Peer Ratings

THE IMPACT OF PURPOSE ON RATING QUALITY AND USER ACCEPTANCE

JIING-LIH FARH
Louisiana State University
ALBERT A. CANNELLA, Jr.
Texas A&M University
ARTHUR G. BEDEIAN
Louisiana State University

Using a quasi-experimental design, the effects of purpose (evaluative vs. developmental) on both peer-rating quality and user acceptance were examined. Subjects were 65 undergraduates divided into 11 project groups. Six groups conducted peer ratings for evaluative (i.e., grading) purposes, whereas the remaining 5 did so for the purpose of providing developmental feedback. Peer ratings conducted for evaluative purposes tended to contain greater halo and to be more lenient, less differentiating, less reliable, and less valid than those performed for developmental purposes. User acceptance as measured by recommendation for future use was more favorable under the developmental than the evaluative conditions. These results suggest that the quality of peer ratings and user acceptance are highly susceptible to the influence of rating contexts and that peer ratings are more useful for developmental than for evaluative purposes. Implications of these results for future peer-appraisal practices and research are discussed.

It has been repeatedly suggested that peer assessment as a source of performance appraisal has high reliability and validity (e.g., Bernardin & Beatty, 1984; Love, 1981; Mumford, 1983). A close examination of the relevant literature indicates that the stated advantages of peer assessment in terms of

We would like to express our appreciation to Pat Wright and Greg Dobbins for their comments on an earlier version of this article. We also are indebted to two anonymous reviewers for their suggestions regarding new analyses and other revisions. *Requests for reprints should be addressed to Jiing-Lih Farh, Department of Management, Louisiana State University, Baton Rouge, LA 70803-6312.*

Group & Organization Studies, Vol. 16 No. 4, December 1991 367-386

high reliability and validity are largely limited to peer nominations and peer rankings and do not generalize to peer ratings. For example, Kane and Lawler (1978) reviewed research on three methods of peer assessment: peer nominations, peer ratings, and peer rankings. They found that among the three methods, peer nomination appears to have the highest reliability and validity. They also found that although peer rating is the most useful among the three methods for feedback purposes, it produces the least valid, reliable, and unbiased measurements. Indeed, in the 14 studies that Kane and Lawler (1978) reviewed, the median values of reliability and validity for peer rating were .45 and .35, respectively. These discouraging results led them to conclude that only limited faith could be placed in the discriminability of peer ratings. A subsequent study by Love (1981) likewise confirmed the finding that peer ratings are less reliable and valid than peer nominations and peer rankings. To date, research has yet to examine factors that may contribute to the low reliability and validity of peer ratings.

Perhaps the low reliability and validity of peer ratings should not come as a surprise to researchers after all. As typically defined, peer rating is the process of having group members rate each other on a given set of performance or personal characteristics by using a specific set of rating scales (Kane & Lawler, 1978). As such, peer ratings are likely to exhibit virtually every form of rater bias that has been documented with supervisor and self-appraisals (e.g., Bernardin & Beatty, 1984; Thornton, 1980). However, there has been no research that has directly compared the validity and reliability of peer ratings under different rating conditions (Kane & Lawler, 1978).

Another issue that may raise a question about the usefulness of peer ratings is user acceptance. Although several studies have documented that peer ratings suffer from poor user acceptance (e.g., Cederblom & Lounsbury, 1980; Love, 1981), few studies have directly compared user acceptance in different rating contexts. A notable exception is McEvoy and Buller's (1987) examination of user acceptance under evaluative versus developmental conditions. Because this study, however, was conducted in a field setting without rigorous experimental controls, its results were susceptible to numerous threats to internal validity (Cook & Campbell, 1979).

The present study was designed to address the issues mentioned above in an experimental setting. More specifically, this study examined the effect of peer-rating purpose (evaluative vs. developmental) on (a) rating quality (defined as high reliability, high validity, and free of rating biases such as leniency, halo, and restriction of range) and (b) user acceptance.

RATING PURPOSE AND RATING QUALITY

LENIENCY

Recent performance appraisal research has shown that rating quality is highly dependent on the context in which an appraisal is conducted (e.g., Bernardin & Beatty, 1984; Farh & Werbel, 1985). One contextual factor that has received much research attention is the purpose or consequence of an appraisal. Research involving supervisor ratings shows that raters take the purpose of an appraisal into account when completing performance evaluations. For example, Longenecker, Sims, and Gioia (1987) reported that managers were typically more lenient in rating their subordinates when their evaluations would be used to determine compensation than for other purposes. As a second example, Williams, DeNisi, Blencoe, and Cafferty (1985) found that performance ratings for promotion or salary decisions were more lenient than ratings for training referral decisions. This research indicates that supervisory ratings conducted for feedback or developmental purposes are less prone to leniency bias than are appraisals conducted for administrative purposes, such as determining employee compensation and promotability.

Specific reasons that supervisors are typically more lenient in performance ratings used for administrative purposes are not difficult to construe. Kane (1980) observed that rating errors are often prompted by apprehensions naturally present when one person evaluates another. This apprehensiveness is at least partially a function of a rating's purpose and, hence, likely consequences. Other things being equal, the more severe the perceived consequences of a negative rating, the greater the incentive for the rater to be lenient. Thus leniency bias reflects an increased supervisor concern over the potential negative ramifications of an unfavorable appraisal (Fisher, 1979; Longenecker et al., 1987). Indeed, Bernardin (1980) has reported evidence of subordinate retaliation for harsh supervisor evaluations, citing a relationship between the ratings a supervisor gives a subordinate and the sub-ordinate's description of the supervisor's leadership behavior.

Concern over potential negative ramifications is likewise present in peer ratings, especially when the ratings are collected for evaluative purposes. In some environments (e.g., a unionized workplace), peers may be unwilling to evaluate each other critically because they may feel that appraisal is their manager's job and that they should protect their peers by not providing negative data about them (Mohrman, Resnick-West, & Lawler, 1989). Moreover, the concern for retaliation is more serious for peer ratings than for

supervisory ratings because the typical peer-rating procedure requires peers to evaluate each other. Research has shown that student subjects lowered subsequent ratings of their classmates by giving them more harsh ratings after learning that the classmates had rated them negatively (DeNisi, Randolph, & Blencoe, 1983; Koeck & Guthrie, 1975). It would thus seem that when peer ratings are conducted for feedback or research purposes, the perceived consequences are less severe and so is the concern over potential negative ramifications. The following hypothesis is thus suggested.

Hypothesis 1: Peer ratings performed for evaluative purposes will be more lenient than those performed for developmental purposes.

DISCRIMINABILITY

When confronted with the apprehension of conducting peer ratings for evaluative purposes, an easy alternative for raters would be to appraise their peers' achievements the same or nearly the same, regardless of actual performance. This is known as *uniformity bias.* Drawing on Kane's (1980) discussion of motivated errors, one would anticipate that this alternative would become more attractive as the perceived consequences of an appraisal become more severe.

Although no studies have investigated this reasoning, related research supports its plausibility. Using undergraduate students to rate written descriptions of supermarket checker performance, Zedeck and Cascio (1982) found a greater willingness to discriminate between ratees when appraisals were to be used for developmental purposes than for merit-raise (i.e., evaluative) purposes. This therefore suggests a second hypothesis:

Hypothesis 2: Peer ratings performed for evaluative purposes will have less discriminability (across ratees and within raters) than those performed for developmental purposes.

HALO ERROR

Halo error in the context of performance appraisal may be defined as a tendency of raters to allow a general impression to affect their ratings of individual dimensions, resulting in high interdimension correlations or low interdimension variance (Cooper, 1981). Researchers have identified several principal sources of halo: undersampling, engulfing, insufficient concreteness, insufficient rater motivation and knowledge, and cognitive distortion (Bernardin & Beatty, 1984; Cooper, 1981). Among these sources, insufficient

rater motivation is closely related to rating purpose. As discussed above, when peers are asked to rate each other for evaluative purposes, they may be reluctant to provide accurate information about each other. If so, we may expect that unmotivated raters expend insufficient effort to differentiate ratings across dimensions within ratees. Therefore, the following hypothesis is advanced.

Hypothesis 3: Peer ratings performed for evaluative purposes will have more halo error than those performed for developmental purposes.

INTERRATER RELIABILITY

Interrater reliability has a direct bearing on the validity and, thus, the usefulness of peer ratings. If such ratings are unreliable, their stated advantages (e.g., multiple raters and awareness of true performance) are nonexistent. Because previous research has shown that peer ratings tend to exhibit low reliability, establishing the reliability of peer ratings, especially under different purposes, is thus clearly important.

As stated earlier, peers are reluctant to provide accurate (especially negative) information about each other under evaluative purposes because of a concern about potential negative ramifications of such actions. This unwillingness to evaluate ratees on the basis of actual performance tends to introduce errors into the rating process and thus make ratings unreliable. In contrast, when peer ratings are collected for developmental purposes, raters are less concerned about rating consequences and more willing to rate each other on the basis of their actual performance. Under such conditions, we should expect higher interrater agreement within ratees and more reliable ratings. Accordingly, a fourth hypothesis is suggested:

Hypothesis 4: Peer ratings performed for evaluative purposes will have lower interrater reliability than those performed for developmental purposes.

USER ACCEPTANCE

The question of user acceptance of peer ratings has been investigated in several studies (e.g., Cederblom & Lounsbury, 1980; Fedor & Bettenhausen, 1989; Love, 1981; McEvoy & Buller, 1987). Of these studies, only two have investigated user acceptance of peer ratings under a purpose manipulation. McEvoy and Buller (1987), who conducted a field study with hourly employees of a food-processing plant, found user acceptance of peer ratings to be

more favorable when ratings were used for developmental (i.e., counseling) rather than evaluative (e.g., wage) purposes. Because rating purposes were not manipulated in McEvoy and Buller's study (in fact, both types of peer rating were conducted concurrently in the plant at the time of the user survey), their results are subject to a variety of alternative explanations (Cook & Campbell, 1979). A second study, conducted by Fedor and Bettenhausen (1989), has likewise investigated the impact of appraisal purposes on the acceptance of peer ratings and found contradictory results. The results of this study, however, are not comparable to either McEvoy and Buller's (1987) or to our investigation for several reasons. First, its subjects (i.e., undergraduate students) were not "peer" in a strict sense, because they did not work together on a common undertaking. Second, the subjects did not mutually rate each other. Third, no future interactions among subjects were expected, because the peer rating was conducted at the end of a semester. Thus drawing on McEvoy and Buller (1987) alone, as well as prior theory (DeNisi & Mitchell, 1978), a final hypothesis is offered:

Hypothesis 5: User acceptance of peer ratings performed for developmental purposes will be more favorable than acceptance of those performed for evaluative purposes.

STUDY OVERVIEW

Sixty-seven undergraduate students enrolled in two sections of an organizational behavior course participated in this study. Both sections were taught on the same days by the same instructor using the same teaching method. As part of the course's requirements, students were assigned to groups of 6 or 7 members in order to complete three team projects. A peer-evaluation procedure was introduced and experimentally manipulated across sections. Because it was impossible to assign students randomly to sections (treatment conditions), one section was assigned to an evaluative (i.e., grading) condition, the other to a developmental (i.e., feedback) condition. The developmental (i.e., feedback) treatment condition had 32 subjects (16 female, 16 male); the evaluative (i.e., grading treatment condition) had 35 subjects (7 female, 28 male).

The strategy of using student subjects in a classroom setting, of course, raises the question of external validity. This issue can be addressed in three ways. First, the primary goal of this study is to demonstrate that peer-rating quality and user acceptance are sensitive to an evaluation's purpose. Given

this goal, it is unnecessary to show that such sensitivity occurs with a specified frequency in field settings; one needs only to show that it is possible for such an effect to take place (Mook, 1983). Second, as Locke (1986) and others have pointed out, the generalizability of a laboratory study to field settings hinges on its similarity to the latter setting in terms of essential attributes. In designing the present study, we took pains in ensuring that reasonable similarity on essential attributes existed. For example, the assigned group projects required a considerable amount of coordinated team effort; to simulate the ongoing nature of existing groups, peer ratings were conducted following a *second* group project to ensure that a fair amount of group interaction had already occurred, and subjects were fully aware that future interactions would take place in the third project; the treatment manipulation (explained later) had a meaning to students that was similar to the meaning it would have to employees. Third, the applicability of laboratory findings to problems of real organizations may be underestimated. Researchers have systematically reviewed the literature in industrial and organizational psychology (including performance appraisal) and compared findings found in laboratory studies with those found in field studies (Locke, 1986). The overall conclusion is that the direction of effect found in field and laboratory studies is either highly similar or virtually identical (Locke, 1986). The preceding points, taken together, suggest that the results obtained in this study are likely generalizable to field settings.

METHOD

PROCEDURE

During the course's first class period, students were informed that (a) they would be assigned to groups to complete *three* team projects, (b) the projects were to be an integral part of the course, (c) the projects would make up a significant proportion of the course grade, and (d) the team-project grade would be their individual grades. Students were then asked to complete a brief biographical form, requesting information about their age, gender, grade point average (GPA), major, class standing (sophomore, junior, etc.), and employment status (full-time, part-time, or unemployed). This information was used by the instructor to divide the students into 6- or 7-member groups that were demographically balanced.

The first team project was assigned in the second week of class. It required students to use Mintzberg's managerial framework (Mintzberg, 1973) to

interview three working managers and prepare a written report. Peer evaluations were not mentioned at this time. All groups were given 2 weeks to complete the project. Except for their first meeting, which took place during class time, all other group meetings were conducted outside class. When the project was completed, team members were asked to complete a 20-item questionnaire designed to assess intragroup processes. This questionnaire was designed to serve as a pretest to check for an equivalence of group processes across the two treatment conditions.

The second team project was assigned 2 weeks after the first project was completed. In announcing the project, students were informed that peer evaluations, in which they would be asked to evaluate each member of their group (including themselves), would be conducted at the project's completion. The instructor deliberately kept the announcement brief, not explaining the peer evaluation's purpose. The second project required the teams to analyze a complex organizational behavior case and prepare a detailed written analysis. On the project's due date, students were asked to complete a peer-evaluation form under one of two instructional sets. All students in both the developmental and evaluative conditions completed the forms.

Except for the instructional sets that induced the purpose manipulation, the peer evaluation forms were identical for both treatment conditions. The forms requested students to rate all team members on 13 items (described below). To simplify the rating procedure, each form had team members' names printed as column headings. One week after the peer evaluations, students were fed back their true peer-rating results in writing by the instructor. For each rating item, the students were provided with their own peer ratings along with the average ratings for their team. Immediately after the feedback, students were asked to complete a short questionnaire assessing their perceptions of the peer-evaluation procedure.

EXPERIMENTAL MANIPULATION

The experimental manipulation was performed by means of instructional sets. For subjects in the developmental (i.e., feedback) condition, instructions were as follows:

> The purpose of this evaluation is to provide each group member with information about his or her contribution to the second team project as perceived by other group members. Your evaluation is confidential and will not be seen by other group members. Only the summarized results of the evaluation will be returned to group members to help them function more effectively in future group settings. The results will not affect your grade. Your individual grades will be determined solely by the group grade on the project.

For subjects in the evaluative (i.e., grading) condition, the instructions were as follows:

> Ideally, each group member contributes equally to a group project. In reality, however, the contributions of individual group members may vary substantially. The purpose of this peer evaluation is to assess the relative contribution of each group member to the second team project. Your responses on this questionnaire will be kept confidential and will be seen only by the class instructor. The results will be used to adjust individual group members' grades on the project. Individuals who are consistently rated poorly by their peers will receive fewer points on the project than those who are consistently favorably evaluated by their peers.

It should be noted that because the experimental purpose manipulation took place after the second team project was already completed, it could not have altered team members' behavior during the project. Discounting for diffusion or imitation of treatments and assuming that the teams in the two treatment conditions were equivalent prior to the peer evaluations, any differences in the quality of the peer ratings could thus be attributed to the effect of the experimental purpose manipulation rather than to the differing group processes possibly caused by the manipulation.

Shortly after the experiment was completed, a third group project was assigned to the class. All subjects were informed by the instructor that a peer-rating procedure would follow the project for grading purposes. A simplified peer-rating form was used by the instructor, and the results were unavailable for this study.

DEPENDENT VARIABLES

Pretest variable. A 20-item pretest questionnaire, adapted from Staw (1975), assessed team members' perceptions of various types of group processes including group cohesiveness, group communication, perceived self- and teammate motivation, and the openness of the group to change and different ideas. Responses to each item were coded from 1 to 11, with greater values reflecting higher scores.

Rating dimensions. A peer-rating form was modeled after a classification system of group member roles (behaviors) developed by Benne and Sheats (1948). The rating form contained 13 items tapping three dimensions of group member roles. Each item was measured on a 7-point scale (1 = *strongly disagree,* 7 = *strongly agree*). The first dimension, *task performance,* represented member roles related to the accomplishment of the group task. It was

measured by five items: offering valuable ideas or suggestions to the project, completing their fair share of work, actively participating in group activities, coordinating group activities, and attending every group meeting. Coefficient alpha for this dimension in the present study was .95.

The second dimension, *group maintenance,* focused on member roles geared toward the functioning of the assigned teams. It was measured by five items: encouraging cohesiveness and warmth between group members, having an overall positive attitude toward the group, encouraging participation by all members, helping reduce conflict and tension, and helping the group set goals and standards. Coefficient alpha for this dimension in the present study was .95.

The third dimension, *individual orientation,* refers to individual members' behavior directed toward the satisfaction of their own needs. According to Benne and Sheats (1948), this dimension is irrelevant to both task performance and the functioning of teams as groups. The dimension included three items: never calling attention to self by boasting or acting superior, being agreeable and willing to listen to other's suggestions, and never interrupting others. Coefficient alpha for this dimension in the present study was .90.

Peer rating. Peer rating for a ratee was defined as the average of the ratings given to that ratee by members of his or her group on each of the three rating dimensions.

Discriminability. Consistent with Bernardin and Beatty (1984), discriminability is operationalized as the standard deviations of ratings across ratees within raters. A high score thus indicates that a rater assigned different ratings to different team members, whereas a low score indicates that a rater gave the same or similar ratings to all or most team members. This index was calculated separately for each rating dimension. The intraclass correlation reliability for this index was .90 for the developmental purpose condition and .63 for the evaluative purpose condition.

Interrater reliability. The interrater reliability of the mean peer ratings for the two treatment conditions was assessed through the intraclass correlation coefficient (Shrout & Fleiss, 1979).

Halo error. Halo was calculated within raters. For each rater, we first calculated the standard deviation across the three dimensions for each ratee. We then calculated the mean of the preceding measure across all ratees for

that rater and defined it as an indicator of halo. Therefore, halo in this study was operationalized as low standard deviations across rating dimensions within ratees; higher scores (large standard deviations) indicate less halo and vice versa (Bernardin & Beatty, 1984).

User acceptance. Consistent with Cederblom and Lounsbury (1980) and McEvoy and Buller (1987), user acceptance of the peer-evaluation procedure was assessed by team members' responses to the item "I strongly recommend that the procedure be used in future classes" on a 7-point scale (1 = *low* and 7 = *high*). In addition, team members' perceptions of the value of the peer-evaluation procedure were measured by the following items using similar scales: (a) perceived positive effect of peer ratings on team morale, (b) usefulness of peer ratings as feedback, (c) perceived accuracy of peer ratings received, and (d) extent to which peer ratings were affected by friendship bias (reverse scoring).

RESULTS

EQUIVALENCE OF TREATMENT CONDITIONS

Because the treatment conditions were confounded by class section, we first examined the equivalence of the groups on the study's pretest measures. Given that the pretest measures were correlated, multivariate analysis of variance (MANOVA) was used to test the overall difference between the two groups. The result indicated that the two groups were not significantly different. Independent t tests also found no significant difference between the groups on each pretest measure (all $ps < .05$). Moreover, the subjects were similar across treatment conditions on GPA, age, major, and employment status. Although the percentage of female students was significantly different between conditions (50% in the developmental condition, 20% in the evaluative condition), gender was not considered a confound, because preliminary analyses indicated no main effects for Gender or any Gender × Condition interactions.

Subjects in the two treatment conditions were thus homogeneous on pretest measures and demographic variables except gender. Although it is impossible to rule out all threats related to selection without the benefit of randomization, the available evidence points to the relative equivalence of the subjects in the two treatment conditions prior to the experimental manipulation.

MANIPULATION CHECK

Previous research (e.g., Farh & Werbel, 1985) has shown that students' self-ratings of performance are more lenient when an appraisal is being conducted for grading as contrasted with research purposes. To check the effectiveness of the experimental manipulation, MANOVA was conducted to compare the self-ratings for the two treatment conditions. Results of this analysis indicated a significant difference in the means of self-ratings between the two conditions ($F[3, 57] = 4.73, p < .01$). Results of subsequent t tests showed that self-ratings obtained under the evaluative condition were more lenient than those obtained under the developmental condition for each of the three dimensions (all ps $< .05$). Thus the experimental manipulation was effective.

LENIENCY BIAS

To examine Hypothesis 1, MANOVA was performed on the peer ratings by treatment. Results of this analysis showed that peer ratings were significantly different between the two conditions ($F[3, 63] = 11.33, p < .01$). A series of t tests was then performed, with the results shown in Table 1, indicating that subjects in the evaluative condition received more lenient peer ratings on all three rating dimensions than did subjects in the developmental condition. Thus Hypothesis 1 is strongly supported.

DISCRIMINABILITY

To test Hypothesis 2, MANOVA was performed on the discriminability indices by treatment. Results indicated that the treatment had a significant effect on discriminability indices ($F(3, 63) = 4.66, p < .01$). Results of t tests (also presented in Table 1) show that peer ratings obtained under the developmental condition are more discriminating for all three rating dimensions than those obtained under the evaluative condition. Thus Hypothesis 2 is strongly supported.

HALO ERROR

To test Hypothesis 3, that halo error is greater in the evaluative condition than in the developmental condition, the mean interdimension standard deviations were compared through a t test. The result (also reported in Table 1) shows that ratings collected in the developmental condition had significantly

TABLE 1
Means, Standard Deviations, and Results of Significance Tests of Peer Ratings, Discriminability Index, and Halo by Appraisal Purpose

Measures	*Developmental* (n = 32)		*Evaluative* (n = 35)		*Mean Difference Test* (t values)
	M	SD	M	SD	
			Peer ratings		
Task performance	5.19	1.09	5.91	0.95	2.88**
Group maintenance	4.89	1.11	5.97	0.70	4.69**
Individual orientation	5.35	0.92	6.27	0.53	4.92**
			Discriminability index		
Task performance	1.11	0.80	0.68	0.78	2.22*
Group maintenance	0.84	0.83	0.41	0.58	2.45**
Individual orientation	0.59	0.73	0.10	0.24	3.62**
			Halo		
Across the three rating dimensions	0.60	0.35	0.39	0.39	2.25*

*$p < .05$; **$p < .01$.

greater interdimension standard deviations than those collected in the evaluative condition (Ms = 0.60 vs. 0.39, $p < .01$). Thus Hypothesis 3 is strongly supported.

INTERRATER RELIABILITY

Table 2 shows the intraclass correlation coefficients and 95% confidence intervals for each rating dimension under the two treatment conditions. Confidence intervals were calculated in accordance with the method suggested by Shrout and Fleiss (1979). Whereas all three intraclass coefficients were highly significant for the developmental condition, only one of the three (task performance) reached significance for the evaluative condition. The average intraclass coefficient for the developmental condition was .73, in contrast with .36 for the evaluative condition. Only one of the three confidence intervals for the developmental condition, however, lies completely outside the confidence interval for the evaluative condition. These results provide qualified support for Hypothesis 3.

TABLE 2
Intraclass Correlations Coefficients for Peer Ratings by Appraisal Purpose

Dimension	*Reliability*	*95% Confidence Interval*	
		Lower Bound	*Upper Bound*
Task performance			
Developmental	.771**	.554	.844
Evaluative	.606**	.483	.721
Group maintenance			
Developmental	.797**	.658	.880
Evaluative	.316	−.155	.596
Individual orientation			
Developmental	.615**	.350	.772
Evaluative	.149	−.437	.497

**$p < .01$.

USER ACCEPTANCE

To test Hypothesis 4, a *t* test was performed on user recommendations for the future use of the peer-evaluation procedure by treatment condition. Results (shown in Table 3) indicate that subjects in the developmental condition had a more favorable recommendation for the future use of peer ratings than did subjects in the evaluative condition. Thus Hypothesis 4 was supported.

Table 3 also contains the results concerning team members' perceptions of the value and accuracy of the peer-rating procedure, indicating that there are no significant differences between the two conditions on perceived effect on morale, usefulness of ratings as feedback, perceived accuracy, and friendship bias.

CONVERGENT VALIDITY

Because both self-ratings and peer ratings were available in this study, the convergent validities of the self-ratings versus peer ratings were examined. Table 4 presents the correlations between self-ratings and peer ratings by treatment conditions. The results show that self-ratings and peer ratings are significantly correlated for each rating dimension in the developmental condition, but none of the correlations is significant in the evaluative condition. However, the difference in the size of the correlations between the two

TABLE 3
Means, Standard Deviations, and Results of Significance Tests of User Reactions to Peer Ratings by Appraisal Purpose

	Developmental (n = *29*)		*Evaluative* (n = *35*)		
Dimension	M	SD	M	SD	*Mean Difference Test* (t *values*)
Recommendation for future use	5.24	1.50	4.43	1.93	1.89*
Positive effect on morale	5.14	0.99	4.74	1.24	1.43
Usefulness of ratings as feedback	4.93	1.69	4.94	1.68	0.03
Perceived accuracy	4.62	1.72	5.11	1.57	1.19
Free of friendship bias	4.69	1.63	5.09	1.58	0.98

*$p < .05$, one-tailed test.

TABLE 4
Convergent Validity Between Self-Ratings and Peer Ratings by Appraisal Purpose

	Developmental (n = *30*)	*Evaluative* (n = *31*)	Z *values*
Task performance	.65*	.24	2.30*
Group maintenance	.49*	.11	1.85
Individual orientation	.46*	.14	1.55

*$p < .05$.

treatment conditions reached significance ($p < .05$) for only one of the three rating dimensions, arguably because of small sample size.

DISCUSSION

The findings of this study indicate that the quality of peer ratings is very sensitive to the contexts in which the ratings are obtained. When peer ratings were conducted for evaluative (i.e., grading) purposes, peer raters tended to rate each other more leniently and to assign similar ratings across ratees as

well as across dimensions, regardless of actual performance. These rating biases resulted in a severe restriction in range, lower interrater reliability, and low convergent validity. In contrast, when peer ratings were collected for developmental (i.e., feedback) purposes, raters were less lenient in their evaluations and were more willing to differentiate among their peers as well as across rating dimensions, resulting in higher interrater reliability and convergent validity.

It is worth noting that although peer ratings are susceptible to the influence of contextual factors, other peer-assessment methods (especially peer nomination) are more immune to such influence. Research by Hollander (1957) showed that peer nomination had respectable reliability under different administration conditions. These results may have occurred because peer nomination includes a forced comparison procedure that effectively eliminates leniency and uniformity biases. Despite the favorable psychometric properties of peer nomination, its value for providing developmental feedback is quite limited because it does not evaluate ratee performance against absolute standards (Kane & Lawler, 1978).

Regarding user acceptance, it was found that team members in the developmental condition recommended their procedure more strongly than those in the evaluative condition. This finding reconfirmed and extended the McEvoy and Buller (1987) study, in that it demonstrated, with rigorous experimental control and with a different sample (college students as compared with hourly workers), that user acceptance is indeed lower in an evaluative condition than in a developmental condition.

Concerning peer members' perceptions of the value and accuracy of the peer-rating procedure, appraisal purpose had no significant effect on any of the measures. It is worth noting that although team members in the developmental condition gave a more favorable recommendation for the peer-rating procedure, they tended to perceive the rating results as less accurate than those in the evaluative condition did. These seemingly contradictory results, however, can be explained by self-serving biases (Miller & Ross, 1975). In this study, all raters completed the user reactions survey immediately after they were shown their peer-appraisal results. Because raters in the evaluative condition received a more lenient rating than those in the developmental condition, their responses on the survey may have been biased by the differing rating results. To verify this possibility, we examined the correlations between peer ratings received, perceived accuracy, and recommendation for future use for the entire sample. The results showed that perceived accuracy was positively correlated with peer ratings received ($rs = .28$ to $.41$ for the three rating dimensions, all $ps < .05$), but recommendation for future

use was uncorrelated with peer ratings received (rs = –.08 to –.03). When the data were analyzed within each treatment condition, a similar pattern of results also emerged. These results suggest that although perceived accuracy was tainted by self-serving biases, recommendation for future use was relatively free of such biases and thus a better measure of user reactions.

Two study limitations should be mentioned. First, one could still question the present study's external validity because of its student sample and classroom setting, notwithstanding our effort to address this issue through our research design. One may point to the fact that student groups formed in classroom settings for completing team projects are invariably temporary in nature, and their frequency of interaction is also limited, whereas in existing work groups, employees typically have more information about each other, and their concern over the long-term effects of peer appraisals may be more serious. If this is indeed true, the presence of significant effects in our study thus argues even more strongly for the efficacy of the investigated relationship in field settings. We, however, realize that the issue of generalizability is ultimately an empirical question and thus call for further systematic investigation of this issue in field settings.

The second limitation of this study has to do with the peer-rating instrument employed. It consisted of 13 items, each referring to member behavior in a group setting. The items were each anchored by all-purpose terms (strongly agree, strongly disagree, etc.). The rating form thus resembled traditional graphic rating scales, which are known to be vulnerable to rating biases. Moreover, in completing the peer-rating instrument, raters were not instructed to provide any justification for their assessments. Thus the impact of rating purposes (consequences) on rating results remains to be assessed where performance dimensions are well defined, sophisticated rating instruments are employed, and justification for ratings given to others are required.

If the preceding limitations are kept in mind, this study offers several important implications for the future use of peer ratings in organization settings. First, the study suggests that in team-oriented work settings, organizations should consider peer ratings as a valuable source of information for developmental purposes. Under these settings peers often have very significant and sometimes unique information about the behavior of their fellow workers (Kane & Lawler, 1979). Our study demonstrates that when peer ratings are collected and used for developmental purposes, peers are receptive to such a procedure and, more important, the rating results are reliable and valid.

Second, this study suggests that organizations should be cautious about using peer ratings for evaluative purposes (e.g., salary and promotion deci-

sions). Not only does the procedure have lower user acceptance, but the resultant ratings also suffer from leniency, restriction of range, low reliability, and low validity. One way to address this issue is to use peer nominations in place of peer ratings in such situations. As noted earlier, peer nominations essentially force raters to differentiate among each other and thus avoid leniency and uniformity biases. Although peer nominations have been criticized for lack of feedback value, providing feedback is not the central concern of evaluative appraisal anyway. In fact, the inherent conflict between evaluative and developmental appraisals has prompted researchers to suggest that two separate appraisals should be used, one for evaluative and one for developmental purposes (e.g., Meyer, Kay, & French, 1965).

Using peer nomination in place of peer ratings for evaluative purposes does not address the issue of low user acceptance. It has been suggested that low user acceptance may result from (a) a belief that evaluative appraisal is the manager's job and peers should protect each other, (b) the fear of retaliation from those who received less-than-expected ratings, (c) a perception that peer ratings are invalid and represent merely a popularity contest, and (d) a concern about the long-term negative ramifications of the appraisal on group morale (e.g., Mohrman et al., 1989). Future research should move beyond documenting these concerns and begin to inquire how these concerns could be effectively addressed through various intervention strategies, such as rater training, fostering rater trust, communication efforts, built-in procedural safeguards, and rating format improvements.

In summary, the present study demonstrates that peer-rating purpose has a significant impact on user acceptance and the quality of the resulting ratings. Peer ratings collected under developmental purposes not only have higher user acceptance but also tend to be more reliable, more valid, and less susceptible to rating biases than those collected under evaluative purposes. These findings suggest that peer ratings in an appropriate context can be a useful means of appraisal. Given the societal trend toward team-oriented, high-involvement work arrangements, peer appraisal is likely to play an increasingly important role in the future. More research is needed to delineate the various conditions that may influence the effective use of peer-performance assessments.

REFERENCES

Benne, K., & Sheats, P. (1948). Functional roles of group members. *Journal of Social Issues, 2,* 42-47.

Bernardin, H. J. (1980). The effect of reciprocal leniency on the relationship between consideration scores from the LBDQ and performance ratings. *Proceedings of the 40th Annual Meeting of the Academy of Management,* 131-136 (Detroit).

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work.* Belmont, CA: Wadsworth.

Cederblom, D., & Lounsbury, J. W. (1980). An investigation of user acceptance of peer evaluations. *Personnel Psychology, 33,* 567-579.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Skokie, IL: Rand McNally.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90,* 218-244.

DeNisi, A. S., & Mitchell, J. L. (1978). An analysis of peer ratings as predictors and criterion measures and a proposed new application. *Academy of Management Review, 3,* 369-374.

DeNisi, A. S., Randolph, W. A., & Blencoe, A. G. (1983). Potential problems with peer ratings. *Academy of Management Journal, 26,* 457-464.

Farh, J. L., & Werbel, J. D. (1985). The effects of purpose of the appraisal and expectation of validation on the quality of self-appraisals. *Journal of Applied Psychology, 71,* 527-529.

Fedor, D. B., & Bettenhausen, K. L. (1989). The impact of purpose, participant preconceptions, and rating level on acceptance of peer evaluations. *Group & Organization Studies, 14,* 182-197.

Fisher, C. D. (1979). Transmission of positive and negative feedback to subordinates: A laboratory investigation. *Journal of Applied Psychology, 64,* 533-540.

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology, 41,* 85-90.

Kane, J. (1980). *Alternative approaches to the control of systematic error in performance appraisals.* Paper presented at the First Annual Scientist-Practitioner Conference in Industrial/ Organizational Psychology, Norfolk, VA.

Kane, J. S., & Lawler, E. E. (1978). Methods of peer assessment. *Psychological Bulletin, 85,* 555-586.

Kane, J. S., & Lawler, E. E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. Staw (Ed.), *Research in organizational behavior* (Vol. 1). Greenwich, CT: JAI Press.

Koeck, R., & Guthrie, G. M. (1975). Reciprocity in impression formation. *Journal of Social Psychology, 54,* 31-41.

Locke, E. A. (1986). *Generalizing from laboratory to field settings.* Lexington, MA: Lexington Books.

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of performance appraisal. *Academy of Management Executive, 1,* 183-193.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology, 66,* 451-457.

McEvoy, G. W., & Buller, P. F. (1987). User acceptance of peer appraisals in an industrial setting. *Personnel Psychology, 40,* 785-797.

Meyer, H. H., Kay, E., & French, J. (1965). Split roles in performance appraisal. *Harvard Business Review, 43,* 123-129.

Miller, D., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82,* 213-225.

Mintzberg, H. (1973). *The nature of managerial work.* New York: Harper & Row.

Mohrman, A. M., Resnick-West, S. M., & Lawler, E. E. (1989). *Designing performance appraisal systems: Aligning appraisals and organizational realities.* San Francisco: Jossey-Bass.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379-387.
Mumford, M. D. (1983). Social comparison theory and the evaluation of peer evaluations: A review and some applied implications. *Personnel Psychology, 36,* 867-881.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.
Staw, B. M. (1975). Attribution of the "causes" of performance: A general alternative interpretation of cross-sectional research on organizations. *Organizational Behavior and Human Performance, 13,* 414-432.
Thornton, G. C. (1980). Psychometric properties of self-appraisals of job performance. *Personnel Psychology, 33,* 263-271.
Williams, K. J., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The role of appraisal purpose in information acquisition and utilization. *Organizational Behavior and Human Decision Processes, 35,* 314-339.
Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. *Journal of Applied Psychology, 67,* 752-758.

Jiing-Lih Farh is an associate professor of management in the Department of Management at Louisiana State University. He received his Ph.D. degree in organizational behavior and personnel from Indiana University at Bloomington in 1983. His current research interests include performance appraisal, task design, goal setting, reward systems, and comparative management.

Albert A. Cannella, Jr., is an assistant professor of management at Texas A&M University. He received his Ph.D. in strategic management from Columbia University. His main areas of interest are strategic leadership, individual influences on organizational outcomes, and strategy implementation.

Arthur G. Bedeian is the Ralph and Kacoo Olinde Distinguished Professor of Management at Louisiana State University, where he specializes in organizational design and the study of behavior in organizations.