Using a source monitoring technique to reduce the effects of performance expectations on work behavior ratings
by David Paul Evans

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 examined the use of a source monitoring technique in a behavioral rating task to determine whether this would reduce the performance cue bias. Two . hundred and twenty-three participants were given positive or negative feedback regarding the performance of a work group, and following observation of the work group, were randomly placed in either a source monitoring training group or a control group with no instruction. Afterwards, both groups completed a work-behavior questionnaire. As was predicted, the control group (but not the source monitoring 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 the control group (but not the source monitoring group) raters were found to be systematically biased by performance information as well. Implications for performance appraisal theory, research, and practice are discussed.

USING A SOURCE MONITORING TECHNIQUE TO REDUCE THE EFFECTS OF PERFORMANCE EXPECTATIONS ON WORK BEHAVIOR RATINGS

by

David Paul Evans

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

of

Master of Science

In

Applied Psychology

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 2002

APPROVAL

Of a thesis submitted by

David Paul Evans

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 _____ 7-22-02
(Signature) (Date)

Approved for the Department of Psychology

A. Michael Babcock _____ 7/22/02
(Signature) (Date)

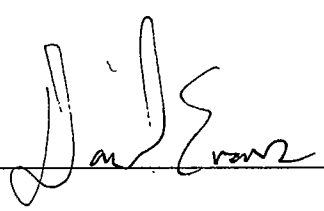
Approved for the College of Graduate Studies

Bruce R. Mcleod _____ 7-22-02
(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 7-22-02

ACKNOWLEDGEMENTS

I would like to thank Dr. Richard Martell for his constant input and guidance throughout the course of this project, my committee members Dr. Chuck Pierce and Dr. Dan Moshavi for their time and feedback, and finally my friends and family 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 examined the use of a source monitoring technique in a behavioral rating task to determine whether this would reduce the performance cue bias. Two hundred and twenty-three participants were given positive or negative feedback regarding the performance of a work group, and following observation of the work group, were randomly placed in either a source monitoring training group or a control group with no instruction. Afterwards, both groups completed a work-behavior questionnaire. As was predicted, the control group (but not the source monitoring 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 the control group (but not the source monitoring group) raters were found to be systematically biased by performance information as well. Implications for performance appraisal theory, research, and practice are discussed.

INTRODUCTION

Implicit Theories of Performance and the "Performance Cue Bias"

Understanding the nature of biases that can arise in work performance evaluations is of importance to researchers and practitioners, alike. One source of bias, the performance cue bias, has been the subject of numerous investigations (e.g., Martell & Willis, 1993; Martell & Guzzo, 1991; Staw, 1975). The performance cue bias was first identified by Staw (1975) in a study in which he led individuals to believe that their work group had performed "very well" or "very poorly" on a business simulation exercise. Group process ratings and self-evaluations collected afterwards 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. This finding suggests that an individual's knowledge of group performance may serve as a cue, causing individuals to ascribe cue-consistent attributes to the target of their evaluation.

More recent research on the performance cue bias demonstrates that rater expectations can introduce bias into work behavior ratings (Martell & Guzzo, 1991; Martell & Willis, 1993; Martell, Guzzo, & Willis, 1995; Martell & Leavitt, in press). Specifically, observers led to believe that a group did quite well (versus quite poorly) attributed more effective (and fewer ineffective) behaviors to this group. In organizational settings, these cues are often readily available and can consist of performance Knowledge taken from a variety of sources, such as references, employee

resumes, and prior interviews or performance appraisals. Ultimately, the performance-cue bias is problematic in two specific ways. First, as Staw (1975) pointed out, "significant correlations between performance and self-report data may only be reflecting the respondents 'theories' of organizational performance rather than actual events" (p.417). That is, 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." Also, as discussed by Murphy (1991) and Lord (1985), bias in behavioral ratings is especially troublesome, given the importance of providing accurate performance ratings feedback for training, employee evaluations, and assessment center judgments.

Anatomy of A Bias

Establishing a viable explanation for why bias is introduced in work behavior ratings has been the focus of several studies (Baltes & Parker, 2000; Martell & Guzzo, 1991; Martell & Willis, 1993; Martell & Leavitt, in press). Performance on a behavioral rating instrument is fully captured by two components (Snodgrass & Corwin, 1988) – memory strength and decision criterion. Memory strength is the amount of information stored in memory and is assessed by looking at an individual's ability to distinguish between behaviors they have observed and those they have not. Decision criterion provides a measure of the individuals decision criteria in force when deciding whether or not a behavior has occurred. Raters may adopt a decision criterion that is either too liberal (i.e., a bias to say a behavior did occur), too conservative (i.e., a bias to say a

behavior did not occur), or neutral (i.e., equal likelihood of saying the behavior did or did not occur).

Past research has found that performance cue bias in behavioral ratings is due to a systematic response bias in raters' decision criterion, such that a more-liberal decision criteria is adopted for behavior consistent with a rater's performance expectation. There is no evidence that memory strength plays a role (Martell & Guzzo, 1991; Martell & Willis, 1993). Rater's decision criteria is generally thought to be guided by the perceived familiarity of the behavior (Whittlesea, 1993). That is, performance expectations trigger a response bias in which raters rely on feelings of familiarity to determine whether consistent versus inconsistent behaviors were observed. A greater feeling of familiarity for expectancy-consistent behaviors (and a lower feeling of familiarity for inconsistent behaviors) causes raters to adopt a too-liberal decision criteria when judging consistent behavior and a too-conservative decision criteria for inconsistent behaviors. Thus, raters' decision criteria and it's familiarity component are key to understanding the impact of performance-based expectations and, importantly, how the performance-cue bias in work behavior ratings may be reduced.

Source Monitoring Training

Current research on source monitoring has shown some promising avenues for better understanding the ways in which individuals recollect behaviors from their past. Source monitoring refers to the set of processes involved in making attributions about the origin (or source) of memories. According to the Source Monitoring Framework (SMF)

outlined by Johnson, Hashtroudi, & Lindsay (1993), later activation of previously formed memory representations (along with activation of other information at test) results in mental experiences that range from general feelings of familiarity to concrete memories for specific features of prior events such as their perceptual or contextual detail. Thus, from a source monitoring perspective, individuals may misinterpret the initial source of a memory. For example, Dodson and Johnson (1993) observed more source misattribution errors when participants made judgments on the basis of feelings of familiarity rather than on memory for the specific details of the prior stimulus event. Indeed, Lindsay and Johnson (1989) propose that heuristic processes similar to decision criteria (e.g. heuristic judgments involved criteria such as "if familiarity level is above X, the event probably happened," or "if the amount of perceptual detail exceeds X, the event was probably perceived") are employed when making social judgments. If so, relying on feelings of familiarity rather than a concrete memorial representation of a prior stimulus event may introduce bias in judgment (see Whittlesea, 1993; Whittlesea, Jacoby, & Girard, 1990).

Past research investigating familiarity in recognition memory (Tulving, 1985; Mandler, 1980; Gardiner, 1988; Rajaram, 1993; Roediger, 1996) points to two different ways in which an individual can recognize a past event. First, there is there is Remembering that something has been seen before. A Remember judgment occurs when an individual is certain of seeing something before due to an ability to specifically remember the source of the memory and mentally relive it, in essence pinpointing where it came from. It might include memories of what an individual was thinking about at the time they saw it, such as an association or image formed, something of personal

significance one was reminded of, or something about the physical aspects of the behavior. Second, there is Knowing that something has been seen before. A Know judgment occurs when an individual is certain of seeing something before because of strong feelings of familiarity that are accompanied by the stimulus event. There is the absence of conscious recollection so nothing comes to mind when trying to remember the source of the memory, you just know it happened. For example, if a professor asked his colleague about a conference he had attended while in graduate school, he might accurately recollect that he did indeed drive to the conference, but without any details of the trip—the weather, traffic conditions, or if there were any travel companions. This is the essence of a Know judgment.

On The Status of Know Versus Remember Judgments

As noted by Rajaram (1993), Remember responses are akin to measures of actual recollection, whereas Know responses are measures of familiarity. The distinction between these two bases of recognition memory is supported by an abundance of research. A series of experiments reported by Gardiner and his colleagues (1988; Gardiner & Java, 1990; Gardiner & Parkin, 1990) have identified numerous factors (viz., levels of processing, age, word stem/fragment completion, and generate vs. read) that produce dissociations between Remember and Know responses. Accordingly, it is widely accepted that Remember and Know judgments are exclusive of one another; that is, recognition of an item can only be based on either a "Remember" or "Know" response, not both (Gardiner & Parkin, 1990; Jacoby & Whitehouse, 1989).

It is important to note that Remember and Know do not differ in levels of confidence (Rush & Beauvais, 1981; Gardiner & Java, 1990; Rajaram, 1993; Parkin & Walter, 1992). That is, an individual's confidence does not differ from a Remember versus a Know response. For example, Gardiner and Java (1990) and Rajaram (1993) substituted "sure" and "unsure" confidence judgments for "Remember" and "Know" responses and found that both levels of an independent variable, which had previously affected "Know" and "Remember" responses, did not affect levels of "unsure" and "sure" judgments. Findings by Parkin and Walter (1992) further validated the notion that Know and Remember responses do not differ in confidence levels. They found that older adults did not register a distinction between confidence judgments when compared to younger adults even though they had previously done so with Remember and Know judgments. Finally, Rush and Beauvais (1981) demonstrated the robustness of this difference by finding out that when they encouraged participants to rate only the items that they felt confident in providing an accurate rating for, they still obtained a performance-cue bias.

Hypotheses

As described earlier, response bias can be seen as a mechanism by which raters rely on feelings of familiarity (Know judgments) to determine if a behavior has been observed previously. Accordingly, there is greater familiarity associated with behaviors that are consistent with an individuals expectations or implicit theories of effective work group behaviors, which can therefore lead to systematic decisional biases. A logical inference from this research is that by controlling the bases of raters' behavior recognition ratings

by informing them to only attribute Remember (and not Know) judgments to the group, we can significantly reduce the performance-cue bias. In other words, by eliminating the use of Know judgments and having raters ascribe behaviors to the group that they have only specific recollections for, the familiarity based response bias will be no longer exert the influence it once did. This use of source monitoring could be a powerful force in helping to reduce the bias that is exhibited in an individual's evaluation of another's performance and could provoke individuals to think about their own memory in a way that they might not normally be accustomed to.

Accordingly, this article will report the results of a study aimed at addressing two key questions: Can raters learn to distinguish between Know and Remember judgments in work behavior ratings? And, if so, can the biasing effects of performance-based expectations on behavioral ratings be reduced by instructing raters to report only behaviors that are remembered (and not behaviors that are simply familiar)? Thus, the following hypotheses will be tested.

Hypothesis 1. (Hit rates).

Source monitoring training will significantly reduce the performance-cue bias in work behavior ratings. Specifically, when provided with positive versus negative performance-related information, raters will attribute more effective and fewer ineffective behaviors to the work group, whereas trained-raters will be significantly less influenced by the nature of the performance information.

Hypothesis 2. (Decision criteria).

When provided with positive versus negative performance-related information, raters will adopt a more liberal decision criteria for effective behaviors and a more conservative decision criteria for ineffective behaviors, whereas the decision criteria adopted by trained raters will be significantly less influenced by the nature of the performance information.

Hypothesis 3. (False alarm rates).

When provided with positive versus negative performance-related information, raters will attribute to the work group more effective and fewer ineffective behaviors that did not occur, whereas the false alarm rates of trained raters will be significantly less influenced by the nature of the performance information.

Hypothesis 4. (Memory strength).

When provided with positive versus negative performance-related information, no significant differences in memory strength will be observed in non-trained raters versus trained-raters.

METHOD

Participants

Two hundred and twenty-three undergraduate students enrolled in Introductory Psychology classes at a mid-size state University in the Northern Rockies participated for extra credit and partial course requirements in return for course credit.

Design and Procedure

The design of this study was a 2 x 2 x 2 mixed factorial, with performance expectation (positive or negative) and source monitoring training (R/K, none) as between-subjects factors and behavioral type (effective, ineffective) as a within-subjects factor. Participants were randomly assigned to conditions. Informed consent was given before the start of the experiment and a debriefing form will be passed out and read aloud at the conclusion of the study.

Participants watched a 14-minute videotape that depicts a group of five men attempting to build a bridge out of ropes and planks so they can transport themselves and a box across a body of water. Before observing the group, participants were given positive or negative information regarding the group's performance and instructed to pay careful attention to the work group. At the conclusion of the videotape, participants completed a manipulation check and were then randomly assigned to a source monitoring training or no training control group.

Source Monitoring Training Group

The experimental group were instructed on how to distinguish between a Know and a Remember judgment (i.e. w/ Remember judgments, we confidently recognize something because we are able to "bring it back to mind" and actually picture the source of the memory. At other times, we confidently recognize something on the basis of its familiarity. That is, although we are unable to "bring it back to mind" and actually picture it, we are still sure that it occurred because of the strong feelings of familiarity associated with the event. This would be a Know judgment). Examples were given to help clarify the differences between the two. (i.e. Imagine, for example, seeing someone on the street whom you had met last week at a party. You are aware of having a mental image of meeting the person that you can now "see" in your mind. This might include, for example, your memory of talking to the person, what you were thinking about at the time you met the person or where the two of you were standing at the party.) Information was also provided to ensure that the differences in Know and Remember judgments would not be confused with levels of confidence. After completion of the questionnaire, participants filled out a short test to ensure their understanding of the differences between Know and Remember judgments.

No Training Control Group

The control group did not receive these instructions. Instead, they were given the task of free-writing on the topic of "Why We Work" for exactly four and a half minutes (the same amount of time that the R/K training took to complete). This task was chosen

because of it's consistency with the overall theme of the study and to ensure that any latency effects between the conditions would remain equal.

Independent Variables

Performance Expectation

Participants were told that the group had performed quite good or quite poor. The performance information was given prior to observing the group. Specific instructions were as follows: "You might like to know that the group you are about to view was rated by a panel of experts specially trained in the observation of group performance. These experts rated the group on several aspects of group performance, including such things as the time it took to cross the bridge, the number of participants to reach the other side, and the time to get the box across. Compared with other groups performing this task, the group you are about to observe was judged to have worked together exceptionally well (poorly) given the circumstances they had to deal with. They were judged to be in the 90^{th} (10^{th}) percentile compared to all other groups which indicates that their performance was ranked better (worse) than 90 (10) out of 100 groups. In other words, the performance of this group was quite good (poor) in comparison with other groups."

Training

To reduce the biasing effects of the performance feedback, groups randomly received either Know/Remember training or a control group task to fill the same duration of time as the training.

Behavior Type

All participants were asked to judge whether effective and ineffective behaviors of the task-performing group did or did not occur.

Dependent Variables

Behavior rating instrument

All participants were presented with a behavioral rating instrument compromised of 20 behaviors (11 effective, 9 ineffective) that appear in the video. In addition, there were 20 behaviors (11 effective, 9 ineffective) included that were not depicted in the video. Some sample behavioral items are as follows: "...a group member carefully explained to the others how to best position their feet on the bridge to avoid falling" (effective, did occur); "Each member volunteered to perform a specific task in order to build the bridge in an orderly fashion" (effective, did not occur). For the control group, a behavior checklist was used with participants either circling a "Y" indicating that the behavior occurred or a "N" indicating that the given behavior did not occur. For the experimental group, a behavior checklist was used with the instructions of only circling "Y" when the recollection of that behavior fits a Remember judgement.

Participants' behavior ratings were first transformed into hit rates and false-alarm rates. A hit is defined as a "yes" response to a behavior that occurred in the videotape and the hit rate, HR, is the conditional probability of responding "yes" to a present behavior:

$$\text{Hit rate} = P(\text{yes}/\text{present behavior}).$$

Overall hit rates were calculated for each participant by treating correctly identified effective and ineffective behaviors that were present as hits. The overall hit rate ranges from 0.0 to 1.0, with higher values indicating a greater percentage of correctly identified behaviors. A false alarm is defined as a "yes" response to a behavior that was not present and the false alarm rate, FAR, is the conditional probability of responding "yes" to a not present behavior.

False alarm rate = P(yes/not present behavior).

Accordingly, overall false alarm rates were calculated by treating falsely identified effective and ineffective behaviors that were not present as false alarms. The overall false alarm rate ranges 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), hit rates were transformed 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. False alarm rates were transformed 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.

Next, measures of decision criteria (Br) and memory strength (Pr) were calculated. Br ranges from 0.0 to 1.0. A Br of .50 indicates that raters are relying on a neutral decision criterion when judging whether a behavior was observed, whereas scores greater than .50 indicate a too-liberal decision criterion and scores less than .50 indicate a too-conservative decision criterion. Br is computed as follows:

Br = False alarm rate/1-(Hit rate – False Alarm rate).

Pr is a measure of memory strength and ranges from –1.0 (no memory) to +1.0 (perfect memory). It is computed as follows:

$$Pr = HR - FAR.$$

The overall hit and false alarm rates of individuals and groups were transformed into measures of decision criteria and memory strength.

Manipulation checks. A manipulation check was used to determine whether the participants actually internalized the performance feedback given to them. In addition, a 6-item test of Know or Remember judgments was administered to the experimental training group to ensure that the participants understood the distinction between the two bases of recognition memory judgment.

RESULTS

Manipulation Checks

To confirm the effectiveness of the performance expectation manipulation, participants evaluated the group's performance using a 5-point rating scale with endpoints labeled (1) "extremely good" to (5) "extremely poor." Univariate analysis of variance (ANOVA) revealed a significant main effect for performance expectation, $F(1,222) = 273.16$, $p < .001$, which confirmed participants provided with positive performance information rated the group's overall performance more favorably (M = 2.80) than participants given negative performance information ($M = 4.00$). In addition, although prior pilot testing ensured that participants sufficiently understood the differences between "Remember" and "Know", we thought it beneficial to determine the extent to which study participants in the reality-monitoring conditions succeeded in learning the key features of the training program. The results of a 6-item test of Know versus Remember judgments revealed that participants held a very good level of understanding. The overall mean score of 5.23 (out of 6.00) points provides evidence of participants' competency to correctly distinguish between the two bases of recognition memory judgments.

Next, a series of repeated measures ANOVAs were conducted, with performance expectation and rater training as between-subjects factors and behavior type as the within-subject factor. Hypotheses 1-3 predicted significant three-way interactions. Accordingly, to interpret the nature of the predicted interactions, planned cell-wise

comparisons (using Bonferroni tests at $p < .0125$ for each of the four contrasts which ensures an overall error rate of .05) were conducted and the resulting effect-size estimates were examined. Means and standard deviations for hit rates, decision criteria, false alarm rates and memory strength appear in Table 1. Complete ANOVA results are presented in Table 2.

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

	Effective Work Behavior				Ineffective Work Behavior			
	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
No Training								
Positive expectation (n=52)	.68 (.17)	.45 (.21)	.23 (.12)	.45 (.18)	.57 (.20)	.37 (.19)	.22 (.10)	.34 (.20)
Negative expectation (n=60)	.45 (.15)	.15 (.14)	.09 (.08)	.36 (.12)	.72 (.20)	.62 (.22)	.41 (.16)	.31 (.20)
Reality Monitoring Training								
Positive expectation (n=55)	.54 (.15)	.21 (.17)	.11 (.09)	.43 (.13)	.48 (.17)	.27 (.17)	.19 (.13)	.28 (.17)
Negative expectation (n=57)	.44 (.15)	.14 (.12)	.08 (.07)	.36 (.13)	.47 (.20)	.31 (.19)	.23 (.14)	.24 (.20)

Note. Standard deviations appear in parentheses.
[a] Mean values range from 0 (no "present" behaviors reported) to 1.0 (all "present" behaviors reported).
[b] Br > .50 indicates a liberal decision criterion; Br <.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] Pr values range from -1.0 (no memory) to +1.0 (perfect memory).

Table 2.
Analysis of Variance Results of Behavior Ratings

		Hit Rates	Decision Criteria	False Alarm Rate	Memory Strength
Source	df	F	F	F	F
	Between Subjects				
Performance Expectation (A)	1	7.25**	1.13	.75	10.92***
Rater Training (B)	1	58.06***	91.69***	50.89***	5.15*
(A) X (B)	1	.29	.08	.65	.90
	Within Subjects				
Behavior Type (C)	1	2.95	74.06***	135.84***	37.57***
(A) X (C)	1	41.19***	82.19***	74.96***	1.85
(B) X (C)	1	7.32**	4.90*	3.49	2.16
(A) X (B) X (C)	1	14.31***	35.77***	33.37***	.19
Error	220				

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

Behavioral ratings (hit rates)

The 2 x 2 x 2 repeated measures ANOVA produced a performance expectation x behavior type interaction, $F(1,220) = 41.19$, $p < .001$, which demonstrates the effects of the performance-cue bias. Also, as predicted, this analysis included a performance expectation x behavior type x rater-training interaction, $F(1,220) = 14.31$, $p < .001$. The nature of the three-way interaction, depicted in Figure 1, provides support for hypothesis 1. Specifically, when given positive versus negative performance-related information, untrained raters ascribed significantly more effective behaviors, $t(110) = 6.08$, $\eta^2 = .25$, and significantly fewer ineffective behaviors, $t(110) = 3.96$, $\eta^2 = .125$, to the group. These findings replicate previous research (Martell & Guzzo, 1991; Martell & Guzzo, 1995; Martell & Willis, 1993). In contrast, performance feedback had no significant effects on the hit rates of participants in the source monitoring condition. In this condition, the effects of performance-related information were significantly reduced for effective behavior ratings, $t(110) = 2.65$, $\eta^2 = .059$, and eliminated for ineffective behavior ratings, $t(110) = 0.26$, $\eta^2 = .00$.

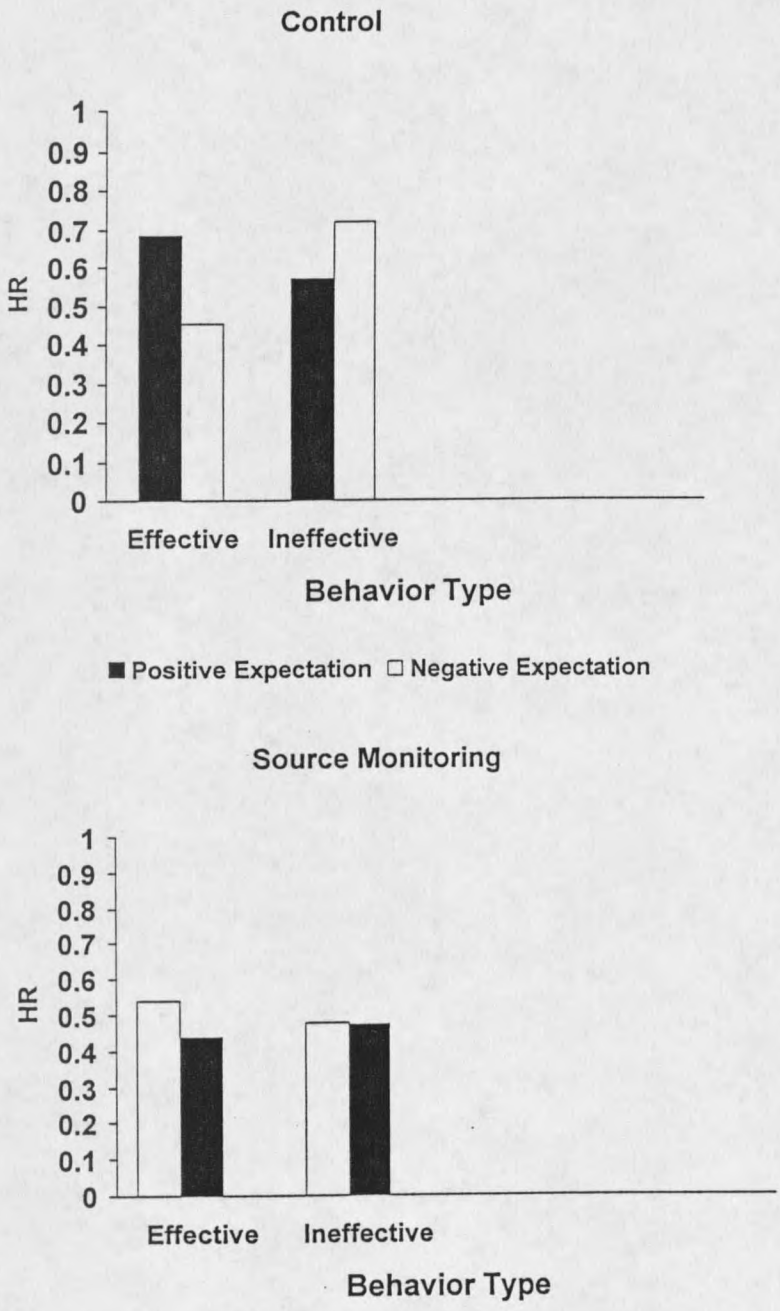


Figure 1: Hit rates as a function of performance expectations, behavior type, and instructional set (source-monitoring vs. control condition).

Decision Criteria (Br)

The 2 x 2 x 2 repeated measures ANOVA produced a performance expectation x behavior type interaction, $F(1,220) = 82.19$, $p < .001$, and, as predicted, a performance expectation x behavior type x rater-training interaction, $F(1,220) = 35.77$, $p < .001$. The nature of the three-way interaction, depicted in Figure 2, provides support for hypothesis 2. Specifically, when given positive versus negative performance-related information, untrained raters adopted a significantly more liberal decision criteria when rating effective behaviors, $t(110) = 7.93$, $\eta^2 = .36$, and a significantly more conservative decision criteria when judging ineffective behaviors $t(110) = 6.61$, $\eta^2 = .28$. These findings replicate previous research (Martell & Guzzo, 1991; Martell & Guzzo, 1995; Martell & Willis, 1993). In contrast, performance feedback had no significant effects on the decision criteria adopted by raters in the source monitoring conditions; neither the effective behavior ratings, $t(110) = 1.85$, $\eta^2 = .03$, nor the ineffective behavior ratings, $t(110) = 1.06$, $\eta^2 = .01$, were significant.

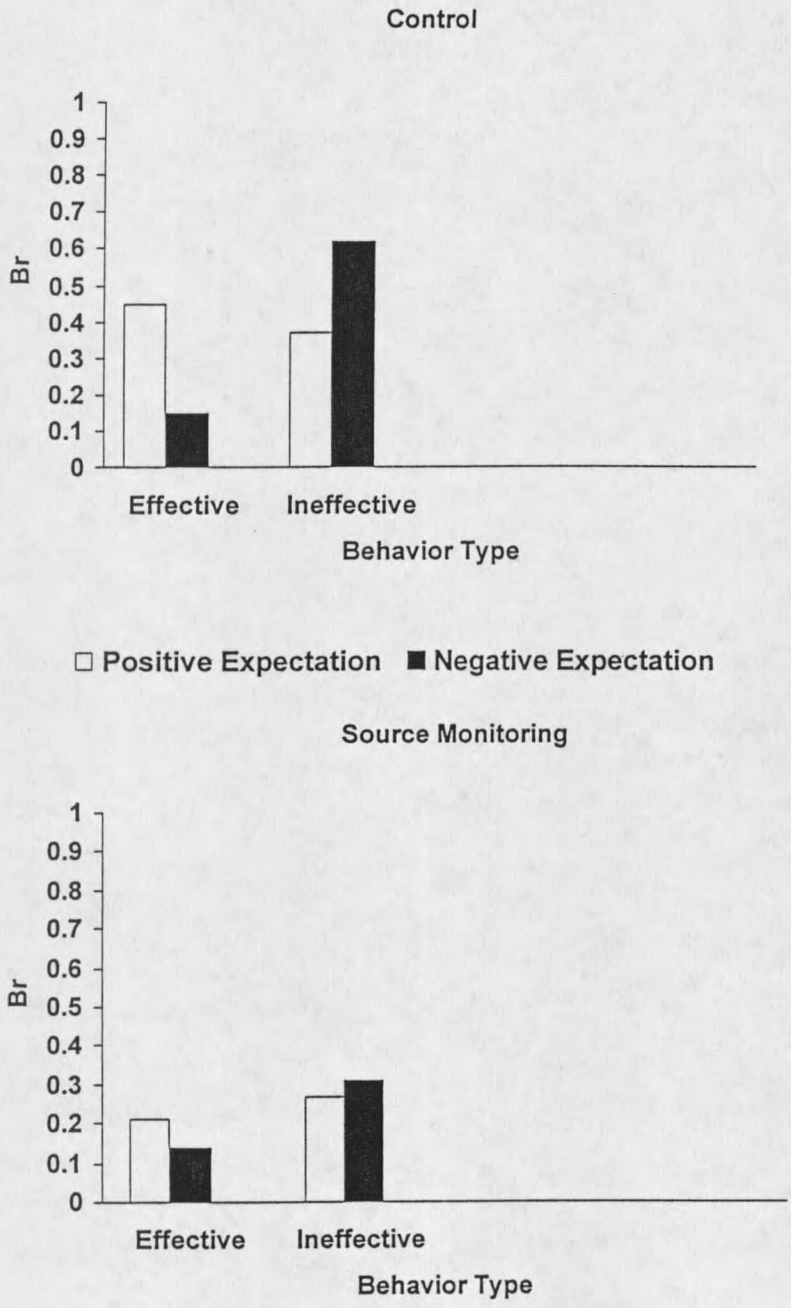


Figure 2: Response bias as a function of performance expectations, behavior type, and instructional set (source-monitoring vs. control condition).

False Alarm Rates

The 2 x 2 x 2 repeated measures ANOVA produced a performance expectation x behavior type interaction, $F(1,218) = 74.97$, $p < .001$, and as predicted, a performance expectation x behavior type x rater-training interaction, $F(1,218) = 33.38$, $p < .001$. The nature of the three-way interaction, depicted in Figure 3, provides support for hypothesis 3. Specifically, when given positive versus negative performance-related information, untrained raters made significantly more false alarm errors for effective behaviors $t(110) = 6.36$, $\eta^2 = .26$, than ineffective behaviors $t(110) = 8.63$, $\eta^2 = .40$. These findings replicate previous research (Martell & Guzzo, 1991; Martell & Guzzo, 1995; Martell & Willis, 1993). In contrast, performance feedback had no significant effects on false alarm errors for raters in the source monitoring conditions, neither the effective behaviors ratings, $t(110) = 1.36$, $\eta^2 = .016$, nor ineffective behaviors ratings, $t(110) = 1.81$, $\eta^2 = .028$ were significant.

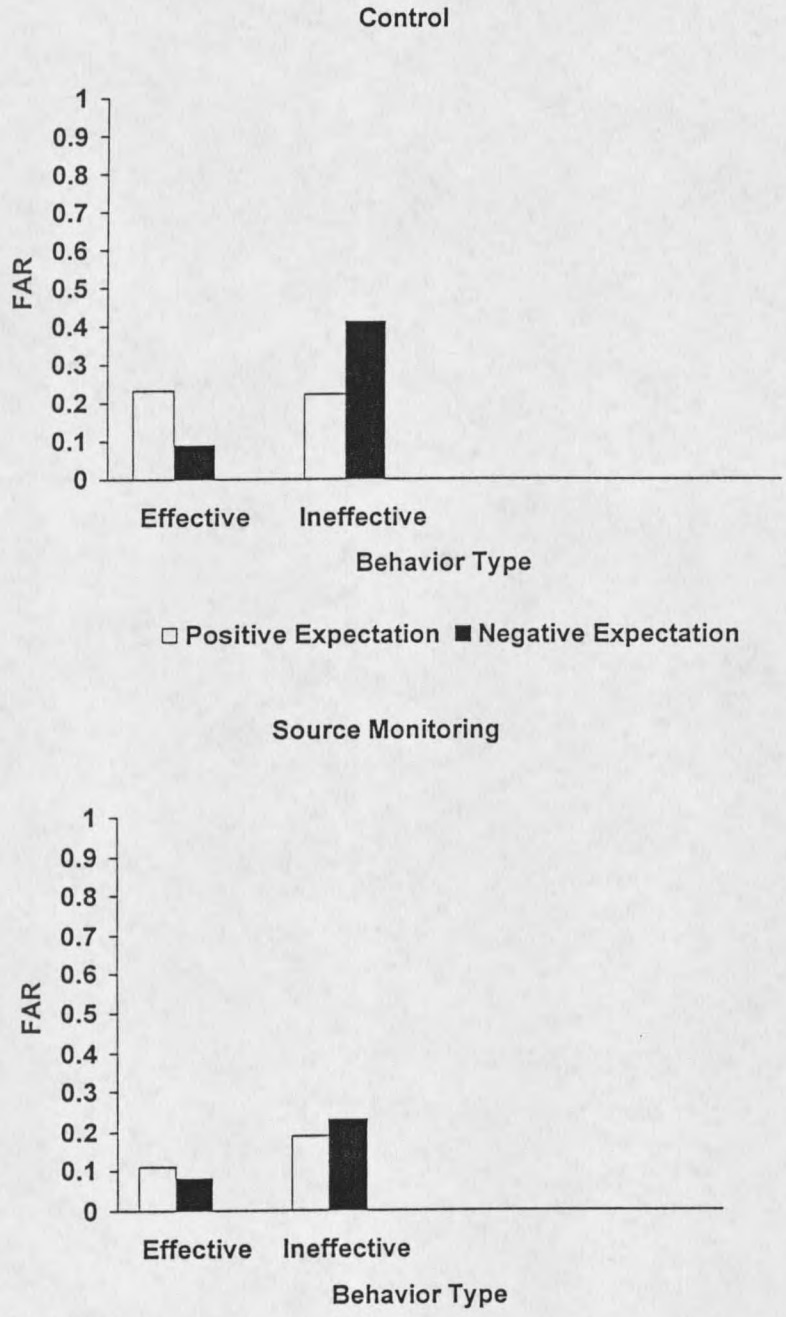


Figure 3: False alarm rates as a function of performance expectations, behavior type, and instructional set (source-monitoring vs. control condition).

Memory Sensitivity (Pr)

As predicted, the 2 x 2 x 2 repeated measures ANOVA did not produce a performance expectation x behavior type interaction, $F(1,220) = 1.85$, $p = .175$, and, as predicted, did not produce a performance expectation x behavior type x rater-training interaction, $F(1,220) = .190$, $p = .663$. The absence of this provides support for hypothesis 4 and further supports the Martell and Guzzo (1991) model which states that observed differences in hit rates are not mediated by memory strength.

DISCUSSION

Summary of Results

Confirming Hypothesis 1, a significant interaction of performance expectation and behavior type did occur, as did a three-way interaction of performance expectation, behavior type, and rater-training. A performance cue bias was exhibited in the untrained condition in which individuals systematically ascribed more cue-consistent behaviors as having happened than cue-inconsistent behaviors in response to the given performance outcome. As predicted, though, the performance cue bias was significantly reduced when individuals in the source-monitoring training condition were asked to judge whether work behaviors had taken place or not. Therefore, knowledge of performance outcomes led untrained (but not trained) raters to adjust their responses to fit an implicit theory of group performance. What this demonstrates is that when source-monitoring training is administered to raters before they fill out a work behavior rating form, they will be less likely to succumb to the biasing effects of performance expectations. In other words, when raters adjust their decision criteria to only say that a behavior occurred when they can specifically *remember* the details of the behavior rather than just the feelings of *familiarity* that accompany it, they can successfully reduce the impact that knowledge of performance outcomes brings

Supporting Hypothesis 2, the significant three-way interaction of performance expectation, behavior type, and rater-training on decision criterion scores (Br) suggests that individuals trained in source-monitoring do not systematically vary their decision

criteria for effective and ineffective behaviors based on performance expectations given to them. Rather, they tended to adopt a conservative decision criterion in general and exhibited almost equal levels of response for both effective and ineffective behaviors. As in previous research, the decision criterion of untrained raters varied as a function of feedback given. That is, individuals adopted a more liberal decision criterion for behaviors consistent with their expectations and a more conservative one for behaviors inconsistent with their expectations (Martell & Willis, 1991; Martell & Willis, 1993).

In support of Hypothesis 3, the significant three-way interaction of performance expectation, behavior type, and rater-training demonstrates that false alarm rates of individuals trained in source-monitoring were not significantly affected by the performance cue bias. Furthermore, overall false alarm rates for trained raters were lower than for untrained raters, which demonstrates that knowledge of source-monitoring techniques helps individuals to resist identifying behaviors that did not occur, but are compatible with performance expectations.

Finally, backing up the Martell and Guzzo (1991) model, a two-way interaction of performance expectation and behavior type was not significant due to the fact that observed differences in hit rates are not mediated by memory strength. Further, Hypothesis 4 was supported by the lack of a three-way interaction between performance expectation, behavior type, and rater-training. This demonstrates that the source-monitoring training administered to participants did not affect what they actually held in their memory, but rather how they went about using their decision criteria when it came time to make a rating judgment.

Implication for theory development

This study replicates previous research which has found that individuals tend to attribute more effective and fewer ineffective behaviors to a group when they believe that the group had performed well (Martell & Guzzo, 1991; Martell & Willis, 1993; Martell & Leavitt, in press). This study also extends previous research on the performance cue effect (Baltes & Parker, 2000; Martell & Leavitt, in press) by demonstrating that it is possible to reduce the effects of performance expectations on behavioral ratings. What that tells us is that behaviorally-based recognition decisions are a function of either a *Remember* or a *Know* judgment. That is, when an individual has to decide whether a behavior had taken place or not, the individual will either respond with yes (if it fits a Remember of Know judgment) or no (if it fits neither). The expectation was that when participants are asked to refrain from making familiarity-based judgments on past behaviors, and instead rely on specific tangible memories, their biases would be significantly reduced. We found that by introducing the Remember/Know paradigm to participants and instructing them to only answer that a behavior occurred if it fit that of a "Remember" judgment, we could successfully (and significantly) reduce the effects of performance expectations on work behavior ratings. In addition, untrained raters were found to be less accurate in their appraisals and had higher false alarm rates for behaviors that were consistent with their performance expectations. These results were obtained despite previous research, which found that false recognition of associates was an extremely robust phenomena that could not be successfully reduced (McDermott & Roediger, 1998).

The main benefit of this source-monitoring training was the stabilizing abilities it brought to a raters decision criterion being used at the time of judgment. It encouraged them to avoid using familiarity as a basis for assuming a behavior happened and to rely less on their own implicit theories of group processes and outcomes. It appears that a source-monitoring test requires the person to search for various details of a memorial experience and assign a feature weight to each of the details, and then decide whether sufficient information exists to label the memorial experience as seen; while a yes/no test requires the person to set a threshold criterion which determines the amount of information that is necessary before labeling something seen. In essence, the source test simply requires a much more thorough analysis than a yes/no recognition test. By forcing the rater to critically examine the cognitive components that lead to source identification rather than simply allowing them to make threshold-based judgments, this tool is a major asset in the reduction of memorial based biases.

Another finding is that due to the nature of the source-monitoring training, a raters decision criterion will tend to be more conservative than without the training. This is a result of the participant being instructed to scrutinize their actual memory of the behavior and only report that it took place if they can vividly re-imagine it. Since sometimes the actual behaviors that did take place will be overlooked due to the participants inability to relive the behavior under the specific terms of a "Remember judgment", they will have a tendency to be slightly more conservative in their appraisal.

Implications for theory-testing

A major concern in the realm of theory-testing is over the notion that self-report data used to develop and test theories of organizational behavior might be flawed in a systematic way. The findings of this study back up research done by Staw (1975), who initially investigated the performance-cue bias, by further enlightening the psychological community to the notion that an individual's implicit theories can help to shape recollections made by research participants. That is, an individual's memories can often be the result of our judgments and, therefore, may invalidate the answers given on self-report measures. The results obtained in this study would lean toward the conclusion of informing research participants to be consciously aware of how they are arriving at their decisions and what source their memories are being obtained from.

Implications for appraisal

What these results also tell us then is that there are viable options for reducing bias in work behavior ratings. The removal of these effects in performance appraisals are important for organizations to insure that their selection and promotion systems are accurate and fair. The ease of this intervention suggests that it would be a useful addition to a manager/supervisor's evaluation abilities. It also would be a valuable asset to other situations in which raters are biased by implicit theories of performance, stereotypes, or other cognitive schemata. For example, prior research has shown that sex and race-related stereotypes can bias personnel decisions and behavioral ratings (Heilman, 1995).

A Remember/Know judgment intervention would force raters to rely less on feelings of familiarity and potentially reduce the effects of such stereotypes on subsequent behavioral ratings.

Potential Limitations

Although this study found that the performance cue effect could be successfully reduced in work ratings, it is important to note the boundaries of this study and the potential limitations induced by a laboratory setting. First, in situations where there is a long delay between the observation of a behavior and the time of rating, there may be less of an ability on the participants part to accurately recollect behaviors that fit the guidelines of a Remember judgment. This decreased ability might lead to one of two things. First, the individual might be more inclined to mistakenly say that a behavior was recognized as a Remember judgment due to the lapse of time in between the events. Second, the individual might adopt an extremely conservative decision criterion that is the result of their inability to effectively "relive" many of their memories for the behaviors presented. A second limitation of this study is due to the fact that inexperienced raters were used and may have impaired the effectiveness of the source monitoring training. Experienced raters may be more likely to Remember highly diagnostic behaviors and use them as a basis for later behavioral ratings. Future research that involves actual trained raters might find an even greater reduction of this bias in work ratings.

Conclusion

There is a growing body of research contributing to our understanding of how performance-cue biases affect behavior ratings and the possible ways to reduce such errors (Baltes & Parker, 2000a; Martell & Leavitt, in press). The present research extends these findings by introducing a training program, which can significantly reduce the amount of bias that is elicited by performance expectations. Not only is encouraging source monitoring in raters effective, it is also a relatively basic and easily understandable task for the average person. In closing, we believe this study makes a significant contribution to existing research. It shows that an intervention can be developed that successfully reduces the performance cue effect.

REFERENCES CITED

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

Gardiner, J.M. (1988). Functional aspects of recollective experience. Memory and Cognition, 16, 309-313.

Gardiner & Java (1990). Recollective experience in word and nonword recognition. Memory and Cognition, 18, 23-30.

Gardiner & Parkin (1990). Attention and recollective experience in recognition memory. Memory and Cognition, 18, 579-583.

Heilman M. (1995). Sex stereotypes and their effects in the workplace: What we Know and what we don't Know. Journal of Social Behavior and Personality, 10, 3-26.

Jacoby, L.L. (1983). Remembering the data: Analysing interactive processes in reading. Journal of Verbal Learning and Verbal Behavior, 22, 485-508.

Jacoby, L.L. & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. Journal of Experimental Psychology: General, 3, 306-340.

Jacoby, L.L., Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. Journal of Experimental Psychology: General, 118(2), 126-135.

Johnson, M.K., Hashtroudi, S., & Lindsay, D.S. (1993). Source monitoring. Psychological Bulletin, 114(1), 3-28.

Lindsay, D.S. & Johnson, M.K. (1989). The eyewitness suggestibility effect and memory for source. Memory and Cognition, 17(3), 349-358.

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

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. Psychological Review, 87, 252-271.

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. & 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.

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., & 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. & Leavitt, K.N. (in press). Reducing the performance-cue bias in work behavior ratings: Can groups help? Journal of Applied Psychology.

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

Parkin, A.J. & Walter, B. (1992). Recollective experience, normal aging, and frontal dysfunction. Psychology and Aging, 7, 290-298.

Rajaram, S. (1993). Remembering and Knowing: Two means of access to the personal past. Memory and Cognition, 21(1), 89-102.

Roediger III, H.L. (1996). Memory illusions. Journal of Memory and Language, 35, 76-100.

Rush, M.C. & Beauvais, L.L. (1981). A critical analysis of format-induced versus subject-imposed bias in leadership ratings. Journal of Applied Psychology, 66(6), 722-727.

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

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.