1 and 3 and the Van Duzer and McMartin measure at Times 2 and 4. When all scores on a particular measure are aggregated, the sample includes data collected at all four time points and data from both groups of participants. This approach controls for group differences, changes in the measured characteristics with time, and order effects, such as novelty.

Measures

CATME-B. Participants used the web-based CATME-B measure to provide two sets of peer evaluations. As in Study 1, there was a high degree of intercorrelation among the five CATME performance dimensions (mean $r = .76$). Consequently, we formed a composite index representing an

overall rating as the mean rating across the five dimensions (coefficient $\alpha = .94$). In addition, the correlation between the CATME rating composite across the two administrations was .44. We averaged the composite measures from the two administrations to form an overall CATME rating composite for each participant.

Van Duzer and McMartin (VM) Scale. The peer-evaluation measure presented by Van Duzer and McMartin (2000) is comprised of 11 rating items. Respondents provided ratings on 4-point scales of agreement (1 = *Disagree*, 2 = *Tend to Disagree*, 3 = *Tend to Agree*, 4 = *Agree*). Examination of the interitem correlations indicated a high degree of overlap across items (mean $r = .48$). Thus, we formed a composite score as the mean rating across the 11 items. The coefficient alpha for the composite was .91. The composite scores from first and second VM administrations had a correlation of .55. We averaged the composite measures from the two administrations to form an overall VM rating composite for each participant.

Course Grades. Final course grades were based upon a total of 16 requirements. Each requirement was differentially weighted such that the total number of points obtainable was 1,000. Of the total points, up to 50 points were assigned on the basis of class participation, which was directly influenced by the peer evaluations. Thus, we excluded the participation points from the final course score, resulting in a maximum possible score of 950.

Results

Descriptive statistics and intercorrelations for all study variables are presented in Table 5. It is interesting to note that, for both the CATME-B and VM, ratings decreased between the first and second administration. We next examined the correlation of each of the two overall rating composites with final course grades. As expected, both the CATME-B and VM measures were significantly related to final course grades ($r = .51$ for the CATME-B; $r = .60$ for the VM). It is also interesting that the correlation between the composite peer ratings and the final course score increases as the semester progresses. The peer rating correlates with final course score .27 at administration 1 with the CATME-B; .38 at administration 1 with the VM scale; .54 at administration 2 with the CATME-B; and .61 at administration 2 with the VM scale. In addition, the overall mean ratings decrease over administrations, while rating variability increases.

The results indicate a high level of convergence between the CATME-B and VM composites ($r = .64$, $p < .01$). In addition, the five components that make up the CATME ratings demonstrated differential relationships with the eleven individual VM items (with correlations ranging from −.14 to .67. $M = .30$; MDN = .30; $SD = .28$). Thus, we also examined this set of correlations for coherence. As can be seen in Table 6, the pattern of correlations was highly supportive of convergence across the two measures. Specifically, all of the VM items (with the exception of item 6) demonstrated significant correlations with one or more of the CATME items. The VM items appear to be most highly related, however, to the "Contributing to the Team's Work" CATME dimension.

Discussion of Study 2

The results of Study 2 provide additional validity evidence for the CATME-B scale. As expected, the CATME-B scale demonstrates a high degree of convergence with another published peer-evaluation measure and a significant relationship with final course grades in a course requiring a high level of

TABLE 5
Descriptive Statistics for and Correlations Among All Study Variables in Study 2

	M	*SD*	CATME-B 1[a]	CATME-B 2[b]	CATME-B composite	VM 1[a]	VM 2[b]	VM composite	Course points
CATME-B 1[a]	4.20	.38	—						
CATME-B 2[b]	4.07	.59	.44	—					
CATME-B composite	4.14	.41	.77	.91	—				
VM 1[a]	3.72	.19	.53	.53	.62	—			
VM 2[b]	3.64	.37	.30	.58	.55	.55	—		
VM composite	3.68	.25	.42	.63	.64	.79	.95	—	
Course points	822.74	53.55	.27	.54	.51	.38	.61	.60	—

Notes. $N = 98$.
[a] First administration.
[b] Second administration.
All correlations are significant ($p < .01$).

TABLE 6
Correlations Between VM Items and CATME-B Dimensions in Study 2

	VM items	CATME-B Dimensions				
		Contributing to the team's work	Interacting with teammates	Keeping the team on track	Expecting quality	Having relevant KSAs
1	Failed to do an equal share of the work. (R)	.57**	.44**	.49**	.38**	.43**
2	Kept an open mind/was willing to consider other's ideas.	.17	.37**	.13	.02	.14
3	Was fully engaged in discussions during meetings.	.56**	.47**	.51**	.34**	.46**
4	Took a leadership role in some aspects of the project.	.65**	.45**	.57**	.42**	.54**
5	Helped group overcome differences to reach effective solutions.	.64**	.59**	.63**	.49**	.62**
6	Often tried to excessively dominate group discussions. (R)	-.11	.05	-.07	-.14	-.07
7	Contributed useful ideas that helped the group succeed.	.66**	.52**	.64**	.48**	.54**
8	Encouraged group to complete the project on a timely basis.	.55**	.37**	.51**	.32**	.37**
9	Delivered work when promised/needed.	.55**	.40**	.45**	.41**	.29**
10	Had difficulty negotiating issues with members of the group. (R)	.23*	.27**	.31**	.09	.17
11	Communicated ideas clearly/effectively.	.67**	.57**	.64**	.45**	.59**

Note. $N = 398$.
(R) = reverse scored.
*$p < .05$. **$p < .01$.

team interaction. It should also be noted the CATME-B instrument offers a more efficient measure of peer performance than does the Van Duzer and McMartin (2000) instrument. That is, the CATME-B measure requires five ratings per teammate versus eleven (a difference that is quickly compounded as the number of teammates or times of measurement increase) and offers more differentiation across multiple aspects of performance.

STUDY 3: CATME SCORES AS PREDICTORS OF TEAMMATES' ATTITUDES

The final study examines how the peer ratings that students receive on the CATME-B instrument are related to teammates' attitudes toward the students. We expected that scores on the CATME-B would be positively associated with the degree to which teammates like the student and would want to work with the student again. These are important outcomes of team-members' contributions and behaviors. Student teams sometimes barely hold together to finish a one-semester project; yet in the workplace, *team viability*, which is the ability of teams to continue working cooperatively, is an important dimension of team effectiveness (Hackman, 1987). It is, therefore, important that students learn to cooperate in ways that create sustainable working relationships.

Research has found that liking and interpersonal attraction are related to team viability and team-member satisfaction (Barrick, Stewart, Neubert, & Mount, 1998). Workload sharing (similar to the CATME-B dimension "Contributing to the Team's Work"), and communication and flexibility (related to the CATME-B dimensions of "Interacting with Teammates" and "Keeping the Team on Track") are highly correlated with social cohesion (meaning that team members want to stay in the group), and social cohesion is a good predictor of team viability (Barrick et al., 1998).

We, therefore, expected that there would be variation in the pattern of correlations among the five dimensions of CATME-B and students' liking of their teammates and the degree to which they would want to work with that teammate again. For example, we thought that the correlation between students' scores on "Interacting with Teammates" would be the strongest predictor of whether they liked the student because high scores on that dimension of CATME-B would indicate interpersonal effectiveness, which would make someone more pleasant to be around. We expected that scores on "Contributing to the Team's Work" would be positively associated with teammates wanting to work with the student again because a peer who contributes a high quantity and quality of work increases the chances of the team getting a good grade and reduces the workload for teammates.

Method

Participants and Procedure

Participants were 570 students working in 113 teams of 3–7 people (mode = 3) in 19 sections of

junior- and senior-level (3rd and 4th year) management and marketing courses at a large university in Georgia between fall 2007 and summer 2010. Students completed the CATME-B instrument on-line as part of the requirements for teamwork in their classes. Students were asked follow-up questions on-line to measure the dependent variables.

Measures

Liking. We used two items from Jehn and Mannix (2001) and created a third item to measure the extent to which teammates like the student. These were (1) I like this person as an individual; (2) I consider this person to be a friend; and (3) I enjoy spending time with this person. Students answered using the scale: 1 = *Strongly Disagree*, 2 = *Disagree*, 3 = *Neither Agree nor Disagree*, 4 = *Agree*, 5 = *Strongly Agree*. The coefficient alpha for these items was .86.

Future. We created three items to measure the extent to which teammates would want to work with the student again in the future. These were (1) I would gladly work with this individual in the future; (2) If I were selecting members for a future work team, I would pick this person; and (3) I would avoid working with this person in the future (reverse scored). The response choices were the same as above. The coefficient alpha for these items was .92.

Results

For our analysis, we only used data for which we had at least three teammates' ratings of an individual team member (self-ratings were not used), which resulted in a sample size of 358. We conducted an exploratory factor analysis with the six items for Future and Liking and the items loaded as expected on two separate factors. Descriptive statistics and intercorrelations for all study variables are presented in Table 7.

We then computed a mean (across teammates) rating on each of the five CATME-B dimensions for each ratee. Next, we regressed both the "Liking" composite and the "Future" composite on the set of mean ratings for the five CATME-B dimensions. Results of the regression analyses are presented in Table 8. As expected, ratings on the five CATME-B dimensions accounted for a significant proportion of variance in both the "Liking" ($R^2 = .26$) and "Future" ($R^2 = .58$) composites. Interestingly, the ratings accounted for significantly more variance in the "Future" composite than the "Liking" composite. This suggests that peers might base their intentions to work with an individual again more on the quality of work in the present group than whether they like the individual.

As expected, different dimensions of CATME-B had different relationships with the dependent variables. "Interacting with Teammates" ($\beta = .16$, $p < .05$) and "Keeping the Team on Track" ($\beta = .25$, $p < .05$) were significant predictors of teammates "Liking" the student. The results were stronger for the "Future" variable. "Contributing to the Team's Work" ($\beta = .45$, $p < .01$) and "Interacting with Teammates" ($\beta = .26$, $p < .01$) were significant predictors of teammates wanting to work with that student again in the "Future."

Discussion of Study 3

The results of Study 3 provide evidence for the predictive validity of the CATME-B instrument. The study showed that scores on different dimensions of the CATME-B have different relationships with teammates' attitudes toward students in two important areas: liking the students and wanting to work with them again in the future. The finding that teammates who were rated higher on "Keeping the Team on Track," were more liked by their teammates was somewhat surprising. "Keeping the Team on Track" involves monitoring teammates and the external environment for conditions

TABLE 7
Descriptive Statistics for and Correlations Among All Study Variables in Study 3

	M	*SD*	1	2	3	4	5	6	7
1. Future composite	4.22	.69	1.00						
2. Liking composite	4.25	.47	.63	1.00					
3. Contributing to the team's work	4.02	.62	.74	.39	1.00				
4. Interacting with teammates	4.05	.53	.66	.38	.71	1.00			
5. Keeping the team on track	3.99	.56	.68	.40	.82	.72	1.00		
6. Expecting quality	4.06	.54	.61	.37	.75	.65	.77	1.00	
7. Having relevant KSAs	4.17	.51	.61	.33	.77	.65	.78	.70	1.00

Note. $N = 358$.
All correlations are significant ($p < .01$).

TABLE 8
Regression Equations for Study 3

Dependent variable	Independent variables	*R*	R^2	Beta
Liking composite				
	Contributing to the team's work			.158
	Interacting with teammates			.159*
	Keeping the team on track			.248*
	Expecting quality			.118
	Having relevant KSAs			−.140
		.509	.259	
Future composite				
	Contributing to the team's work			.452**
	Interacting with teammates			.265**
	Keeping the team on track			.113
	Expecting quality			.039
	Having relevant KSAs			−.044
		.760	.578	

Note. *$p < .05$. **$p < .01$.

that could affect the team's success and taking action, such as providing constructive feedback when problems are discovered. We often hear anecdotally from our students that they are reluctant to monitor their teammates and to discuss problems with them for fear of creating social conflict and being disliked. Therefore, we speculate that perhaps it is when team members like one another and feel comfortable with their relationship with each other that they might feel freer to engage in monitoring and feedback behaviors that might be more stressful if the relationship were not as strong.

The strongest predictor of teammates wanting to work with a student again was, not surprisingly, ratings on "Contributing to the Team's Work." Having teammates who do a large quantity and quality of work makes it easier for students to earn high grades with less effort, and team members who score high in this area are not the low contributors who are often resented for their "free riding." Scores on "Interacting with Teammates" also strongly predicted the degree to which teammates would want to work with the student again. Students who do well in this area would make the teamwork experience more pleasant and make others feel more valued.

GENERAL DISCUSSION

In the work reported here, we developed and tested a behaviorally anchored rating scale version (CATME-B) of the previously published Comprehensive Assessment of Team Member Effectiveness (CATME). This instrument is for self- and peer evaluation by team members and may be useful to instructors who use team learning methods and need a relatively short and simple instrument that is closely aligned with the literature on team-member effectiveness. The instrument uses behavioral descriptions to anchor three levels of performance in five categories of team-member performance.

In general, our results provide a high level of support for the CATME-B instrument as a peer-evaluation measure. Specifically, the instrument demonstrates equivalent psychometric characteristics to the much longer Likert version of CATME. In addition, it demonstrates a high degree of convergence with another peer-evaluation measure and a significant relationship with final course grades in a course requiring a high level of team interaction. Ratings on the instrument are also associated with the degree to which teammates like the student and would want to work with him or her in the future.

Features Added to Enhance Usability

Because the initial testing of CATME-B showed that the instrument was viable, we created additional features to make the on-line system a more useful tool for students and instructors. Security features were added so that team members could complete the survey from any computer with Internet access using a password-protected log-in, increasing both convenience and privacy for students. This matters because research shows that it is important for peer evaluations to be confidential (Bamberger, Erev, Kimmel, & Oref-Chen, 2005).

A comments field was added so that students could type comments that are only visible to their instructors. Instructors could insist that their students use the comments field to justify their ratings, in order to increase accountability for providing accurate ratings (Mero, Guidice, & Brownlee, 2007). Comments can also facilitate discussions between instructors and students.

To lessen the time required for instructors to review and interpret the data, the system calculates average ratings for each student in each of the five CATME categories, compares those to the team average, and computes a ratio of each team member's score to the team average. This results in "grade adjustment factors" (two scores are computed—one that includes the self-rating one that does not). Instructors can use one of these scores to adjust grades, providing an alternative for the many instructors who have used point-distribution systems in the past. Instructors who do not use the scores for grading can quickly scan them to see which students had high or low ratings relative to their teammates.

The system flags seven "exceptional conditions" in the rating patterns that may warrant the instructor's attention. Some are team-level situations, such as patterns indicating that the team has experienced conflict or split into cliques. Others are individual-level situations, such as high- or low-performing team members or situations in which a team member rates him- or herself high and all teammates low.

Instructors can use the system to electronically provide feedback to students. It provides a visual display of the student's self-rating, the average of how teammates rated the student, and the team average for each of the five categories measured by the instrument. Instructors control whether and when they release this feedback to their students.

To familiarize students with the BARS rating format and the teamwork behaviors measured by the CATME-B instrument, we developed a practice-rating exercise. Students rate four fictitious team members whose performance is described in written scenarios and then receive feedback on how their ratings compare to those of expert raters. Repeated use of a peer-evaluation system increases students' confidence and skills in rating their peers (Brutus, Donia, & Ronen, In press). The practice-rating exercise should help students to improve their confidence and rating skill before they rate their actual teammates. As noted below, an important area for future research is to develop web-based rater training and team-skills training to expand the potential for this instrument to positively impact student learning.

To provide additional support for managing teamwork, the web interface links with the Team-Maker tool. Instructors choose the criteria they will use to assign students to teams and how each will be weighted. Team-Maker collects the information from students, creates teams, and gives students the names and e-mail addresses of their teammates and the times that team members said they would be available for meetings (Layton, Loughry, Ohland, & Ricco, 2010).

The new instrument and added features appear to meet the need for a practical peer-rating tool that instructors find useful. More than 2,600 faculty users (over 20% of whom are in business-related disciplines) at over 575 universities currently use the system. More than 120,000 unique students have used the system; many of whom have used it in multiple courses. Instructors who want to use the system can request a free account at www.catme.org.

Implications for Management Education

By offering practical solutions for problems that instructors and students face when they are involved with teamwork, we believe that this tool meets the challenge issued by Bell (2010: 7) to engage in work that will "matter beyond citations and impact factors." The tool has the potential to make teamwork in management classes less frustrating for students and instructors, teach students about teamwork, and, by reducing teamwork problems that interfere with learning, facilitate learning about other management topics.

This system can also facilitate teamwork in other higher education disciplines. Because management is the discipline to which people from many fields turn for science-based advice on how to manage people and workflow, and management scholars conduct a large amount of the research on teamwork, it is appropriate for management education research to provide solutions that can also apply to other educational contexts.

Challenges, Limitations, and Future Research Needs

Our results reflect several problems that are familiar in peer-evaluation research. Although we designed the instrument so that a "3" score would be satisfactory performance and a "5" would be excellent performance, the average score for all five dimensions was approximately 4.2 across our three studies. Students in these studies did not

appear to use the full range of the scale, resulting in a restriction of range problem with the data. Although this is a common problem in peer-evaluation research for a variety of reasons, including social pressures to give high ratings (Taggar & Brown, 2006), it reduces the accuracy of the peer ratings and makes it difficult to obtain strong correlations with criterion variables. However, using self- and peer evaluations is likely to encourage members to contribute more effectively to their teams, even when there is not a large variance in the scores and most students do not have a significant grade adjustment as a result of the peer evaluations (Johnston & Miles, 2004). If the system deters free riding and quickly draws the instructors' attention to those few teams with problems, it will be a useful tool for managing student teams.

There were also high correlations among the five factors in the CATME-B instrument. Research finds that team members who are highly skilled in areas related to the team's work also display better social skills and contribute more to team discussions (Sonnetag & Volmer, 2009). Thus, there is probably a substantial amount of real correlation among CATME-B categories due to some team members being stronger or weaker contributors in many or all of the five areas measured by the instrument. For example, team members with strong relevant knowledge, skills, and abilities are more likely to contribute highly to the team's work than students who lack the necessary skills to contribute. The high correlations among the CATME-B dimensions, however, may also indicate the presence of halo error, which occurs when peers' perceptions of a teammate as a good or bad team member affect their ratings in specific areas. A meta-analysis found that correlations among different dimensions of job performance rated by peers are inflated by 63% due to halo error (Viswesvaran et al., 2005).

The magnitude of the rater effects observed in this research also warrants discussion. In essence, these are main effect differences across raters, which, in conjunction with the smaller Ratee $\times$ Rater interaction effects, indicate that while raters are fairly consistent in their rank-ordering of ratees, there are mean differences across ratees. One explanation could be that while raters are fairly consistent in the pattern of ratings they assign, there may be calibration issues. In other words, some raters may be more lenient or severe in their ratings than others. Rater training that focuses on calibrating raters to performance level would be an appropriate mechanism for reducing these effects. Frame-of-reference training is an established technique for training raters in performance appraisal settings that could help to make self- and peer appraisals more accurate in educational settings (Woehr, 1994; Woehr & Huffcutt, 1994). The BARS format of this instrument will facilitate that training.

An alternative explanation for the observed rater effects is that they represent real differences in the behavior of a ratee with respect to different team members. The fact that rater effects are most substantial for the dimensions "Interacting with Teammates" and "Keeping the Team on Track" supports this interpretation because these dimensions are more idiosyncratic to specific teammates (i.e., a team member may not interact with all teammates in the same way). Thus, rater effects may represent real differences and not rating error. Unfortunately, in the present study, there is no way to determine which explanation for the observed rater effects is more correct. Even if it could be shown that some of the rater effects represented real observed differences in behavior rather than rater error, rater training could still be useful to facilitate student learning.

Instructors may be able to teach students to rate more accurately by training students about the dimensions of teamwork represented in the rating scale and giving students practice using the ratings instrument before they use the instrument to rate their teammates. By learning to rate teamwork, students should better understand how to effectively contribute to teams, thus building teamwork skills. The timely feedback that the web-based CATME-B system provides may also enhance students' learning of team skills, particularly when students have the opportunity to practice their skills in several time periods and receive multiple rounds of feedback from the system (Fellenz, 2006; Hess, 2007). Brutus and Donia (2010) showed that using peer evaluations can improve students' team skills. Future research should examine whether this self- and peer-evaluation tool helps students learn team skills, and whether rater training adds to this learning.

In addition to facilitating student learning, training business students in teamwork may encourage them to use more effective team processes, resulting in better teamwork experiences (Bacon et al., 1999). Having more positive team experiences would enable students to enjoy teamwork more. One study shows that mature communication, accountable interdependence, psychological safety, common purpose, and role clarity within student teams account for 71.7% of variance in students' attitudes' toward teamwork (Ulloa & Adams, 2004).

Although, with training, students may develop the ability to accurately rate teammates' perfor-

mance, training will not motivate students to rate accurately. The audience to whom a rater feels accountable has a big impact on rating patterns (Mero et al., 2007). To improve the accuracy of peer ratings and improve the reliability and validity of peer-evaluation scores, future research should identify ways to change the conditions that motivate team members to rate their teammates based upon considerations other than their contributions to the team. Past research has found that both students and employees in organizations are reluctant to evaluate their peers, particularly for administrative purposes, such as adjusting grades or determining merit-based pay (Bettenhausen & Fedor, 1997; Sheppard et al., 2004). Individuals who are required to participate in peer-evaluation systems often resist the systems because they are concerned that peer evaluations will be biased by friendships, popularity, jealousy, or revenge. Recent research suggests that these concerns may be well-founded (Taggar & Brown, 2006). Students who received positive peer ratings then liked their teammates better and rated them higher on subsequent peer evaluations; whereas students who received negative peer ratings liked their teammates less and gave them lower ratings on subsequent peer evaluations. This was the case even though only aggregated feedback from multiple raters was provided.

In Study 3 we viewed students liking one another and wanting to work with one another in the future as outcomes of members' team-related interactions and contributions to the team. It is possible, however, that some students had relationships with one another prior to working together on the team, or interactions with one another outside of the team context during the semester, that could have affected ratings on these variables at the end of the semester. Future studies that measure students' contact with one another outside of the team context could help to determine the degree to which interpersonal affect among team members is a source of bias in peer-evaluation scores versus an outcome of team members' behavior and team contributions.

CONCLUSIONS

As management educators, we should do more than just "use" teams in the classroom: We should leverage them as a context within which to teach about teams and teamwork. The research presented here offers a tool that instructors may be able to use to achieve this goal.

Instructors could use the instrument before teams begin their work to set expectations for team-member behavior. The instrument can also be used multiple times as the teams' work progresses, to provide feedback and hold students accountable for their team contributions. This could help to make teamwork less frustrating and more rewarding for both students and instructors.

APPENDIX A

Comprehensive Assessment of Team Member Effectiveness—Likert Short Version

Contributing to the Team's Work
- Did a fair share of the team's work.
- Fulfilled responsibilities to the team.
- Completed work in a timely manner.
- Came to team meetings prepared.
- Did work that was complete and accurate.
- Made important contributions to the team's final product.
- Kept trying when faced with difficult situations.
- Offered to help teammates when it was appropriate.

Interacting With Teammates
- Communicated effectively.
- Facilitated effective communication in the team.
- Exchanged information with teammates in a timely manner.
- Provided encouragement to other team members.
- Expressed enthusiasm about working as a team.
- Heard what teammates had to say about issues that affected the team.
- Got team input on important matters before going ahead.
- Accepted feedback about strengths and weaknesses from teammates.
- Used teammates' feedback to improve performance.
- Let other team members help when it was necessary.

Keeping the Team on Track
- Stayed aware of fellow team members' progress.
- Assessed whether the team was making progress as expected.
- Stayed aware of external factors that influenced team performance.
- Provided constructive feedback to others on the team.
- Motivated others on the team to do their best.
- Made sure that everyone on the team understood important information.
- Helped the team to plan and organize its work.

Expecting Quality
- Expected the team to succeed.
- Believed that the team could produce high-quality work.
- Believed that the team should achieve high standards.
- Cared that the team produced high-quality work.

Having Relevant Knowledge, Skills, and Abilities (KSAs)
- Had the skills and expertise to do excellent work.
- Had the skills and abilities that were necessary to do a good job.
- Had enough knowledge of teammates' jobs to be able to fill in if necessary.
- Knew how to do the jobs of other team members.

APPENDIX B

Comprehensive Assessment of Team Member Effectiveness—Behaviorally Anchored Rating Scale (BARS) Version

	Your name					← **Write the names of the people on your team including your own name.** **This self and peer evaluation asks about how you and each of your teammates contributed to the team during the time period you are evaluating. For each way of contributing, please read the behaviors that describe a "1", "3," and "5" rating. Then confidentially rate yourself and your teammates.**
Contributing to the Team's Work	**5**	**5**	**5**	**5**	**5**	• Does more or higher-quality work than expected. • Makes important contributions that improve the team's work. • Helps to complete the work of teammates who are having difficulty.
	4	**4**	**4**	**4**	**4**	Demonstrates behaviors described in both 3 and 5.
	3	**3**	**3**	**3**	**3**	• Completes a fair share of the team's work with acceptable quality. • Keeps commitments and completes assignments on time. • Fills in for teammates when it is easy or important.
	2	**2**	**2**	**2**	**2**	Demonstrates behaviors described in both 1 and 3.
	1	**1**	**1**	**1**	**1**	• Does not do a fair share of the team's work. Delivers sloppy or incomplete work. • Misses deadlines. Is late, unprepared, or absent for team meetings. • Does not assist teammates. Quits if the work becomes difficult.
Interacting with Teammates	**5**	**5**	**5**	**5**	**5**	• Asks for and shows an interest in teammates' ideas and contributions. • Improves communication among teammates. Provides encouragement or enthusiasm to the team. • Asks teammates for feedback and uses their suggestions to improve.
	4	**4**	**4**	**4**	**4**	Demonstrates behaviors described in both 3 and 5.
	3	**3**	**3**	**3**	**3**	• Listens to teammates and respects their contributions. • Communicates clearly. Shares information with teammates. Participates fully in team activities. • Respects and responds to feedback from teammates.
	2	**2**	**2**	**2**	**2**	Demonstrates behaviors described in both 1 and 3.
	1	**1**	**1**	**1**	**1**	• Interrupts, ignores, bosses, or makes fun of teammates. • Takes actions that affect teammates without their input. Does not share information. • Complains, makes excuses, or does not interact with teammates. Accepts no help or advice.
Keeping the Team on Track	**5**	**5**	**5**	**5**	**5**	• Watches conditions affecting the team and monitors the team's progress. • Makes sure that teammates are making appropriate progress. • Gives teammates specific, timely, and constructive feedback.
	4	**4**	**4**	**4**	**4**	Demonstrates behaviors described in both 3 and 5.
	3	**3**	**3**	**3**	**3**	• Notices changes that influence the team's success. • Knows what everyone on the team should be doing and notices problems. • Alerts teammates or suggests solutions when the team's success is threatened.
	2	**2**	**2**	**2**	**2**	Demonstrates behaviors described in both 1 and 3.
	1	**1**	**1**	**1**	**1**	• Is unaware of whether the team is meeting its goals. • Does not pay attention to teammates' progress. • Avoids discussing team problems, even when they are obvious.
Expecting Quality	**5**	**5**	**5**	**5**	**5**	• Motivates the team to do excellent work. • Cares that the team does outstanding work, even if there is no additional reward. • Believes that the team can do excellent work.
	4	**4**	**4**	**4**	**4**	Demonstrates behaviors described in both 3 and 5.
	3	**3**	**3**	**3**	**3**	• Encourages the team to do good work that meets all requirements. • Wants the team to perform well enough to earn all available rewards. • Believes that the team can fully meet its responsibilities.
	2	**2**	**2**	**2**	**2**	Demonstrates behaviors described in both 1 and 3.
	1	**1**	**1**	**1**	**1**	• Satisfied even if the team does not meet assigned standards. • Wants the team to avoid work, even if it hurts the team. • Doubts that the team can meet its requirements.
Having Relevant Knowledge, Skills, and Abilities	**5**	**5**	**5**	**5**	**5**	• Demonstrates the knowledge, skills, and abilities to do excellent work. • Acquires new knowledge or skills to improve the team's performance. • Able to perform the role of any team member if necessary.
	4	**4**	**4**	**4**	**4**	Demonstrates behaviors described in both 3 and 5.
	3	**3**	**3**	**3**	**3**	• Has sufficient knowledge, skills, and abilities to contribute to the team's work. • Acquires knowledge or skills needed to meet requirements. • Able to perform some of the tasks normally done by other team members.
	2	**2**	**2**	**2**	**2**	Demonstrates behaviors described in both 1 and 3.
	1	**1**	**1**	**1**	**1**	• Missing basic qualifications needed to be a member of the team. • Unable or unwilling to develop knowledge or skills to contribute to the team. • Unable to perform any of the duties of other team members.

REFERENCES

ABET. 2000. ***Criteria for Accrediting Engineering Programs.*** Accreditation Board for Engineering and Technology [Abet], Baltimore, MD.

Aggarwal, P., & O'Brien, C. L. 2008. Social loafing on group projects structural antecedents and effect on student satisfaction. ***Journal of Marketing Education,*** 30: 255–264.

Alsop, R. 2002. Playing well with others. ***The Wall Street Journal,*** September 9, 2002: R11.

Bacon, D. R., Stewart, K. A., & Silver, W. S. 1999. Lessons from the best and worst student team experiences: How a teacher can make the difference. ***Journal of Management Education,*** 23: 467–488.

Baker, D. F. 2008. Peer assessment in small groups: A comparison of methods. ***Journal of Management Education,*** 33: 183–209.

Bamberger, P. A., Erev, I., Kimmel, M., & Oref-Chen, T. 2005. Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity. ***Group and Organization Management,*** 30: 344–377.

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. 1998. Rating member ability and personality to work-team processes and team effectiveness. ***Journal of Applied Psychology,*** 83: 377–391.

Bell, M. P. 2010. From the editors: What if . . . ? Diversity, scholarship, and impact. ***Academy of Management Learning and Education,*** 9: 5–10.

Bettenhausen, K. L., & Fedor, D. B. 1997. Peer and upward appraisals: A comparison of their benefits and problems. ***Group and Organization Management,*** 22: 236–263.

Boni, A. A., Weingart, L. R., & Evenson, S. 2009. Innovation in an academic setting: Designing and leading a business through market-focused, interdisciplinary teams. ***Academy of Management Learning and Education,*** 8: 407–417.

Brennan, R. L. 1994. Variance components in generalizability theory. In C. R. Reynolds (Ed.), ***Cognitive assessment: A multidisciplinary perspective:*** 175–207. New York: Plenum Press.

Brennan, R. L. 2000. Performance assessments from the perspective of generalizability theory. ***Applied Psychological Measurement,*** 24: 339–353.

Brooks, C. M., & Ammons, J. L. 2003. Free riding in group projects and the effects of timing, frequency, and specificity of criteria in peer assessments. ***Journal of Education for Business,*** 78: 268–272.

Brutus, S., & Donia, M. B. 2010. Improving the effectiveness of students in groups with a centralized peer evaluation system. ***Academy of Management Learning and Education,*** 9: 652–662.

Brutus, S., Donia, M., & Ronen, S. In press. Can business students learn to evaluate better? Evidence from repeated exposure to a peer evaluation process. ***Academy of Management Learning and Education.***

Burdett, J. 2003. Making groups work: University students' perceptions. ***International Education Journal,*** 4: 177–190.

Burdett, J., & Hastie, B. 2009. Predicting satisfaction with group work assignments. ***Journal of University Teaching and Learning Practice,*** 6(1): 62–71.

Calloway School of Business and Accountancy of Wake Forest University. ***A Report on Recruiters' Perceptions of Undergraduate Business Schools and Students.*** February 2004.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. 1973. The development and evaluation of behaviorally based rating scales. ***Journal of Applied Psychology,*** 57: 15–22.

Chapman, K. J., & van Auken, S. 2001. Creating positive group project experiences: An examination of the role of the instructor on students' perceptions of group projects. ***Journal of Marketing Education,*** 23: 117–127.

Chen, G., Donahue, L. M., & Klimoski, R. J. 2004. Training undergraduates to work in organizational teams. ***Academy of Management Learning and Education,*** 3: 27–40.

Chen, Y., & Lou, H. 2004. Students' perceptions of peer evaluation: An expectancy perspective. ***Journal of Education for Business,*** 79: 275–282.

Conway, J. M., & Huffcutt, A. I. 1997. Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisory, peer, and self-ratings. ***Human Performance,*** 10: 331–360.

Cronbach, L. J., Glesser, G. C., Nanda, H., & Rajaratnam, N. 1972. ***The dependability of behavioral measurements: Theory of generalizability for scores and profiles.*** New York: John Wiley & Sons.

Davis, D., Trevisan, M., Gerlick, R., Davis, H., McCormack, J., Beyerlein, S., Thompson, P., Howe, S., Leiffer, P., & Brackin, P. 2010. Assessing team member citizenship in capstone engineering design courses. ***International Journal of Engineering Education,*** 26: 1–13.

Dochy, F., Segers, M., & Sluijsmans, D. 1999. The use of self-, peer and co-assessment in higher education: A review. ***Studies in Higher Education,*** 24: 331–350.

Dominick, P. G., Reilly, R. R., & McGourty, J. W. 1997. The effects of peer feedback on team member behavior. ***Group and Organization Management,*** 22: 508–520.

Druskat, V. U., & Wolff, S. B. 1999. Effects and timing of developmental peer appraisals in self-managing work groups. ***Journal of Applied Psychology,*** 84: 58–74.

Erez, A., LePine, J. A., & Elms, H. 2002. Effects of rotated leadership and peer evaluation on the functioning and effectiveness of self-managed teams: A quasi-experiment. ***Personnel Psychology,*** 55: 929–248.

Felder, R. M., & Brent, R. 2007. Cooperative learning. In P. A. Mabrouk, ed., ***Active learning: Models from the analytical sciences,*** ACS Symposium Series, 970: 34–53. Washington, DC: American Chemical Society.

Fellenz, M. R. 2006. Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions. ***Journal of Management Education,*** 30: 570–591.

Gatfield, T. 1999. Examining student satisfaction with group projects and peer assessment. ***Assessment and Evaluation in Higher Education,*** 24: 365–377.

Gueldenzoph, L. E., & May, G. L. 2002. Collaborative peer evaluation: Best practices for group member assessments. ***Business Communication Quarterly,*** 65: 9–20.

Hackman, J. R. 1987. The design of work teams. In ***Handbook of organizational behavior:*** 315–342. Englewood Cliffs, NJ: Prentice Hall.

Hansen, R. S. 2006. Benefits and problems with student teams: Suggestions for improving team projects. ***Journal of Education for Business,*** 82: 11–19.

Harris, M. M., & Schaubroeck, J. 1988. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. ***Personnel Psychology,*** 41: 43–62.

Harris, T. C., & Barnes-Farrell, J. L. 1997. Components of teamwork: Impact on evaluations of contributions to work team effectiveness. ***Journal of Applied Social Psychology,*** 27: 1694–1715.

Hedge, J. W., Bruskiewicz, K. T., Logan, K. K., Hanson, M. A., & Buck, D. 1999. Crew resource management team and individual job analysis and rating scale development for air force tanker crews. ***Technical Report:*** 336. Minneapolis, MN: Personnel Decisions Research Institutes, Inc.

Hess, P. W. 2007. Enhancing leadership skill development by creating practice/feedback opportunities in the classroom. ***Journal of Management Education,*** 31: 195–213.

Hooijberg, R., & Lane, N. 2009. Using multisource feedback coaching effectively in executive education. ***Academy of Management Learning and Education,*** 8: 483–493.

Hughes, R. L., & Jones, S. K. 2011. Developing and assessing college student teamwork skills. ***New Directions for Institutional Research,*** 2011: 53–64.

Inderrieden, E. J., Allen, R. E., & Keaveny, T. J. 2004. Managerial discretion in the use of self-ratings in an appraisal system: The antecedents and consequences. ***Journal of Managerial Issues,*** 16: 460–482.

Jassawalla, A., Sashittal, H., & Malshe, A. 2009. Students' perceptions of social loafing: It's antecedents and consequences in undergraduate business classroom teams. ***Academy of Management Learning and Education,*** 8: 42–54.

Jehn, K. A., & Mannix, E. A. 2001. The dynamic nature of conflict: A longitudinal study of intragroup conflict and group performance. ***Academy of Management Journal,*** 44: 238–251.

Johnston, L., & Miles, L. 2004. Assessing contributions to group assignments. ***Assessment and Evaluation in Higher Education,*** 29: 751–768.

Kaufman, D. B., Felder, R. M., & Fuller, H. 2000. Accounting for individual effort in cooperative learning teams. ***Journal of Engineering Education,*** 89: 133–140.

Kolb, A. Y., & Kolb, D. A. 2005. Learning styles and learning spaces: Enhancing experiential learning in higher education. ***Academy of Management Learning and Education,*** 4: 193–212.

Kruger, J., & Dunning, D. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. ***Journal of Personality and Social Psychology,*** 77: 1121–1134.

Layton, R. A., Loughry, M. L., Ohland, M. W., & Ricco, G. D. 2010. Design and validation of a web-based system for assigning members to teams using instructor-specified criteria. ***Advances in Engineering Education,*** 2: 1–28.

London, M., Smither, J. W., & Adsit, D. J. 1997. Accountability: The Achilles' heel of multisource feedback. ***Group and Organization Management,*** 22: 162–184.

Loughry, M. L., Ohland, M. W., & Moore, D. D. 2007. Development of a theory-based assessment of team member effectiveness. ***Educational Psychological Measurement,*** 67: 505–524.

Loyd, D. L., Kern, M. C., & Thompson, L. 2005. Classroom research: Bridging the ivory divide. ***Academy of Management Learning and Education,*** 4: 8–21.

MacDonald, H. A., & Sulsky, L. M. 2009. Rating formats and rater training redux: A context-specific approach for enhancing the effectiveness of performance management. ***Canadian Journal of Behavioural Science,*** 41: 227–240.

McCorkle, D. E., Reardon, J., Alexander, J. F., Kling, N. D., Harris, R. C., & Iyer, R. V. 1999. Undergraduate marketing students, group projects, and teamwork: The good, the bad, and the ugly? ***Journal of Marketing Education,*** 21: 106–117.

McGourty, J., & De Meuse, K. P. 2001. ***The team developer: An assessment and skill building program. Student guidebook.*** New York: John Wiley & Sons.

Mero, N. P., Guidice, R. M., & Brownlee, A. L. 2007. Accountability in a performance appraisal context: The effect of audience and form of accounting on rater response and behavior. ***Journal of Management,*** 33: 223–252.

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (Eds.). 2004. ***Team-based learning: A transformative use of small groups in college teaching.*** Sterling, VA: Stylus Publishing.

Millis, B. J., & Cottell, P.G., Jr. 1998. ***Cooperative Learning for Higher Education Faculty.*** Phoenix, AZ: American Council on Education/Oryx Press.

Neuman, G. A., & Wright, J. 1999. Team effectiveness: Beyond skills and cognitive ability. ***Journal of Applied Psychology,*** 84: 376–389.

Oakley, B., Felder, R. M., Brent, R., & Elhajj, I. 2004. Turning student groups into effective teams. ***Journal of Student-Centered Learning,*** 2: 9–34.

Offerman, L. R., Bailey, J. R., Vasilopoulos, N. L., Seal, C., & Sass, M. 2004. The relative contribution of emotional competence and cognitive ability to individual and team performance. ***Human Performance,*** 17: 219–243.

Ohland, M. W., Layton, R. A., Loughry, M. L., & Yuhasz, A. G. 2005. Effects of behavioral anchors on peer evaluation reliability. ***Journal of Engineering Education,*** 94: 319–326.

Paswan, A. K., & Gollakota, K. 2004. Dimensions of peer evaluation, overall satisfaction, and overall evaluation: An investigation in a group task environment. ***Journal of Education for Business,*** 79: 225–231.

Pfaff, E., & Huddleston, P. 2003. Does it matter if I hate teamwork? What impacts student attitudes toward teamwork. ***Journal of Marketing Education,*** 25: 37–45.

Pope, N. K. L. 2005. The impact of stress in self- and peer assessment. ***Assessment & Evaluation in Higher Education,*** 30: 51–63.

Raelin, J. 2006. Does action learning promote collaborative leadership? ***Academy of Management Learning and Education,*** 5: 152–168.

Rosenstein, R., & Dickinson, T. L. (1996, August). The teamwork components model: An analysis using structural equation modeling. In Rosenstein, R., (Chair), ***Advances in definitional team research.*** Symposium conducted at the annual meeting of the American Psychological Association, Toronto.

Saavedra, R. & Kwun, S. K. 1993. Peer evaluation in self-managing work groups. ***Journal of Applied Psychology,*** 78: 450–462.

Salas, E., Sims, D. E., & Burke, C. S. 2005. Is there a "big five" in teamwork? ***Small Group Research,*** 36, 555–599.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. 1984. Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. ***Personnel Psychology,*** 37: 407–422.

Sharma, S., & Weathers, D. 2003. Assessing generalizability of scales used in cross-national research. ***Research in Marketing, 20:*** 287–295.

Shavelson, R. J., & Webb, N. M. 1991. ***Generalizability Theory.*** Newbury Park: Sage Publications, Inc.

Sheppard, S., Chen, H. L., Schaeffer, E., Steinbeck, R., Neumann, H., & Ko, P. 2004. ***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.

Shore, T. H., Shore, L. M., & Thornton, G. C., III 1992. Construct validity of self-and peer evaluations of performance dimensions in an assessment center. ***Journal of Applied Psychology,*** 77: 42–54.

Smith, P. C., & Kendall, L. M. 1963. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. ***Journal of Applied Psychology,*** 47: 149–155.

Sonnetag, S., & Volmer, J. 2009. Individual-level predictors of task-related teamwork processes: The role of expertise and self-efficacy in team meetings. ***Group & Organization Management,*** 34: 37–66.

Stevens, M. J., & Campion, M. A. 1999. Staffing work teams: Development and validation of a selection test for teamwork settings. ***Journal of Management,*** 25: 207–228.

Stevens, M. J., & Campion, M. A. 1994. The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. ***Journal of Management,*** 20: 503–530.

Taggar, S., & Brown, T. C. 2006. Interpersonal affect and peer rating bias in teams. ***Small Group Research,*** 37: 87–111.

Taggar, S., & Brown, T. C. 2001. Problem-solving team behaviors: Development and validation of BOS and a hierarchical factor structure. ***Small Group Research,*** 32: 698–726.

The Wall Street Journal/Harris Interactive Survey. 2002. The rating criteria. ***The Wall Street Journal,*** September 9, 2002: R5.

Thomas, G., Martin, D., & Pleasants, K. 2011. Using self- and peer-assessment to enhance students' future-learning in higher education. ***Journal of University Teaching & Learning Practice,*** 8(1): article 5.

Ulloa, B. C. R., & Adams, S. G. 2004. Attitude toward teamwork and effective teaming. ***Team Performance Management,*** 10: 145–151.

Van Duzer, E., & McMartin, F. 2000. Methods to improve the validity and sensitivity of a self/peer assessment instrument. ***IEEE Transactions on Education,*** 43: 153–158.

Verzat, C., Byrne, J., & Fayolle, A. 2009. Tangling with spaghetti: Pedagogical lessons from games. ***Academy of Management Learning and Education,*** 8: 356–369.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. 1996. Comparative analysis of the reliability of job performance ratings. ***Journal of Applied Psychology,*** 81: 557–574.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. 2005. Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. ***Journal of Applied Psychology,*** 90: 108–131.

Walker, A. 2001. British psychology students' perceptions of group-work and peer assessment. ***Psychology Learning and Teaching,*** 1: 28–36.

Wang, L., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. 2009. Assessing teamwork and collaboration in high school students: A multimethod approach. ***Canadian Journal of School Psychology,*** 24, 108–124.

Willcoxson, L. E. 2006. "It's not fair!": Assessing the dynamics and resourcing of teamwork. ***Journal of Management Education,*** 30: 798–808.

Woehr, D. J. 1994. Understanding frame-of-reference training: The impact of training on the recall of performance information. ***Journal of Applied Psychology,*** 79: 525–534.

Woehr, D. J., & Huffcutt, A. I. 1994. Rater training for performance appraisal: A quantitative review. ***Journal of Occupational and Organizational Psychology,*** 67: 189–206.

Young, C. B., & Henquinet, J. A. 2000. A conceptual framework for designing group projects. ***Journal of Education for Business,*** 76: 56–60.

Zantow, K., Knowlton, D. S., & Sharp, D. C. 2005. More than fun and games: Reconsidering the virtues of strategic management simulations. ***Academy of Management Learning and Education,*** 4: 451–458.

Zhang, B., & Ohland, M.W. 2009. How to assign individualized scores on a group project: An empirical evaluation. ***Applied Measurement in Education,*** 22: 290–308.

Zhu, J., Chen, S., & Lu, Q. 2010. Measuring member performance in multi-functional R&D teams: An empirical study with GAHP analysis. ***2010 IEEE International Conference on Industrial Engineering and Engineering Management IEEM):*** 507–5111.

Matthew W. Ohland is a professor of engineering education at Purdue University. Ohland earned his PhD from the University of Florida in Civil Engineering. His research interests include team assignment, peer evaluation, the longitudinal study of engineering students, and the adoption of active and collaborative teaching methods.

Misty L. Loughry is a professor of management at Georgia Southern University. Loughry earned her PhD in management from the University of Florida. Her research interests include organizational control, especially peer control, and teamwork, including peer evaluation of team-member contributions and peer influences in team settings.

David J. Woehr (PhD, Georgia Tech) is professor and chair of the Department of Management at the University of North Carolina at Charlotte. He is an associate editor for *Human Performance* and serves on the editorial boards for *Organizational Research Methods*, and the *European Journal of Work and Organizational Psychology*.

Lisa G. Bullard is a teaching professor in chemical and biomolecular engineering at North Carolina State University. Bullard received her PhD in chemical engineering from Carnegie Mellon University. Dr. Bullard's research interests lie in the area of educational scholarship, including teaching and advising effectiveness, academic integrity, and process design.

Richard M. Felder is Hoechst Celanese Professor Emeritus of Chemical Engineering at North Carolina State University. Felder received his PhD in chemical engineering from Princeton in 1966 and coauthored *Elementary Principles of Chemical Processes* (3rd Edition, Wiley, 2005) and roughly 300 papers on chemical process engineering and engineering education.

Cynthia J. Finelli holds a PhD in electrical engineering from the University of Michigan, where she currently is director of the Center for Research on Learning and Teaching in Engineering and research associate professor of engineering education. Her research interests include studying faculty motivation to adopt effective teaching practices and exploring ethical decision making in engineering students.

Richard A. Layton is an associate professor of mechanical engineering at Rose-Hulman Institute of Technology with a PhD from the University of Washington. His professional work includes student teaming, persistence, migration, and retention of engineering undergraduates, as well as consulting in data visualization and graph design. Layton is also a singer and songwriter.

Hal R. Pomeranz is the lead developer of the CATME web interface. He is founder and technical lead of Deer Run Associates, a company specializing in open systems and security. Pomeranz holds a BA in mathematics with a minor in computer science from Swarthmore College.

Douglas G. Schmucker earned his PhD in structural engineering from Stanford University. Schmucker currently develops on-line engineering courses as a private consultant and has interests in investigative engineering, service engineering, and engineering teams. He can be reached at doug_schmucker@yahoo.com.