Reducing the performance cue bias in work behavior ratings : do groups help or hurt?
by Keith Norman Leavitt

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Applied Psychology
Montana State University

Abstract:
The present study compared the use of group versus individual raters in a behavioral rating task to determine whether groups would attenuate the performance cue bias. Three hundred and three participants were given positive or negative feedback regarding the performance of a work group, and following observation of the work group, completed a work-behavior questionnaire either individually or in a four-person rating group. As was predicted, individual (but not group) raters were systematically biased to identify behaviors congruent with feedback given, such that they identified more effective and fewer ineffective behaviors when given feedback of relatively good (versus poor) performance. In addition, the false alarm rates and decision criterion of individual (but not group) raters were found to be systematically biased by performance information as well. Two factors that predict the likelihood that groups will attenuate individual level bias—bias magnitude and task perception—are identified. Implications for performance appraisal theory, research, and practice are discussed.

REDUCING THE PERFORMANCE CUE BIAS IN WORK BEHAVIOR RATINGS: DO GROUPS HELP OR HURT?

by

Keith Norman. Leavitt

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Applied Psychology

MONTANA STATE UNIVERSITY
Bozeman, Montana

September 2001

APPROVAL

of a thesis submitted by

Keith Norman Leavitt

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

Richard F. Martell _____ 10-1-01
(Signature) (Date)

Approved for the Department of Psychology

A. Michael Babcock _____ 10-1-01
(Signature) (Date)

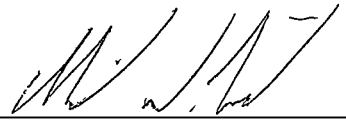
Approved for the College of Graduate Studies

Bruce R. McLeod _____ 10-5-01
(Signature) (Date)

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis in whole or in parts may be granted only by the copyright holder.

Signature 

Date 9.27.01

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to Eric Semes for his tireless hours in the lab, Dr. Richard Martell for his constant input and guidance throughout the course of this project, my committee members Dr. Chuck Pierce and Dr. Wes Lynch for their time and feedback, and finally my parents for their constant support throughout the past two years.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

The present study compared the use of group versus individual raters in a behavioral rating task to determine whether groups would attenuate the performance cue bias. Three hundred and three participants were given positive or negative feedback regarding the performance of a work group, and following observation of the work group, completed a work-behavior questionnaire either individually or in a four-person rating group. As was predicted, individual (but not group) raters were systematically biased to identify behaviors congruent with feedback given, such that they identified more effective and fewer ineffective behaviors when given feedback of relatively good (versus poor) performance. In addition, the false alarm rates and decision criterion of individual (but not group) raters were found to be systematically biased by performance information as well. Two factors that predict the likelihood that groups will attenuate individual level bias—bias magnitude and task perception—are identified. Implications for performance appraisal theory, research, and practice are discussed.

INTRODUCTION

Implicit Theories of Performance and the "Performance Cue Bias"

In a classic study conducted by Staw (1975), business students assigned to four-person groups evaluated their own group processes across nine evaluative dimensions after being falsely told that their group's performance was in the top 10% or bottom 10% of all groups performing the task. Importantly, there were no objective differences in the groups' outcomes or processes. Results indicated that groups told that they had performed "quite well" rated their group's processes more favorably than groups led to believe they had performed "quite poorly." An explanation of the differences in ratings of group processes by the "successful" versus "unsuccessful" groups is that the presence of performance cues (feedback that the group did well or poorly) led group members to make attributions about their group's processes in an attempt to explain the outcomes. Specifically, by relying on an implicit theory, which states that there is a likely relationship between a group's outcome and its processes, individual group members infer that their "process" must be consistent with their "outcome." This would lead "effective" groups to rate their process more favorably then "ineffective" groups. Staw thus concluded that theory development in organizational behavior might be seriously compromised, as: "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 has demonstrated that this performance cue bias also influences behavioral assessments that require observers to judge whether behaviors were observed. Martell and his colleagues (Martell & Guzzo, 1991; Martell & Willis, 1993)

found that performance cues significantly influence responses to even very specific behavioral items. That is, observers attributed significantly more effective behaviors and fewer ineffective behaviors when told that the group's performance was "quite good" versus "quite poor". Thus, performance-related feedback creates an expectation within observers, which will likely bias their subsequent evaluations and behavior ratings. In addition, Martell and Guzzo (1991) demonstrated that performance cues influence behavior ratings even when the performance expectations were presented after the observation of the group, suggesting that the bias associated with feedback did not occur at the time of encoding. A follow-up study by Martell, Guzzo, and Willis (1995) implemented a control condition under which no feedback was given. Under this "no-feedback" condition, raters tended to identify behaviors that were congruent with their own self-generated expectations. Thus, the possibility of a demand characteristic was ruled out, and it was further demonstrated that expectations will bias ratings.

Despite the implications of this performance cue bias for theory development and performance appraisal, only one study to date has been aimed at reducing the biasing effects of performance cues. A recent study by Baltes and Parker (2000) found that a free recall intervention (asking subjects to search recall memory for both effective and ineffective behaviors) was effective in moderating the effect for behavioral ratings. Thus, the performance cue bias was removed by adjusting availability for non-cue consistent behaviors. By forcing raters to make a cognitively effortful judgment, deeper processing occurred and judgments were less affected by performance-related expectations (raters processing at a "deeper" level are spending more time deliberating, and considering information from a "hypothesis testing" rather than from a "hypothesis confirming" perspective). Therefore, there is evidence that the performance cue bias is not inevitable; it can be reduced or eliminated. To date, research on the performance cue bias has depended on the use of individuals as raters, with little attention to whether groups might

be less (or more) affected by performance expectations. The present study examines whether groups as raters might be yet another means of reducing the biasing effects of performance feedback on behavioral ratings.

Groups as Raters

A current trend in organizations is to rely on groups as decision makers. Shea and Guzzo (1987), in a thorough examination of work teams, identified groups as a unique human resources tool; "It is time for the field of personnel and human resources management to discover formal groups as resources in their own right" (p. 323). Shea and Guzzo suggest that one appropriate use of the specific abilities of groups may be in the area of performance appraisal; further attention to groups within the field suggests that the use of groups within organizations for work previously performed by individuals is increasing (Boyett & Conn, 1991; Hackman, 1990).

There is also research to suggest that groups perform differently than individuals as information processors, and may yield assessments that differ from those made by individual performance evaluators (see Hinsz, Vollrath, & Tindale 1997 for a review of research to date). As an example, a "theory of rating" developed by Wherry and Bartlett (1982) identifies factors other than ratee performance that will affect performance evaluation and behavioral ratings, including the observation made by the rater, and later recall by the rater. In one theorem of their model, Wherry and Bartlett (1982) state "plural raters" should be used in a performance appraisal judgment. Multiple raters, as in a group, may reduce error and bias in performance appraisals. Because group members will bring more potentially correct information to the judgment task, it is likely that they will correct individual level biases (such as the performance cue bias).

Empirical tests of the effectiveness of groups in reducing social judgment biases

further suggests the potential efficacy of groups in correcting individual-level bias. Wright, Christie, and Luus (1990) found that group discussion facilitated the use of consensus information that is often ignored by individuals, and thus groups attenuated the "consensus under-utilization effect." Wright and Wells (1985) demonstrated that group discussion attenuated the fundamental attribution error. This occurs because the fundamental attribution error is a statistically small effect. Accordingly, not everyone is affected by the fundamental attribution error, which suggests that in a group context some individuals are unaffected, and thus are in a position to influence the group. Indeed, neither a lengthy deliberation nor simply expecting group discussion alone were sufficient to attenuate the fundamental attribution error; some feature of group process most likely mediates the bias. Thus, it is likely that the extent to which group members possess non-redundant information and some members are unaffected by the bias determines the likelihood that proper causal attribution will be made. Wright and Christie (1989) found that group discussion also eliminated the theory perseverance effect, whereby individuals continue to maintain pervasive social beliefs in the presence of contradictory and disconfirmatory information. This is likely mediated by the expectation of scrutiny of individual judgments, and judgments are defended to the group with supporting evidence.

Group Memory

To understand why groups may help reduce the performance cue bias, the dynamics of the performance cue bias must first be explained. In the Martell and Guzzo (1991) study, memory for effective and ineffective behaviors is fully captured by two theoretically uncorrelated indexes: Memory strength (Pr) and Response bias (Br) (Snodgrass & Corwin, 1988). Memory strength is a measure of behavioral information

stored in memory and ranges from –1.0 to 1.0. Response bias is a measure of the decision criterion in use when deciding whether a behavior was observed previously. A decision criterion can be neutral (Br=.5; no tendency to over or under recognize behaviors), overly liberal (Br>.5; a tendency to over recognize behaviors, including those that did not occur), or overly conservative (Br<.5; a tendency to dismiss behaviors, including those that did occur). In Martell and Guzzo (1991), performance cue biases were not found to be a function of memory strength, but rather due to a systematic response bias, whereby the raters adopted a more liberal decision criterion in identifying behaviors congruent with the performance cue (effective behaviors for good groups, ineffective behaviors for poor groups) and a more conservative decision criterion in identifying behaviors incongruent with the performance cue (Martell & Guzzo, 1991). Martell and Willis (1993) demonstrated that effects of knowledge of performance outcomes were mediated by a systematic response bias, and not a memory limitation in judgments of effective/ineffective work behaviors.

According to group memory research, if groups are to help in the case of the performance cue bias, they must lesson the bias in Br, as well as resist reporting behaviors that did not occur. Hinsz (1990) identifies two types of errors in information processing relevant to performance evaluation: errors of omission (failing to identify behaviors that did occur, or "misses") and errors of commission (identifying behaviors that did not actually occur (false-alarm rates). Vollrath, Sheppard, Hinsz, and Davis (1989) found that groups tend to make fewer errors of commission, and that groups serve to correct errors common to individuals within the group, through greater pooled memory, and by reducing response bias. If a few individual group members can recognize a behavior as a false alarm, then under conditions of high deliberation, groups will correctly dismiss the behavior (Hinsz, 1990).

This model of greater combined resources and bias correction, however, requires

that not all group members are equilaterally affected by the bias. Because individual group members possess somewhat unique information, the potential to identify discrepancies between individual memories and correct error is greater for groups so long as sufficiently detailed group deliberation occurs. Hinsz (1997) found that groups are generally more receptive to feedback than are individuals, thus group members may process information more deeply than they would as individuals when under a consensus decision rule (Hinsz, 1997; Sniezek, 1992), especially on difficult items where deeper deliberation is more likely to occur. In addition, Vollrath, Sheppard, Hinsz, and Davis (1989) confirmed that group recognition memory is greater than that of individuals, and that groups are more frequently able to correctly dismiss false alarm behaviors.

The likelihood of groups attenuating the performance cue bias depends upon the relatively small effect size of the bias. If groups are to help, members must possess slightly different levels of susceptibility to the bias (suggested by the relatively small effect size) and the motivation to more deeply discuss information and tease out/correct individual level errors (likely when the task offers a "correct" solution). According to the model of groups as a human resource proposed by Shea and Guzzo (1987), tasks that require a high degree of group member interdependence for achieving the specified work group goal will yield greater performance than that of individuals working alone. Given a realistic task with a high degree of interdependence (such as evaluating work behavior), group members will need to vest a high degree of accountability for task performance and consequently might outperform individuals in making accurate work behavior ratings. Because it is assumed that there is one correct answer for an item in a behavioral rating task, groups under a consensus decision rule will probably process information more deeply than will individual raters, thus increasing the likelihood of identifying discrepant memories and avoiding false alarms. In addition, because group members will have somewhat different memories and a greater chance for bias correction through

deliberation, the likelihood of performance feedback systematically biasing consensus groups should be lesser than it is for individual raters.

This is not to suggest, however, that groups will serve as a panacea for removing all biases related to social judgment. Groups sometimes amplify individual-level biases (e.g., Janis, 1982; Whyte, 1993). Research has shown groups may help reduce biases in social judgment when members are (a) motivated to share their distinct information, and (b) are able to recognize and make effortful attempts at correcting individual level errors. It is interesting to note, however, that the nature of the actual judgment task itself may prove less important than rater perceptions of the task (Stasser & Stewart, 1992). Greater information sharing and deeper deliberation occurs when group members simply believe that the task has a correct answer, as is likely to be true in a behavioral recognition memory task (see Hinsz, 1990; Kerr, MacCoun, & Kramer, 1996) than if the task is viewed as purely subjective. Thus, heightened discussion and sharing of disseminated information are key to increasing the probability of identifying differences in individual accounts, which increases the likelihood that the correct response to an item will surface. This is what occurred in Stasser and Stewart's (1992) study: when the task was framed as having a single correct answer, groups engaged in longer deliberation and shared more information, improving the accuracy of the group's final judgment. Accordingly, in the context of the performance cue bias, it was reasoned that if one or more group members has the correct response, which is likely given the relatively small magnitude of the performance cue bias, heightened discussion and sharing of disseminated information would ensue given the nature of the recognition memory task; consequently, the behavior ratings made by groups would be less biased by the nature of the performance feedback than the ratings made by individuals.

Hypotheses

Three distinct literatures thus suggest that groups may attenuate the performance cue bias. First, Wherry and Bartlett's theory of rating states that the extent to which group members possess unique information at the onset of a rating task, a general reduction in bias will occur. Because the performance cue bias has been previously shown to be a small effect, it can be expected that individual members will not be equally affected across all items on a behavioral rating item. Secondly, it has been suggested by the reduction of other well understood biases in social judgment that groups can successfully be used as a tool to reduce such biases (Wright & Christie, 1989; Wright, Christie & Luus, 1990; Wright & Wells, 1985). Also, because a work-behavior rating instrument is perceived as a task with a correct solution to each item, effortful deliberation between group members should occur (Stasser & Stewart, 1992). Finally, research on group memory suggests that groups will have a greater overall amount of disseminated information (i.e. greater memory strength to draw from). Given that the performance cue bias is mediated by a small but systematic shift in decision criterion in the direction of feedback (Martell & Guzzo, 1993), individual group members should be differentially affected by the performance cues given. Group members with varying levels of bias in a high deliberation context should recognize and act to correct individual level bias, thus significantly reducing or perhaps even completely eliminating the performance cue effect.

In this study, research participants were given feedback regarding a work group's performance prior to watching a videotape of the group at work. Subjects were then asked to complete a behavioral rating instrument as individuals or in a 4-person group. It is predicted that individuals will demonstrate the performance cue bias, whereas groups

will not. Specifically, it was hypothesized that:

Hypothesis 1 (Hit rates).

Individuals will demonstrate the performance-cue bias, whereas this bias will be substantially reduced by group raters. That is, individual raters will attribute more behaviors consistent with performance feedback and dismiss behaviors inconsistent with performance feedback, whereas group raters will be less influenced by performance-related information.

Hypothesis 2 (Decision criteria).

Individuals will adopt a more liberal decision criterion for behaviors consistent with feedback and a more conservative decision criterion for behaviors inconsistent with feedback. The decision criterion of groups will be significantly less influenced by the nature of the performance-related information.

Hypothesis 3 (False alarm rates).

Individual raters will attribute to the work group more non-occurring behaviors consistent with feedback and fewer non-occurring behaviors inconsistent with feedback, whereas group raters will be significantly less influenced by the nature of the performance information, and report fewer non-occurring behaviors overall.

Hypothesis 4 (Memory Strength).

Memory strength (Pr) will not differ significantly as a function of performance feedback for either group or individual raters. However, the memory strength of group raters will be greater than the memory strength of individuals overall.

METHOD

Participants

Three hundred and three students enrolled in introductory and lower division Psychology courses at a mid-size rural state University in the Northwest region of the United States participated for extra credit and partial course requirements. Participants were randomly assigned to participate individually (N=103 participants) or in four-member groups (N=50 groups.).

Procedure

Participants watched a 14-minute military training videotape, which depicted five men attempting to build a bridge in an effort to transport themselves and a large box across a pool of water (see Guzzo et al., 1986 for a detailed description of the video). Prior to viewing the videotape, participants were provided with performance-related feedback and instructed to pay careful attention to the work group. Immediately after viewing the tape, participants completed a manipulation check, followed by the behavioral rating instrument, which was completed either (a) as an individual or (b) in a four person group. Pilot testing established thirty-five minutes was sufficient to complete the rating task, and all participants completed the task in the allotted time. All participants viewed the videotape in non-interacting groups to control for any possible effects of social facilitation (Zajonc, 1965).

Independent Variables

Performance Feedback

Participants were informed before observing the videotape that the group's performance in the task was rated by "experts" as being very good (in the top quarter) or very poor (in the bottom quarter) relative to other groups in the contest.

Rater

Participants rated the task-performing group either individually or as a four person group. In the group rating condition, a consensus decision rule was in force such that all group members had to agree on each rating. A member of each group was randomly assigned to record the group's ratings.

Behavior Type

All participants were asked to judge whether effective and ineffective behaviors of the task-performing group did or did not occur (a within-subjects measure).

Dependent Variables

Behavioral Ratings

Work group ratings were made with a 40-item behavioral questionnaire. Participants indicated whether each behavior did or did not occur, using a 6-point scale with endpoints labeled (1) "very certain the behavior did not occur" and (6) "very certain the behavior did occur." Twenty of the 40 items depicted behaviors that did occur in the

videotape. Of these, 11 behaviors were effective and 9 were ineffective. Of the twenty behaviors that did not occur, 11 were effective and 9 were ineffective. (For more on the development and classification of the effective and ineffective behaviors, see Martell and Guzzo, 1991).

RESULTS

Manipulation check

To confirm the effectiveness of the performance expectation manipulation, participants individually evaluated the group's overall performance using a 7-point rating scale, with endpoints ranging from (1) "very poor" to (7) "extremely good." Univariate analysis of variance (ANOVA) revealed that participants provided with positive feedback rated the group's overall performance more favorably (M=5.43) than participants given negative feedback (M=3.42), $F(1,149)=180.71$, $p<.001$, confirming that performance information did affect raters' overall impression of the group's performance.

Behavioral ratings (hit rates)

To determine whether performance feedback influenced participants' responses to effective and ineffective behaviors, the work behavior ratings were translated into hit rates. Hit rates for each behavior type (effective and ineffective) were calculated as the total number of "yes" responses to behaviors that did occur divided by the total number of behaviors (N) that did occur (possible scores range from 0 to 1.0), and represent the probability of responding "yes" to a behavior that was previously observed:

$$\text{Hit Rate}=P(\text{yes/observed behavior}).$$

Overall hit rates for effective and ineffective behaviors were calculated for each rater by counting all responses of 4,5, or 6 (expressing confidence in seeing a behavior that did occur) as a hit or "yes" response. Following the recommendations of Snodgrass and Corwin (1988), all scores were transformed prior to analysis by adding .5 to each score,

and dividing by N+1 to eliminate hit rates of 0 or 1.0. Behavioral ratings (hit rates) were compared using a mixed-factor, 2 X 2 X 2 analysis of variance (ANOVA) to test hypothesis 1. Between-subjects factors were feedback given (positive expectation or negative expectation) and rater type (individual or group). The within-subject factor was behavior type rated (effective or ineffective behavior). The ANOVA table for hit rates appears in Table 1.

Table 1. Analysis of Variance Results for Behavior Ratings (Hit Rates)

Source	df	F	η^2
	Between Subjects		
Performance Expectation (A)	1	3.12	.02
Rater Type (B)	1	.28	.00
(A) X (B)	1	.27	.00
Error	149		
	Within Subjects		
Behavior Type (C)	1	22.45^{***}	.13
(A) X (C)	1	21.04^{***}	.12
(B) X (C)	1	7.22^{**}	.04
(A) X (B) X (C)	1	6.99^{**}	.04
Error	149		

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Hypothesis 1 stated that ratings made by individuals (but not groups) would be affected by performance feedback. A significant three-way interaction of Behavior type and Feedback ($F(1,149)=6.99$, $p<.009$) demonstrates the occurrence of the performance cue bias and thus supports hypothesis 1; behavioral ratings varied systematically in accordance with an implicit theory of performance, and differentially across rater type. To clarify further the nature of the interaction, and to determine whether groups are less biased by performance feedback than individuals, simple effects tests were conducted and, where appropriate, followed-up with cell means comparisons.

Individual raters

Results for individual raters only revealed a significant feedback x behavior type interaction ($F(1,101)=31.01$, $p<.001$). ANOVA revealed that the number of effective and ineffective behaviors identified within the individual rater condition varied greatly as a function of feedback given; hence, a performance cue effect was present in individual raters. Cell means were compared with planned t-tests. A significant difference between feedback conditions across effective behaviors ($t(101)=5.94$, $p<.001$), and a significant difference between feedback conditions across ineffective behaviors ($t(101)=3.15$, $p<.002$) demonstrate that the direction of the relationship is congruent with expectation for individuals, consistent with hypothesis 1. Figure 1 shows the relationship between feedback and behavior type.

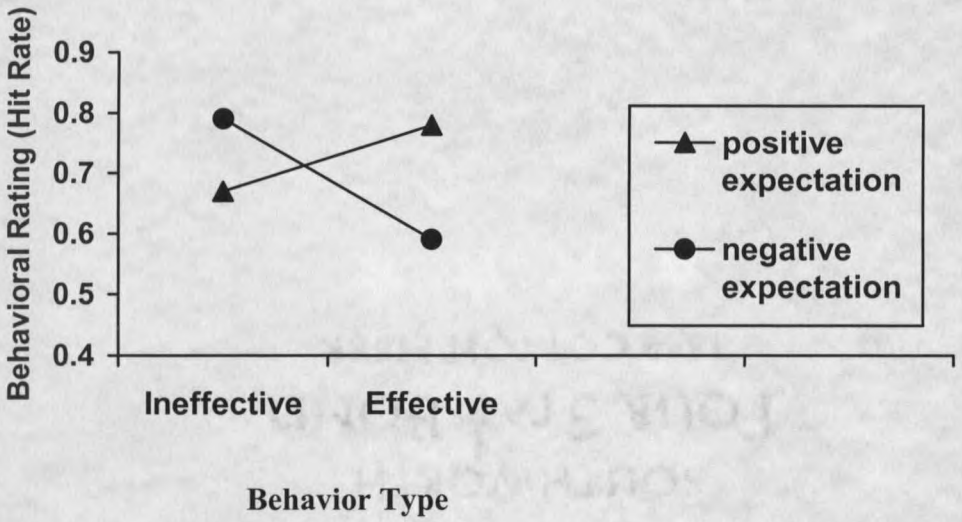


Figure 1. Mean Hit Rate as a function of performance expectation (individual raters only).

Group raters

For group-raters only, results revealed a significant main effect for behavior type only, $F(1,48)=40.69$, $p<.001$, such that overall, more ineffective work behaviors ($M=.79$) were attributed to the group than effective work behaviors ($M=.63$). However, no significant interaction of performance feedback and behavior type on hit rates was found in the group rater condition, $F(1, 48)=2.78$, $p>.10$. Hence, there is compelling evidence that the use of group raters moderates the biasing effect of performance feedback, supporting hypothesis 1.

Mediation: Memory Strength and Decision Criterion

Why might group raters (as opposed to individual raters) be relatively immune to the performance cue bias? An explanation lies within two components of signal detection that drive judgment in recognition memory identified by Snodgrass and Corwin (1988): memory strength (Pr) and decision criterion (Br).

Memory strength (Pr) refers to the overall accuracy of assessment on a behavioral rating task, or the ability to correctly identify behaviors that really did occur, and correctly dismiss those that didn't. Pr can vary from –1.0 (incorrectly dismissing all behaviors which did occur, and incorrectly identifying all behaviors that didn't) to 1.0 (correctly identifying all behaviors which did occur while correctly dismissing all distracter items). The formula for Memory strength is as follows:

Pr = Hit Rate-False Alarm Rate.

A false alarm is defined as a "yes" response to a behavior that did not occur. The false alarm rate (FAR) represents the probability ratio of answering "yes" to an unobserved behavior.

False Alarm rate=P(yes/unobserved behavior).

Similarly to hit rates, false alarm rates for each rater were calculated by considering a response of 4,5, or 6 (answers that indicated confidence in observing the behavior) as a "yes" response for items that did not occur. False alarm rates range from 0 (correctly dismissing all unobserved behaviors) to 1.0 (incorrectly identifying all unobserved behaviors). As with hit rates, a transformation was made prior to analysis to each FAR score by adding a correction of .5 and dividing the total sum by N+1, to eliminate FAR scores of 0 or 1.0 (Snodgrass & Corwin, 1988).

Decision criterion (Br) refers to the response tendency of a rater when deciding

whether or not a behavior was observed. Br ranges from 0.0 (a too-conservative decision criterion, such that one is biased to say a behavior did not occur) to 1.0 (a too-liberal decision criterion, such that one is biased to say a behavior did occur). A Br score of .5 indicates a neutral decision criterion (no response bias).

Br= False Alarm Rate/1 – (Hit Rate – False Alarm Rate)

Both measures of Pr and Br were computed using transformed hit rate and false alarm rate scores, wherein each frequency was added to a correction factor of 0.5 and then divided by N (total number of items)+1, as proposed by Snodgrass and Corwin (1988). Preliminary analyses revealed that Br and Pr scores were uncorrelated for effective (M= +.12) and ineffective(M= -.12) behaviors. Memory strength and decision criterion scores for both group and individual raters are presented in Table 2.

Table 2. Mean Ratings of Work Behavior (Hit Rates), False Alarm Rates, Memory Sensitivity (Pr), and Response Bias (Br)

	Effective Work Behavior				Ineffective Work Behavior			
	Hit Rates[a]	False Alarm Rates[b]	Memory Strength[c]	Response Bias[d]	Hit Rates[a]	False-Alarm Rates[b]	Memory Strength[c]	Response Bias[d]
Individual ratings								
Positive expectation	.78	.32	.47	.59	.67	.28	.39	.48
(n=65)	(.15)	(.20)	(.20)	(.25)	(.19)	(.15)	(.17)	(.23)
Negative expectation	.59	.22	.37	.32	.79	.43	.36	.67
(n=38)	(.18)	(.25)	(.27)	(.23)	(.17)	(.20)	(.20)	(.22)
Group ratings								
Positive expectation	.67	.14	.53	.29	.79	.26	.53	.57
(n=25)	(.12)	(.01)	(.12)	(.16)	(.17)	(.14)	(.20)	(.23)
Negative expectation	.61	.09	.52	.18	.81	.26	.54	.58
(n=25)	(.12)	(.06)	(.12)	(.12)	(.11)	(.12)	(.19)	(.17)

Note. Standard deviations appear in parantheses.

[a] Mean values range from 0 (no observed behaviors reported) to 1.0 (all observed behaviors reported).

[b] Mean values range from 0 (no unobserved behaviors reported) to 1.0 (all unobserved behaviors reported).

[c] Pr values range from -1.0 (no memory) to +1.0 (perfect memory).

[d] Br > .50 indicates a liberal decision criterion; Br <.50 indicates a conservative decision criterion.

Decision Criterion (Br)

A 2 (rater type) x 2 (feedback) x 2 (behavior type) mixed factor ANOVA was conducted for decision criterion scores, to determine if performance information biased the decision criterion of raters. Overall ANOVA results for Decision criterion (Br) are outlined in Table 3.

Table 3. Analysis of Variance Results of Decision Criterion (Br)

Source	df	F	η^2
	Between Subjects		
Performance Expectation (A)	1	4.07^{*}	.02
Rater (B)	1	22.17^{***}	.13
(A) X (B)	1	.07	.00
Error	149		
	Within Subjects		
Behavior Type (C)	1	57.31^{***}	.27
(A) X (C)	1	22.46^{***}	.08
(B) X (C)	1	13.40^{***}	.13
(A) X (B) X (C)	1	7.98^{**}	.05
Error	149		

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

A three way feedback x behavior type x rater interaction suggests that decision criterion towards each behavior type (effective and ineffective) was significantly affected by the nature of the feedback given. Because of the occurrence of a three-way interaction, it is necessary to tease out the nature of the two-way relationship between feedback and behavior type within individual and group rater conditions. Separate simple effects tests of group and individual raters were conducted to determine the nature of the relationship between expectation and behavior type within rater conditions, as well as relevant t-tests to examine the effect of performance feedback on Br for each behavior type at the cell mean level. Results (below) support hypothesis 2.

Individual raters

A significant two-way interaction of behavior type and feedback in the individual condition demonstrates a fluid decision criterion among individuals, which is directed by congruence of the behavior rated with the feedback given ($F(1,101)=35.07$, $p<.001$). Cell means were compared with planned t-tests. As with hit rates, significant differences between feedback conditions across effective and ineffective behaviors demonstrates the probable activation of an implicit theory of performance; those with a positive expectation observed more effective behaviors than those with a negative expectation ($t(101)=5.50$, $p<.001$), and raters with a negative expectation saw significantly more ineffective behaviors than raters with a positive expectation ($t(101)=4.10$, $p<.001$). The relationship between feedback and decision criterion for behavior type in individual raters is detailed in Figure 2.

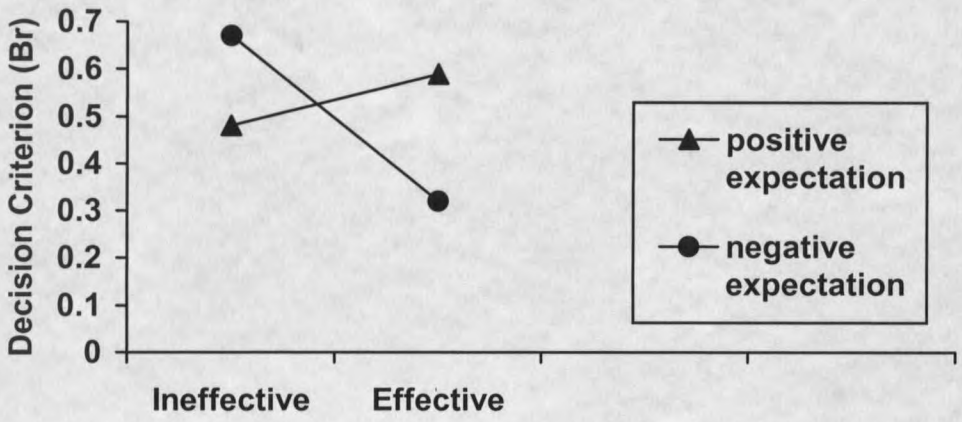


Figure 2. Mean Decision Criterion (Br) as a function of performance expectation (individual raters only).

The congruence of performance feedback and the respective decision criterion for effective and ineffective behaviors suggests that the decision criterion of individual raters is systematically skewed in the direction of expectation; the performance cue effect is likely mediated at the individual rater level by a systematic shift in decision criterion.

Group Raters

A 2 (feedback) X 2 (behavior type) mixed factorial comparison was conducted within the group rater condition. Only a significant main effect for behavior type was detected, $F(1,48)=79.99$, $p<.001$, suggesting that groups adopted a more liberal decision criterion when rating ineffective ($M=.58$) versus effective ($M=.24$) behaviors. No significant interaction of performance feedback and behavior type was found, suggesting

that groups maintained a relatively stable decision criterion across levels of behavior type, despite expectation ($F(1,49)=2.32$, $p>.134$). Thus, the decision criterion of groups in work behavioral ratings appears to remain relatively immune from the effects of performance expectation. Results for Br within the group rating condition closely parallel those for hit rates, bolstering support for the mediation of the performance cue bias in work behavioral ratings through the nature of the decision criterion employed. In summary, hypothesis 2 was supported.

Memory Strength (Pr)

A 2 (rater type) x 2 (feedback) x 2 (behavior type) repeated measures ANOVA was conducted to determine if memory strength (Pr) varied similarly to hit rates, which would suggest memory strength as an element of the performance cue effect in behavioral ratings.

It was hypothesized that groups (through greater combined attention and storage resources) would benefit from greater overall memory strength. Results revealed only a significant effect for rater, $F(1,149)=30.22$, $p<.001$, supporting Hypothesis 4, in that groups had greater memory strength ($M=.53$) than did individuals ($M=.40$). A lack of significant interaction at the three and two-way levels further supports the Martell and Guzzo (1991) model, in that differences in hit rates are not mediated by memory strength ($F(1,149)=.153$, $p> .70$). The range of mean Pr scores across conditions (.36 to .54) suggests that the behavioral rating task was sufficiently difficult (no scores approached the ceiling of 1.0) and still sufficiently fair (no scores were below 0, or the threshold of "no memory"). Thus, all raters were able to correctly differentiate a reasonable proportion of behaviors that did occur from those that did not, suggesting that raters were sufficiently engaged in the task. No significant correlation of memory strength and

decision criteria for effective (r = -.122, p>.1) and ineffective (r = .121, p>.1) behaviors was found. Hypothesis 4 was confirmed. Overall ANOVA results for memory strength are shown in Table 4.

Table 4. Analysis of Variance Results of Memory Strength (Pr)

Source	df	F	η^2
	Between Subjects		
Performance Expectation (A)	1	1.43	.01
Rater (B)	1	30.22***	.16
(A) X (B)	1	1.77	.01
Error	149		
	Within Subjects		
Behavior Type (C)	1	.67	.00
(A) X (C)	1	1.26	.00
(B) X (C)	1	1.30	.00
(A) X (B) X (C)	1	.15	.00
Error	149		

Note. * p <.05, ** p<.01, *** p<.001.

False Alarm Rates

Hypothesis 3 addressed the expectation that individuals (but not groups) would falsely identify behaviors that did not occur in congruence with performance information given, and that group raters would make fewer overall errors of commission than individuals. An ANOVA on false-alarm rates revealed a significant main effect for rater type, demonstrating an across-the-board difference in the false-alarm rates of group and individual raters ($F(1,149)=36.642$, $p<.001$), with groups generally making significantly fewer false alarm errors ($M=.11$) than individuals ($M=.28$). A three-way interaction of feedback x behavior type x rater ($F(1,149)= 4.87$, $p<.03$) was also significant. Suggested is that false alarm errors were consistently affected in the direction of expectation/feedback given, but differentially so across rater type, demonstrating the influence of performance feedback on behaviors falsely reported within individual raters. ANOVA results are presented in Table 5.

Table 5. Analysis of Variance Results of False Alarm Rates

Source	df	F	η^2
	Between Subjects		
Performance Expectation (A)	1	.00	.00
Rater (B)	1	36.64***	.19
(A) X (B)	1	1.28	.00
Error	149		
	Within Subjects		
Behavior Type (C)	1	30.76***	.17
(A) X (C)	1	11.28***	.07
(B) X (C)	1	2.14	.01
(A) X (B) X (C)	1	4.87*	.03
Error	149		

Note. *p <.05, **p<.01, ***p<.001.

To examine further the nature of the three-way interaction, and to determine if groups are significantly less influenced by performance feedback than individuals, subsequent simple effects tests and, where appropriate, cell-wise comparisons were performed at the two-way rater level. As described below, hypothesis 3 was supported.