Journal of Applied Psychology
2002, Vol. 87, No. 6, 1032–1041

0021-9010/02/$5.00 DOI: 10.1037//0021-9010.87.6.1032

Reducing the Performance-Cue Bias in Work Behavior Ratings: Can Groups Help?

Richard F. Martell and Keith N. Leavitt
Montana State University

The authors examined whether the performance-cue bias can be reduced by relying on groups as raters. Study participants ($N = 333$) were provided with feedback regarding the performance of a workgroup and, after observing the group, assigned to an individual or group rater condition to complete a behavioral rating instrument. Results revealed that when provided with positive (vs. negative) feedback, individuals attributed more effective and fewer ineffective behaviors to the workgroup; however, group ratings were unaffected by the feedback. In addition, feedback biased the decision criteria and false alarm rates of individuals but not of groups. Discussion of when groups may attenuate versus amplify bias in performance appraisal judgments emphasizes 2 key elements—bias magnitude and task perception.

An abundance of research demonstrates that individuals' beliefs and expectations can introduce bias in judgments of work performance (Feldman, 1994; Ilgen, Barnes-Farrell, & McKellin, 1993; Motowidlo, 1986). One well-researched example is the *performance-cue bias* that was first investigated by Staw (1975). In this study, Staw led business students to believe that their workgroup had performed "very well" or "very poorly" on a business simulation exercise. Group process ratings (e.g., cohesiveness, task conflict) and self-evaluations (e.g., ability, motivation) collected at the conclusion of the exercise revealed that the bogus performance feedback produced biased evaluations. Specifically, members of high performing groups rated their performance and the quality of the group's interaction more favorably than did members of low performing groups. These findings raised the possibility of causal reversals in which knowledge of group performance serves as a cue, causing individuals to ascribe cue-consistent attributes to the target of their assessment. Indeed, Staw concluded that theory development in organization behavior could be seriously compromised because "significant correlations between performance and self-report data may only be reflecting the respondents 'theories' of organizational performance rather than actual events" (p. 417).

More recent research on the performance-cue bias has examined the effects of observers' expectations on behavioral rating accuracy. As discussed by Murphy (1991) and Lord (1985a), behavioral rating accuracy is an important criterion measure of work performance; it yields a measure of the accuracy with which raters credit individuals with critical work-related behaviors and is used for employee evaluation, training, feedback, and assessment center judgments. This research has found that observers led to believe that a group performed well (vs. poorly) attributed more effective and fewer ineffective behaviors to the group (e.g., Martell & Guzzo, 1991; Martell & Willis, 1993). Moreover, performance-cue biases have been observed in the judgments of experienced managers and in field studies (Fichman, 1991; Gardner, Wright, & Gerhart, 2000; Golden, 1992). The importance of the performance-cue bias was recently affirmed in a study designed to resolve a long-standing methodological issue regarding whether the bias is an artifact, triggered by the desire to render behavioral ratings consistent with experimenter-provided feedback. In contrast to previous studies, Martell, Guzzo, and Willis (1995) provided no performance-based feedback and instead allowed observers to develop their own self-generated evaluations of the group's performance. Results indicated that self-generated evaluations biased observers' behavioral ratings in precisely the same manner as experimenter-provided feedback, demonstrating that the bias is genuine and not an artifact. Despite the implications of the performance-cue bias for theory development and work performance appraisal, the design of interventions aimed at reducing the effects of performance-based expectations on raters' judgments has received scant attention. The sole exception is a recent study by Baltes and Parker (2000), which found that a halo error training program and a structured recall memory intervention eliminated the performance-cue bias. This finding is encouraging and suggests that the effects of performance cues on work behavior ratings can be circumvented.

To date, all previous research has relied on individuals as raters—little is known of the performance-cue bias in group-based judgments. Given the increasing frequency with which groups are employed as decision makers in organizational settings and the finding that groups are less sensitive to some types of judgmental biases, attention to group-based judgments is warranted. The goal of the research reported in this article was to examine whether the performance-cue bias in behavioral ratings can be reduced by relying on groups as opposed to individuals as raters.

Richard F. Martell and Keith N. Leavitt, Department of Psychology, Montana State University.

Keith N. Leavitt conducted this study under the direction of Richard F. Martell as part of the requirements for his master's degree in psychology at Montana State University. We gratefully acknowledge Wesley Lynch and Charles Pierce, who served as committee members, and James Robison-Cox for their helpful advice. Eric Szemes's untiring assistance in conducting this research is much appreciated.

Correspondence concerning this article should be addressed to Richard F. Martell, Department of Psychology, Montana State University, Bozeman, Montana 59717-3440. E-mail: martell@montana.edu

Groups as Raters

Organizations are increasingly relying on groups to do the work typically reserved for individuals (see Borman, Hanson, & Hedge, 1997; Dumaine, 1994; Guzzo & Shea, 1992; Ilgen, 1999; Katzenbach & Smith, 1993; Shea & Guzzo, 1987), including, among other things, human resource management functions. Reviewing the potential for groups to perform various human resource management functions, Shea and Guzzo (1987) concluded that "it is time for the field of personnel and human resources management to discover formal groups as resources in their own right" (p. 323) and pointed to performance appraisal as one such function. The notion that multiple raters may convey an advantage over individuals in the performance appraisal process can be traced to a theory of rating proposed decades ago (Wherry & Bartlett, 1982). The primary aims of the theory were to examine the effects of factors other than actual ratee performance on performance appraisals and to identify methods for eliminating rater-induced bias. The theory devotes particular attention to the impact of plural raters on performance appraisal judgments (Wherry & Bartlett, 1982). A key proposition is that the use of multiple raters can reduce bias and error in performance appraisals (Wherry & Bartlett, 1982; see Theorems 25–27). This prediction rests largely on the belief that multiple raters can bring nonredundant information to the judgment task. More recent thinking suggests that multiple raters, if assembled into a group, may correct individual member errors and biases and, perhaps, bolster individual rater motivation. Consistent with this proposition, group discussion among multiple assessors is a critical component of assessment centers and is thought to be helpful in heightening rater accuracy and decreasing bias (Howard, 1997; Lowry, 1997; Spychalski, Quinones, Gaughler, & Pohley, 1997; Zedeck, 1986; but see Sackett & Wilson, 1982).

Research on social judgment biases, including the correspondence bias (Wright & Wells, 1985), the theory-perseverance effect (Wright & Christie, 1989), the hindsight bias (Stahlberg, Eller, Maas, & Frey, 1995, Experiment 2) and the consensus underutilization effect (Wright, Lüüs, & Christie, 1990), provide further reason to believe that groups may render less biased assessments than individuals. In each of these studies, individual-level bias was attenuated (and often eliminated) by groups. Consider, for example, the correspondence bias whereby perceivers tend to ignore compelling situational explanations when determining the cause of a person's behavior and instead wrongly assume that the behavior is a direct manifestation of a person's underlying dispositional state. The correspondence bias reflects a type of cognitive shortcut in which perceivers too readily succumb to an implicit theory that behavior is caused by the underlying trait or disposition to which it corresponds (Gilbert & Malone, 1995); it reveals heuristic thinking and is a consequence of faulty information processing. However, as with many social judgment biases, the magnitude of the correspondence bias is not large; not all individuals necessarily succumb to the bias. This suggests that in a group context, it is likely that one or more individuals are unaffected by the correspondence bias and thus are in a position to impose their (accurate) beliefs on the group (Hinsz, Tindale & Vollrath, 1997; Wright & Wells, 1985). In addition, given the public nature of group deliberation in which members are exposed to contrary beliefs and alternative positions, groups may more often abandon heuristic thinking and engage in a deeper, more systematic processing of information than would individuals, especially if group members believe there is a necessarily correct answer to the task at hand (Stasser & Stewart, 1992). These features of group discussion provide reason to believe that reliance on implicit theories and the biases they produce in work performance ratings may be less evident in groups than individuals.

Group Memory

To better understand how groups may attenuate the performance-cue bias, it is necessary to pinpoint why positive (vs. negative) expectations cause observers to attribute more effective and fewer ineffective behaviors to the group. This has been done by examining the two components that fully capture performance on a behavioral rating instrument (Snodgrass & Corwin, 1988)—memory strength and decision criteria. Memory strength serves as an index of the amount of information stored in memory, whereas decision criteria provides a measure of an individual's decision criterion in force when deciding whether a behavior has been observed. The latter component is a type of response bias that indicates the extent to which an individual tends to be too liberal or too conservative in setting a decision threshold. Research has shown that memory strength plays no mediating role in the performance-cue bias. Rather, performance expectations trigger a response bias in which a too-liberal (i.e., a bias to say a behavior did occur) decision criterion is adopted when rating expectancy-consistent behaviors, and a too conservative (i.e., a bias to say a behavior did not occur) decision criterion is adopted when rating expectancy-inconsistent behaviors (Baltes & Parker, 2000; Martell & Guzzo, 1991; Martell & Willis, 1993). This causes individuals to endorse behaviors, whether actually observed or not, because they are consistent with implicit theories of how effective versus ineffective work groups behave. Thus, whereas performance expectations have been found to introduce bias in attentional, encoding, and storage processes (e.g., Ilgen et al., 1993), the performance-cue bias is a type of response bias that occurs at time of rating.

If groups are to help, they must lessen the response bias that afflicts individuals and resist reporting behaviors that were not observed. The group memory literature, much of which points to the superiority of groups over individuals (Clark & Stephenson, 1990; Hinsz, 1990; Hinsz, Tindale, & Vollrath, 1997; Vollrath, Sheppard, Hinsz, & Davis, 1989), has suggested that groups may be well-suited for this. For example, Hinsz (1990) applied signal detection analysis to identify why group recognition memory of the details of a videotaped job interview was superior to that of individuals. Results revealed that groups benefited from a pooling of individual member's memories, and thus, the memory strength of groups was greater than that of individuals. Moreover, of greater relevance for the performance-cue bias, the reporting of unobserved behaviors (false alarms) and the degree of bias in raters' decision criteria, especially on more difficult items, were lessened in groups. Overall, compared with individuals, groups were more prone to correction and not amplification of bias and error. A study by Vollrath et al. (1989) used a mock juror paradigm to test recognition memory for trial details and found similar results. Specifically, groups made fewer errors of omission (failure to correctly recognize information that was presented) than did individuals. And also of importance, groups made fewer errors of commission (incorrectly recognizing information that was not presented), which contributed to the somewhat less biased decision criteria that characterized groups as compared with individuals. Although the impact of individuals' versus groups' expectations and beliefs were not explored in these studies, the findings are consistent with the proposition that groups may attenuate individual-level bias.

It would be incorrect, however, to assume that groups are a panacea. Groups sometimes amplify individual biases (e.g., Janis, 1982; Whyte, 1993). As discussed in a review of group information processing (Hinsz et al., 1997), one key to the debiasing potential of groups resides in how group members perceive the nature of the task. When group members believe there is a correct answer, as is likely to be true in a behavioral recognition memory task (see Hinsz, 1990; Kerr, MacCoun, & Kramer, 1996), greater deliberation and information sharing ensues than if the task is seen as purely subjective and thus a matter of judgment. In the absence of a necessarily correct response, groups tend to eschew a deliberative search and instead direct their focus toward quickly reaching a judgment that all members can agree with (Stasser & Stewart, 1992). Heightened discussion and information sharing are, therefore, key to alerting group members to differing accounts of what was observed, which increases the likelihood that the correct response will surface (Hinsz et al., 1997; Stasser & Stewart, 1992). This is what occurred in Stasser and Stewart's (1992) study: When the task was framed as one that contained a single correct answer, groups engaged in greater deliberation and information sharing, which improved the accuracy of the group's final judgment. Thus, we reasoned that if one or more individuals have the correct response, which is likely given the (not large) magnitude of the performance-cue bias,[1] and heightened discussion and information sharing ensues because of the nature of the rating task, the behavioral ratings of groups will be less affected by performance cues.

Hypotheses

Research participants were provided with information regarding the quality of a work group's performance prior to observing a videotape of the group and were then asked to complete a behavioral rating instrument as individuals or as a four-person group. Given prior theory and research, four hypotheses regarding individual versus group differences in the performance-cue bias were tested. We predicted the following of individual and group raters when provided with positive (vs. negative) performance-based information:

> *Hypothesis 1 (hit rates):* Individuals will demonstrate the performance-cue bias, whereas the performance-cue bias will be reduced by groups. That is, individual raters will attribute more effective and fewer ineffective behaviors to the work group, whereas group raters will be significantly less influenced by the nature of the performance information.

> *Hypothesis 2 (decision criterion):* Individual raters will adopt a more liberal decision criterion for effective behaviors and a more conservative decision criterion for ineffective behaviors, whereas the decision criterion adopted by group raters will be significantly less influenced by the nature of the performance information.

> *Hypothesis 3 (false alarm rates):* Individual raters will attribute to the work group more effective and fewer ineffective behaviors that did not occur, whereas group raters will be significantly less influenced by the nature of the performance information.

> *Hypothesis 4 (memory strength):* No significant differences between individual raters and group raters due to the nature of the performance information will be observed; however, the memory strength of group raters will be greater overall.

Method

Participants and Design

Three hundred thirty-three undergraduate students participated in the study in return for course credit. A 2 × 2 × 2 split-plot design was used. The between-subjects variables were performance expectation (positive or negative) and rater (individual or group), and the within-subject variable was behavior type (effective or ineffective). All participants were randomly assigned to experimental conditions, informed consent was obtained, and a debriefing was conducted at the conclusion of each experimental session.

Procedure

Participants observed a 14-min videotape of five men attempting to build a bridge of rope and planks in an effort to get themselves and a box across a pool of water. The videotape is realistic and involving: It is a military training film and depicts college-age recruits working on a task requiring a full range of behaviors that tap into well-known dimensions of group effectiveness, such as leadership ability, quality of communication, and task commitment (see Guzzo, Wagner, Maguire, Herr, & Hawley, 1986; Martell et al., 1995). To control for differential effects due to the presence of others (Zajonc, 1965), all participants observed the videotape in noninteracting groups of 3 to 8. Prior to observing the work group, participants were provided with information regarding the quality of the work group's performance as well as additional information, such as the nature of the study, what the group was attempting to accomplish, and instructions to carefully observe the group's behavior. At the conclusion of the videotape, participants completed a manipulation check and were randomly assigned to an individual or group rater condition. To guard against the possibility of systematic differences in encoding strategies, assignment to groups was not done until after the videotape had been observed. In addition, to obscure the individual–group manipulation, group assignment was done outside the presence of participants in the individual rater condition. Pilot testing had established that 35 min was sufficient time to complete the rating task, and all participants completed the rating task in the allotted time.

---

[1] Although there has been no meta-analytic review to establish the precise magnitude of the performance-cue bias, Lord (1985b, pp. 114–115) presented effect sizes for 11 studies of the performance-cue bias, and we located the reported effects sizes for an additional 12 studies. As expected, the magnitude of the performance-cue bias was not large ($r = .29$). Performance-based feedback accounted for, on average, approximately 8.4% of the variance in performance ratings.

Independent Variables

Performance expectation. Prior to observing the videotape, participants were told that the group had performed extremely well (*positive performance expectation*) or extremely poorly (*negative performance expectation*). Instructions were as follows: "The group you are about to observe was rated by a panel of experts trained in the observation of group performance. Compared with other groups performing this task, the group you are about to observe was judged to be in the top [or bottom] quarter of all groups. That is, the performance of this group was quite good [or quite poor] in comparison with other groups."

Rater. Behavioral ratings were made individually or as a 4-person group. In the group-rater condition, participants were instructed to reach a consensus when deciding whether a behavior had occurred; that is, all group members had to agree on each rating. A member of the group was randomly assigned to record the group's ratings.

Behavior type. All participants judged whether effective and ineffective behaviors occurred in the videotape.

Dependent Measures

Workgroup ratings were made with a 40-item behavioral questionnaire. Participants indicated whether each behavior occurred by using a 6-point scale with endpoints labeled *1 = very certain the behavior did not occur* to *6 = very certain the behavior did occur*. Twenty of the 40 items depicted behaviors that had occurred in the videotape. Of these 20 behaviors, 11 were effective and 9 were ineffective. Of the remaining 20 behaviors that had not occurred, 11 were effective and 9 were ineffective. Some sample items included, "The group loudly encouraged each member crossing the pool" (effective, did occur) and "A group member dropped one of the ropes" (ineffective, did occur). For more on the development and classification of the effective and ineffective behaviors, see Martell and Guzzo (1991).

Participants' behavior ratings were first translated into hit rates and false-alarm rates. A hit is defined as a "yes" response to a behavior that was present in the videotape, and the hit rate is the conditional probability of responding "yes" to a present behavior: Hit rate = *P*(yes/present behavior). Overall hit rates were calculated for each participant by treating effective and ineffective behaviors that were present and rated a *4*, *5*, or *6* (indicating some degree of confidence) as hits. The overall hit rate ranges from 0.0 to 1.0, with higher values indicating a greater percentage of correctly identified behaviors.[2] A false alarm is defined as a "yes" response to a behavior that was not present, and the false alarm rate is the conditional probability of responding "yes" to a behavior that was not present: False alarm rate = *P*(yes/behavior not present). Overall false alarm rates were calculated by treating effective and ineffective behaviors that were not present in the videotape and were assigned a rating of *4*, *5*, or *6* as false alarms. The overall false alarm rate ranged from 0.0 to 1.0, with higher values indicating a greater percentage of incorrectly identified behaviors. Following the recommendation of Snodgrass and Corwin (1988), we transformed hit rates prior to analysis by adding 0.5 to each frequency and dividing by $N + 1$ (N = the number of present behaviors) to eliminate hit rates of 1.0. We also transformed false alarm rates prior to analysis by adding 0.5 to each frequency and dividing by $N + 1$ (N = the number of not present behaviors) to eliminate false alarm rates of 0.0.[3]

Next, measures of decision criteria (Br) and memory strength (Pr) were calculated. Br ranged from 0.0 to 1.0. A Br of .50 indicated that raters were relying on a neutral decision criterion when judging whether a behavior was observed, whereas scores greater than .50 indicated a too-liberal decision criterion, and scores less than .50 indicated a too-conservative decision criterion. Br was computed as follows: Br = False alarm rate/1 − (hit rate − false alarm rate). Pr ranged from −1.0 (no memory) to 1.0 (perfect memory) and is computed as follows: Pr = hit rates − false alarm rates. The overall hit and false alarm rates of individuals and groups were transformed into measures of decision criteria and memory strength. Preliminary analyses revealed that Br and Pr scores for effective and ineffective behaviors were uncorrelated (*M*s = .12 and −.12, respectively). This finding is consistent with prior research by Snodgrass and Corwin (1988), which has demonstrated the independence of these two particular measures of decision criterion and memory strength.

Results

Manipulation Check

So that we could confirm the effectiveness of the performance expectation manipulation, participants evaluated the group's performance by using a 7-point rating scale with the endpoints labeled *1 = extremely poor* and *7 = extremely good*. Analysis of variance (ANOVA) revealed a significant main effect for performance expectation, $F(1, 149) = 180.71$, $p < .01$, which confirmed that participants receiving positive performance information evaluated the group more favorably ($M = 5.43$, $SD = 0.72$) than did participants receiving negative performance information ($M = 3.42$, $SD = 1.00$).

A series of repeated measures ANOVAs were conducted, with performance expectation and rater as between-subjects factors and behavior type as the within-subject factor. Hypotheses 1–3 predict a three-way interaction such that the Performance Expectation × Behavior Type interaction is significant for individual raters but not significant (or less so) for group raters. Hypothesis 4 predicts only a main effect for rater, with greater memory strength for group than individual raters. Means and standard deviations for hit rates, decision criteria, false alarm rates, and memory strength appear in Table 1. Intercorrelations among the study variables are

[2] As with other recent studies of the performance-cue bias (Baltes & Parker, 2000; Martell & Willis, 1993; Martell et al., 1995), measures of overall hit and false alarm rates were calculated by collapsing participants' scale ratings into a dichotomous yes–no judgment rather than averaging participants' confidence ratings. This was done primarily for two reasons. First, it is in keeping with the goal of performance-cue bias research, which is to investigate the effects of observer expectations on the amount of behavior attributed to the group and not overall confidence per se. Second, calculations of memory strength and decision criteria (described next) require that participants' recognition memory performance be collapsed into a dichotomous yes–no rating. What is important is that previous research demonstrates similar findings whether a continuous or dichotomous rating instrument is used and when the continuous nature of the rating scale is eliminated by reducing it to a dichotomous yes–no judgment (see Martell & Willis, 1993; Phillips & Lord, 1982). Nonetheless, dichotomization will necessarily restrict error variance, which may yield overestimates of our resulting effect-sizes.

[3] This correction factor is typically applied when calculating hit rates (and false alarm rates) because "measures . . . are undefined for hit rates of 1.0 or false alarm rates of 0.0 insofar as the corresponding *z*-scores are infinite" (Snodgrass & Corwin, 1988, p. 35). For consistency, the transformation is routinely recommended, even in the absence of zeros and ones.

Table 1
Means and Standard Deviations for Hit Rates, Decision Criteria, False Alarm Rates, and Memory Strength

	Effective work behavior				Ineffective work behavior			
Rater	Hit rates[a]	Decision criteria[b]	False alarm rates[c]	Memory strength[d]	Hit rates[a]	Decision criteria[b]	False alarm rates[c]	Memory strength[d]
Individual								
Positive expectation ($n = 65$)	.78 (.14)	.58 (.24)	.31 (.19)	.47 (.20)	.67 (.18)	.47 (.23)	.28 (.15)	.39 (.17)
Negative expectation ($n = 38$)	.58 (.18)	.31 (.23)	.21 (.25)	.37 (.26)	.78 (.16)	.66 (.22)	.42 (.20)	.36 (.19)
Group								
Positive expectation ($n = 25$)	.66 (.12)	.28 (.16)	.13 (.08)	.53 (.12)	.78 (.16)	.57 (.23)	.26 (.14)	.52 (.20)
Negative expectation ($n = 25$)	.60 (.12)	.18 (.12)	.08 (.06)	.52 (.11)	.80 (.10)	.58 (.16)	.26 (.12)	.54 (.19)

Note. Standard deviations are in parentheses.
[a] Mean values range from 0 (no present behaviors reported) to 1.0 (all present behaviors reported). [b] Decision criteria greater than .50 indicates a liberal decision criterion, decision criteria less than .50 indicates a conservative decision criterion. [c] Mean values range from 0 (no not present behaviors reported) to 1.0 (all not present behaviors reported). [d] Memory strength values range from −1.0 (no memory) to 1.0 (perfect memory).

presented in Table 2. Complete ANOVA results are presented in Table 3.[4]

Behavior Ratings (Hit Rates)

As predicted, the 2 × 2 × 2 repeated measures ANOVA produced a significant Performance Expectation × Behavior Type × Rater interaction, $F(1, 149) = 6.99$, $p < .01$. Simple effects tests were conducted to provide a direct test of whether groups are less sensitive to the performance-cue bias than are individuals, and where appropriate, planned comparisons were conducted. The results support Hypothesis 1.

Individuals as raters. Results revealed a significant Performance Expectation × Behavior Type interaction, $F(1, 101) = 31.01$, $p < .01$. When given positive versus negative performance information, individuals attributed more effective behaviors, $t(101) = 5.94$, $p < .01$, and fewer ineffective behaviors, $t(101) = 3.15$, $p < .01$, to the group.

Groups as raters. Results revealed only a significant main effect for behavior type, $F(1, 48) = 40.69$, $p < .01$, such that more ineffective ($M = .79$) than effective ($M = .63$) work behaviors were attributed to the workgroup. However, the Performance Expectation × Behavior Type interaction, which is essential for establishing the presence of the performance-cue bias, was not significant, $F(1, 48) = 2.78$, $p > .10$. These results demonstrate the absence of a performance-cue bias in group ratings.

Decision Criteria

As predicted, a 2 × 2 × 2 repeated measures ANOVA produced a significant Performance Expectation × Behavior Type × Rater interaction, $F(1, 149) = 7.98$, $p < .01$. Simple effects tests were conducted to interpret the nature of the three-way interaction, and where appropriate, planned comparisons were conducted. The results support Hypothesis 2.

Individuals as raters. Results revealed a significant main effect for behavior type, $F(1, 101) = 9.37$, $p < .01$, and importantly, a significant Performance Expectation × Behavior Type interaction, $F(1, 101) = 35.07$, $p < .01$. When rating effective work behaviors, individuals given positive versus negative performance information adopted a more liberal decision criterion, $t(101) = 5.50$, $p < .01$. However, when rating ineffective behavior, the pattern was reversed such that individuals given positive versus negative performance information adopted a more conservative decision criterion, $t(101) = 4.10$, $p < .01$.

Groups as raters. Results revealed only a significant main effect for behavior type, $F(1, 48) = 79.99$, $p < .01$, such that groups adopted a more conservative decision criterion when rating effective ($M = .24$) versus ineffective ($M = .58$) work behaviors. More central to the predictions of this study, the Performance Expectation × Behavior Type interaction was not significant, $F(1, 48) = 2.32$, $p > .10$.

False Alarm Rates

The 2 × 2 × 2 repeated measures ANOVA produced, as predicted, a significant Performance Expectation × Behavior Type × Rater interaction, $F(1, 149) = 4.87$, $p < .03$. Simple effects tests were conducted to interpret the nature of the three-way interaction, and where appropriate, planned comparisons were conducted. The results support Hypothesis 3.

Individuals as raters. Results revealed a significant Performance Expectation × Behavior Type interaction, $F(1, 101) = 17.21$, $p < .01$. When given positive versus negative performance information, individuals attributed to the group more effective behaviors, $t(101) = 2.21$, $p < .03$, and fewer ineffective behaviors, $t(101) = 4.17$, $p < .01$, that were not present.

Groups as raters. Results revealed only a significant main effect for behavior type, $F(1, 48) = 54.88$, $p < .01$, such that more ineffective ($M = .79$) than effective ($M = .63$) work behaviors that

[4] There was a small departure from the homogeneity of variance assumption on some of the dependent variables in this study. Thus, we also analyzed the results by using a mixed ANOVA model, which assumed two variance estimates, one for individuals and one for groups. This statistical procedure is useful insofar as it circumvents the problem of differences in variability (Pinheiro & Bates, 2000, pp. 208–226). The results replicated all the original ANOVA findings; each of our four hypotheses were confirmed. The resulting test statistics (F values, p values) were all extremely close to those in the original ANOVA except that the effects were a bit stronger. Thus, the original ANOVA provided a more conservative test, and it is these results that are reported.

Table 2
Descriptive Statistics and Intercorrelations Among all Variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. HR+	.68	.17	—							
2. HR−	.74	.17	−.33**	—						
3. FAR+	.22	.20	.42**	−.20**	—					
4. FAR−	.31	.16	−.34**	.35**	−.02	—				
5. Pr+	.46	.20	.42**	−.07	−.65**	−.26**	—			
6. Pr−	.42	.19	−.01	.59**	−.16*	−.55**	.15	—		
7. Br+	.40	.26	.80**	−.41**	.78**	−.20**	−.12	−.19**	—	
8. Br−	.55	.23	−.39**	.82**	−.13	.72**	−.20**	.12	−.38**	—

Note. $N = 153$. HR = hit rate; FAR = false alarm rate; Pr = memory strength; Br = decision criterion; Plus signs indicate effective behaviors; Minus signs indicate ineffective behaviors.
* $p < .05$. ** $p < .01$.

were not present were attributed to the work group. What is important is that the Performance Expectation × Behavior Type interaction was not significant, $F(1, 48) = 1.49$, $p > .10$, which indicates that performance cues had no effect on groups' false alarm rates.

Memory Strength

Finally, a 2 × 2 × 2 repeated measures ANOVA revealed only a significant main effect for rater, $F(1, 149) = 30.22$, $p < .001$, indicating greater memory strength for groups than individuals (*M*s = .53 and .40, respectively). Thus, Hypothesis 4 was supported. In addition, the results yielded further evidence that participants were not responding so as to simply confirm experimenter-induced expectations. If an experimental demand of this type had been operating, it is very unlikely that participants would have bothered to distinguish between expectancy-consistent behaviors that were present versus absent and, as a result, memory strength results would have approximated 0.0. Memory strength scores departed significantly (and substantially) from 0.0, however.

Discussion

Individuals believing that a workgroup had performed well (versus poorly) attributed more effective and fewer ineffective behaviors to the group. This finding replicates previous research (Martell & Guzzo, 1991; Martell & Willis, 1993; Martell et al., 1995). Group-based judgments, however, were unaffected by the nature of the feedback. Thus, groups were effective in eliminating the performance-cue bias. Additional analyses point to two reasons as to why groups were resistant to the performance-cue bias. As in previous research (Baltes & Parker, 2000; Martell & Guzzo, 1991; Martell & Willis, 1993; Martell et al., 1995), individuals adopted a decision criterion that was more liberal for expectancy-consistent behavior and more conservative for expectancy-inconsistent behavior. In contrast, the decision criterion adopted by groups was unaffected by performance expectations. In addition, performance feedback led individuals (but not groups) to commit more errors of commission in the form of higher false alarm rates for behaviors that were consistent with their performance expectations. These results, combined with the findings of Baltes and Parker (2000),

Table 3
Analysis of Variance Results of Behavior Ratings

Source	*df*	Hit rates *F*	Hit rates η^2	Decision criteria *F*	Decision criteria η^2	False alarm rate *F*	False alarm rate η^2	Memory strength *F*	Memory strength η^2
Between subjects									
Performance expectation (A)	1	3.12	.02	4.07*	.02	0.00	.00	1.43	.01
Rater (B)	1	0.28	.00	22.17**	.13	36.64**	.19	30.22**	.16
A × B	1	0.27	.00	0.07	.00	1.28	.00	1.77	.01
Within subjects									
Behavior type (C)	1	22.45**	.13	57.31**	.27	30.76**	.17	0.67	.00
A × C	1	21.04**	.12	22.46**	.08	11.28**	.07	1.26	.00
B × C	1	7.22**	.04	13.40**	.13	2.14	.01	1.30	.00
A × B × C	1	6.99**	.04	7.98**	.05	4.87*	.03	0.15	.00
Error	149								

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

are encouraging and demonstrate that the performance-cue bias can be eliminated.

Implications for Theory Testing

Concerns regarding the performance-cue bias manifest a more fundamental question: Is believing seeing? That is, do a priori beliefs and expectations cause individuals to (mis)report what they think they have seen? This concern is well-founded. Much of what we portend to know from research on organizational behavior and the theories that result is linked inexorably to the basic proposition that individuals' recollections are mostly accurate. But what if sometimes they are not? What if retrospective accounts are systematically biased in a manner that escapes detection? Indeed, the impetus for Staw's (1975) initial investigation of the performance-cue bias was concern regarding the accuracy of self-report data used to develop and test theories of organizational behavior and the belief that performance-based expectations could systematically bias recollection so as to "confirm" such theories. More recent work indicates that concern over potentially biased retrospective accounts has not abated among organizational scholars (see Golden, 1992; Huber & Power, 1985; Schwenk, 1985). This suggests that researchers engaged in theory testing at the individual, group, or organizational level should be wary of the performance-cue bias as a possible contaminant and consider the use of groups as raters.

Implications for Assessment Center Appraisals

Group-based performance evaluation frequently occurs in assessment centers. A reading of Spychalski et al.'s (1997) survey of assessment center practices and other investigations (e.g., Howard, 1997; Lowry, 1997; Zedeck, 1986) have suggested that our study embodied many of the distinguishing features of the modal assessment center and thus permits a degree of generalization. As in our study, assessors often observe a group working on a task, use a behavioral rating instrument during or immediately after observation, and meet as a group to pool their observations in an effort to increase accuracy and reduce bias. During this latter integration stage, a consensus decision rule is typically in place, as was true in our study. The task-specific assessment center described by Lowry (1997), which concentrates on identifying the presence or absence of concrete behaviors (by means of a behavioral checklist) related to successful performance, seems closest to the nature of the rating task faced by our study participants. Furthermore, our use of a videotape (and not a live work sample) is consistent with an increasing reliance on videotapes of assessee performance for observation and evaluation (Bray & Byham, 1991). It is important to note that Ryan et al. (1995) reported no differences in assessment center ratings of a videotaped versus live work sample. With this in mind, our results speak most directly to a concern raised in Zedeck's (1986) information processing model of the assessment center process. Zedeck argued that assessor expectations, which could be based on numerous factors (e.g., successful completion of the exercise; see Guzzo et al., 1986), might contaminate behavioral observations. Specifically, Zedeck raised the possibility of assessors relying on performance-based expectations to simplify the task of observing and remembering assessee behaviors and suggested that research on the dynamics of such cognitive heuristics, as they may operate in a group context, was warranted. In a similar manner, Ilgen et al.'s (1993) review of the contribution of cognitive approaches to performance appraisal research called for greater attention to the "identification of *workgroup* [italics added] . . . factors" that influence rater cognitive processes (p. 362). Our study directly addresses some of these concerns and suggests that the integration session, in which multiple assessors discuss and decide whether behaviors were observed, can play a valuable role in diminishing the biasing effects of assessor expectations. Groups may help reduce the very sort of systematic biases identified by Zedeck (1986) and others (Guzzo et al., 1986; Ilgen et al., 1993) that can arise in assessment center judgments.

Assessing Groups

Some readers may be surprised that the target of evaluation in this study was a workgroup and not an individual. Such surprise is no longer warranted. It seems that with each passing decade, at least since the 1970s, organizational theorists have speculated that groups would challenge individuals for primacy as the basic unit of performance in organizations (e.g., Cummings, 1978; Dumaine, 1994; Reich, 1983). Recent reviews on the status of groups in organizations suggest that groups are well on the way to supplanting individuals as the basic organizational unit (Guzzo & Shea, 1992; Ilgen, 1999; Katzenbach & Smith, 1993). According to Guzzo and Shea (1992), "Today more than ever, organizations rely on work groups, as evidenced by the appearance of project teams, focus groups, autonomous work groups, quality circles, multifunctional work teams, and team CEOs" (p. 270). Not surprisingly, groups are increasingly being hired, appraised, rewarded, and promoted. Thus, with the ascendance of groups as the primary unit of performance, accurate and unbiased measurement of group performance will become a critical challenge. A nice treatment of some of the difficulties associated with evaluating the work performance of groups can be found in Brown (1984). Accordingly, we offer a prediction of our own. Performance appraisal research will evidence a shift, not away from individual assessment entirely, but more in the direction we have taken in assessing the performance of a group.

When Might Groups Help?

Although our study results suggest that groups can be used to avoid the introduction of bias in retrospective accounts, it is important to note that for groups to be an asset and not a potential liability, two conditions are relevant. First, the magnitude of the particular bias cannot be large. When a bias is large, and thus pervasive, group-based judgments may reflect amplification and not attenuation. Fortunately many social judgment biases reside in the small-to-moderate range and, thus, it is not unusual for individuals to sometimes escape their influence. This suggests that in a group context there is an increased probability that one or more members will have an unbiased response. However, the presence of an unbiased response often is not enough; it must surface in the course of group deliberation. Research has found that when a task is believed to have a single correct answer, groups engage in greater deliberation and information sharing, which heightens the probability that the correct response will be identified (Stasser & Stewart, 1992). Thus, task perception is a second important factor. There is reason to believe that these two conditions were met in our study. The performance-cue bias does not yield a statistically large effect and, thus, it is reasonable to presume that some individuals were unaffected by it. In addition, even individuals affected by the performance-cue bias are unlikely to have produced biased responses on all of the 40 behavior ratings; therefore, these group members will have at their disposal a mix of biased and unbiased responses. In addition, because a behavior either occurred or it did not, the recognition memory task possessed a single correct response (see Hinsz, 1990; Kerr et al., 1996). It is interesting to note that group members need only to believe that a task has a necessarily correct answer. When faced with a task that is simply described as having a necessarily correct response, groups will engage in heightened deliberation and information sharing (Stasser & Stewart, 1992). Thus, how a task is perceived, aside from its objective features, matters.

We are mindful, however, that groups are not a panacea and that specific conditions probably must be in place if groups are to be effective in reducing the biases (e.g., contrast assimilation; primacy-recency effects; halo) that can arise in performance appraisals. One of these conditions, that raters believe there is a single correct response, warrants elaboration.

As discussed by Sulsky and Balzer (1988), not all performance appraisals are alike. We agree. Some appraisals are likely to be perceived by raters as having a necessarily correct answer; this is true of behavioral recognition ratings, which require raters to accurately distinguish between behaviors that did and did not occur. The findings of this study, as well as previous research, provide reason to believe that groups can be an asset to such tasks. But what of appraisal judgments that are more subjective insofar as the accuracy of an assessment is, by definition, more uncertain? Groups may be tempted to construe such a task as simply a "matter of judgment" with no necessarily correct or even best answer. If so, groups may abandon a more deliberative strategy and instead engage in heuristic information processing in an effort to more easily reach agreement; this would amplify individual biases. Thus, we propose another potentially key factor—*potential for feedback* regarding the accuracy of a subjective evaluation. Consider, for example, a group charged with evaluating an individual's potential for future success in a work setting. This is a rather subjective assessment. Imagine, however, that the group had been forewarned that it will learn of the accuracy of its assessment sometime in the future. That is, group members will be informed of the fate of individuals whose potential for success had been evaluated by the group. We predict that the potential for feedback—learning the extent to which a group-based judgment proved to be accurate—will foster greater deliberation and information sharing within the group. Therefore, to the extent a (not large) judgmental bias is affecting some but not all group members, the probability of an unbiased response being correctly identified will be heightened. This preliminary formulation is consistent with previous findings that groups are often more responsive to feedback than are individuals (Hinsz et al., 1997; Tindale, 1989) and accords rater motivation a central role in the appraisal process, as recommended by others (Banks & Murphy, 1985; Mero & Motowidlo, 1995).

It is worth making two additional points regarding our results. First, the findings cause us to modify a conclusion rendered in a study by Martell and Borg (1993), namely, that groups necessarily amplify bias in individual raters' decision criteria. Although evidence of a greater (not lesser) response bias in group versus individual judgments was found, the Martell and Borg study differs from the present study in at least two ways. Performance-based feedback was not provided to study participants and thus the decision criteria adopted by individuals hovered close to the neutral point (i.e., .50). In other words, it was not a study of the performance-cue bias. In addition, whereas the present study used a videotape comprised of many behaviors, Martell and Borg relied on a written description of an individual's work performance that contained fewer behaviors. These differences may also explain why group memory strength was superior to individuals in our study but only in the delayed (and not immediate) rating conditions in Martell and Borg. Close inspection reveals that individuals' memory strength in the immediate rating condition was substantially greater in Martell and Borg's study than in ours, with memory strength means equal to .65 and .39, respectively. Given the heightened memory accuracy of individuals, probably due to the greater ease of remembering written behaviors than remembering a videotaped behavioral sequence, there was less to be gained from the pooling of group members' mostly accurate memories in Martell and Borg. However, as individuals' memories began to erode with the passage of time, as evidenced by the declining memory strength score observed in Martell and Borg's delayed rating condition (memory strength = .46), the predicted benefits of pooling individual memories emerged, producing superior group (vs. individual) memory. At this juncture, we are reminded of Kerr et al.'s (1996) observation of how differences in tasks, procedures, or experimental design can modify whether groups attenuate, amplify, or have no effect on individual-level bias. We agree that there appears to be no "simple, global answer to the question, 'Is group judgment more or less biased than individual judgment?' " (Kerr et al., 1996, p. 693).

Second, given that the performance-cue bias is not large, what should be made of its practical importance? This question is far more complex than it appears insofar as the relationship between the size of an effect and its real-world impact is not at all obvious. Indeed, the counterintuitive notion that extremely small effects can yield enormous real-world consequences is well-documented across scientific disciplines (e.g., Merton, 1936; Pinker & Bloom, 1992; Trivers, 1971). Let us consider just one example. Using computer simulation to model a virtual organization, Martell, Lane, and Emrich (1996) demonstrated that exceedingly small amounts of bias in performance ratings can severely dampen one's organizational mobility and dramatically reduce the probability of reaching senior levels of management. These results illustrate the cumulating effects of very small amounts of bias in performance ratings, which, given the factors governing organizational mobility (see Martell et al., 1996; Martell & Robison-Cox, 2001) can disadvantage the targets of bias.

Potential Limitations

Several potential limitations of this study warrant note. Because we relied on student ad hoc groups, as is true of much research on group decision making, questions of generalizability inevitably arise. Although the performance-cue bias is found outside laboratory settings, and research using student samples to explore the dynamics of work performance evaluations seems to enjoy a reasonable degree of external validity (see Dipboye, 1990; Locke, 1986), whether intact groups with an extensive history would have eliminated the performance-cue bias is an important question that must await additional research. In addition, although we correctly anticipated that group's hit rates, false alarm rates, and decision criteria would be unaffected by performance-based feedback, no direct measures of group interaction were collected. Thus, we can only infer the exact nature of group interaction that led to avoiding the performance-cue bias. In the future, audio and video recording devices could be used to provide a more fine-grained analysis of the dynamics that transpire within groups engaged in performance appraisal.

Conclusion

In all, the results of this study suggest a role for groups in theory-testing efforts and performance appraisal, and we look forward to additional research aimed at establishing when groups are an asset or a liability. Surely, they can be both. To this end, we highly recommend Wherry and Bartlett's (1982) often overlooked "theory of rating" as an excellent point of departure and more recent research on task conditions that foster or inhibit group deliberation and information sharing.

References

Baltes, B. B., & Parker, C. P. (2000). Reducing the effects of performance expectations on behavioral ratings. *Organizational Behavior and Human Decision Processes, 82,* 237–267.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38,* 335–345.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48,* 299–337.

Bray, D. W., & Byham, W. C. (1991). Assessment centers and their derivatives. *The Journal of Continuing Higher Education, 39,* 8–11.

Brown, K. A. (1984). Explaining group poor performance: An attributional analysis. *Academy of Management Review, 9,* 54–63.

Clark, N. K., & Stephenson, G. M. (1990). Group remembering. In P. Paulus (Ed.), *Psychology of group influence* (2nd ed., pp. 357–391). Hillsdale, NJ: Erlbaum.

Cummings, T. G. (1978). Self-regulating work groups: A sociotechnical synthesis. *Academy of Management Review, 3,* 625–634.

Dipboye, R. L. (1990). Laboratory vs. field research in industrial and organizational psychology. In C. L. Cooper & L. T. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 5, pp. 1–34). New York: Wiley.

Dumaine, B. (1994). The trouble with teams. *Fortune, 131,* 86–92.

Feldman, J. M. (1994). On the synergy between theory and application: Social cognition and performance appraisal. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition: Applications* (Vol. 2, pp. 339–398). Hillsdale, NJ: Erlbaum.

Fichman, M. (1991). Attribution processes in the use of performance feedback: A field test. *Advances in Information Processing in Organizations, 4,* 182–200.

Gardner, T. M., Wright, P. M., & Gerhart, B. A. (2000). *The HR-firm performance relationship: Can it be in the mind of the beholder?* (Working Paper 00–02). Ithaca, NY: Cornell University, School of Industrial Relations, Center for Advanced HR Studies.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117,* 21–38.

Golden, B. R. (1992). The past is the past—or is it? The use of retrospective accounts as indicators of past strategy. *Academy of Management Journal, 35,* 848–860.

Guzzo, R. A., & Shea, G. P. (1992). Group performance and intergroup relations in organizations. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 269–313). Palo Alto, CA: Consulting Psychologists Press.

Guzzo, R. A., Wagner, D. B., Maguire, E., Herr, B., & Hawley, C. (1986). Implicit theories and the evaluation of group process and performance. *Organizational Behavior and Human Decision Processes, 37,* 279–295.

Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology, 59,* 705–718.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin, 121,* 43–64.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12,* 13–52.

Huber, G. P., & Power, D. J. (1985). Retrospective reports of strategic level managers: Guidelines for increasing their accuracy. *Strategic Management Journal, 6,* 171–180.

Ilgen, D. R. (1999). Teams embedded in organizations: Some implications. *American Psychologist, 54,* 129–139.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54,* 321–368.

Janis, I. L. (1982). *Groupthink* (2nd ed.). Boston: Houghton-Mifflin.

Katzenbach, J. R., & Smith, D. K. (1993). *The wisdom of teams.* Cambridge, MA: Harvard Business School Press.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review, 103,* 687–719.

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

Lord, R. G. (1985a). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66–71.

Lord, R. G. (1985b). An information processing approach to social perception and behavioral measurement in organizational settings. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 7, pp. 87–128). Greenwich, CT: JAI Press.

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality, 12,* 53–62.

Martell, R. F., & Borg, M. R. (1993). A comparison of the behavioral rating accuracy of groups and individuals. *Journal of Applied Psychology, 78,* 43–50.

Martell, R. F., & Guzzo, R. A. (1991). The dynamics of implicit theories of group performance: When and how do they operate? *Organizational Behavior and Human Decision Processes, 50,* 51–74.

Martell, R. F., Guzzo, R. A., & Willis, C. E. (1995). A methodological and substantive note on the performance-cue effect in ratings of work group behavior. *Journal of Applied Psychology, 80,* 191–195.

Martell, R. F., Lane, D. M., & Emrich, C. E. (1996). Male-female differences: A computer simulation. *American Psychologist, 51,* 157–158.

Martell, R. F., & Robison-Cox, J. (2001, May). *How a little bias can hurt women managers a lot.* Paper presented at the Western Psychological Association annual meeting, Maui, HI.

Martell, R. F., & Willis, C. E. (1993). Effects of observers performance expectations on behavior ratings of work groups: Memory or response bias? *Organizational Behavior and Human Decision Processes, 56,* 91–109.

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 78,* 43–50.

Merton, R. (1936). The unanticipated consequences of purposive social action. *American Sociological Review, 1,* 894–904.

Motowildo, S. J. (1986). Information processing in personnel decisions. In G. R. Ferris & K. R. Rowland (Eds.), *Research in personnel and human resource management* (Vol. 4, pp. 1–44). Greenwich, CT: JAI Press.

Murphy, K. R. (1991). Criterion issues in performance appraisal research: Behavioral accuracy versus classification accuracy. *Organizational Behavior and Human Decision Processes, 50,* 45–50.

Phillips, J. S., & Lord, R. G. (1982). Schematic information processing and perceptions of leadership in problem-solving groups. *Journal of Applied Psychology, 67,* 486–492.

Pinheiro, J. C., & Bates, D. D. (2000). *Mixed-effects models in S and S-plus.* New York: Springer-Verlag.

Pinker, S., & Bloom, P. (1992). Natural language and natural selection. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 451–493). Oxford, England: Oxford University Press.

Reich, R. B. (1983). *The next American frontier.* New York: Times Books.

Ryan, A. M., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T., et al. (1995). Direct, indirect, and controlled observation and rating accuracy. *Journal of Applied Psychology, 80,* 664–670.

Sackett, P. R., & Wilson, M. A. (1982). Factors affecting the consensus judgment process in managerial assessment centers. *Journal of Applied Psychology, 67,* 10–17.

Schwenk, C. R. (1985). The use of participant recollection in the modeling of organizational decision processes. *Academy of Management Review, 10,* 496–503.

Shea, G. P., & Guzzo, R. A. (1987). Groups as human resources. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 5, pp. 323–356). Greenwich, CT: JAI Press.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.

Spychalski, A. C., Quinones, M. A., Gaughler, B. B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50,* 71–90.

Stahlberg, D., Eller, F., Maas, A., & Frey, D. (1995). We knew it all along: Hindsight bias in groups. *Organizational Behavior and Human Decision Processes, 63,* 46–58.

Stasser, G., & Stewart, D. (1992). Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology, 63,* 426–434.

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.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497–506.

Tindale, R. S. (1989). Group vs. individual information processing: The effects of outcome feedback on decision making. *Organizational Behavior and Human Decision Processes, 44,* 454–473.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46,* 35–56.

Vollrath, D. A., Sheppard, B. H., Hinsz, V. B., & Davis, J. H. (1989). Memory performance by decision-making groups and individuals. *Organizational Behavior and Human Decision Processes, 43,* 289–300.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521–550.

Whyte, G. (1993). Escalating commitment in individual and group decision making: A prospect theory approach. *Organizational Behavior and Human Decision Processes, 54,* 430–455.

Wright, E. F., & Christie, S. D. (1989). *The impact of group discussion on the theory-perseverance effect.* Unpublished manuscript, St. Francis Xavier University, Antigonish, Nova Scotia, Canada.

Wright, E. F., Lüüs, C. A., & Christie, S. D. (1990). Does group discussion facilitate the use of consensus information in making causal attributions? *Journal of Personality and Social Psychology, 59,* 261–269.

Wright, E. F., & Wells, G. L. (1985). Does group discussion attenuate the dispositional bias? *Journal of Applied Social Psychology, 15,* 531–546.

Zajonc, R. (1965). Social facilitation. *Science, 149,* 269–274.

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in Organizational Behavior, 8,* 259–296.

Received July 31, 2001
Revision received January 31, 2002
Accepted February 5, 2002 ■