A META-ANALYTIC REVIEW OF THE PERFORMANCE-CUE BIAS

by

Ann Michelle Moody

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

of

Master of Science

in

Applied Psychology

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2005
APPROVAL

of a thesis submitted by

Ann Michelle Moody

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.

Dr. Richard F. Martell, Committee Chair 4-05-05

Approved for the Department of Psychology

Dr. Wesley C. Lynch, Psychology Department Head 4-05-05

Approved for the College of Graduate Studies

Dr. Bruce R. McLeod, Dean of College of Graduate Studies 4-05-05
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 in parts may be granted only by the copyright holder.

Ann Moody
14 April 2005
TABLE OF CONTENTS

1. INTRODUCTION ......................................................................................................... 1

   Potential Moderators of the Performance-Cue Bias ..................................................... 3
   Timing of Performance Cue..................................................................................... 4
   Delay ....................................................................................................................... 5
   Rater Perspective .................................................................................................... 5
   Target ...................................................................................................................... 6
   Rating Category ...................................................................................................... 6
   Behavior Rating ...................................................................................................... 7
   Behavior .................................................................................................................. 8
   Training ................................................................................................................... 8

2. METHOD .................................................................................................................... 11

   Sample of Studies ....................................................................................................... 11
   Criteria for Inclusion ................................................................................................... 11
   Coding Scheme ........................................................................................................... 12
   Calculation of Effect Sizes .......................................................................................... 13
   Analytic Issues ............................................................................................................ 14

3. RESULTS .................................................................................................................... 16

   Characteristics of Studies ........................................................................................ 16
   Overall effect of Performance-Cue Bias ..................................................................... 17
   Moderators of the Performance-Cue Bias ................................................................. 17
   Stimulus Modality .................................................................................................... 17
   Timing of Performance-cue .................................................................................... 18
   Delay ....................................................................................................................... 18
   Rater Perspective .................................................................................................. 19
   Target of rating ..................................................................................................... 19
   Rating Category .................................................................................................... 19
   Behavior Ratings ................................................................................................... 20
   Behavior ................................................................................................................ 20
   Training ................................................................................................................. 21
   Publication Type ................................................................................................... 21
   Manipulation Check .............................................................................................. 21

4. DISCUSSION .............................................................................................................. 23

REFERENCES CITED ................................................................................................ 29
v

LIST OF TABLES

Table                                                                                                                               Page

1.   Summary of Study Characteristics........................................................................ 16

2.   Performance-Cue Bias As a Function of Moderators ........................................... 22
ABSTRACT

This review is a quantitative synthesis of the Performance-Cue Bias literature (43 effect sizes, from 43 independent samples, and 4,013 participants). The purpose of this synthesis was to test the overall magnitude of the Performance-Cue bias as well as 9 possible moderators of this phenomenon. Across all moderators, raters who were given positive performance cues rated the target more favorably than raters given negative performance cues. Five potential moderators were identified and shown to significantly influence the Performance-Cue Bias. These significant moderators were; timing of the performance information, target of the rating, type of rating instrument, type of behavior that was being rated, as well as whether participants were part of a training intervention. I discuss the practical ramifications for the identification of these moderators as well as the moderators that were hypothesized to have an impact but did not.
INTRODUCTION

Over 30 years ago Barry Staw (1975) published a study that challenged the way that research in organizations had been conducted and called for a reform in the way self reports were viewed. Conducting Cross sectional (correlation data) was a convenient method used to collect information on how individuals, groups, and organizations were performing. Though collecting survey data at one point in time may be convenient, it is a method unable to pin point the causes of performance, the effects of performance, and other variables that may influence performance. Staw drew upon attribution theory to create, and test, a specific hypothesis in which level of performance was manipulated. Using performance as an independent variable Staw hypothesized that individuals would create attributions of characteristics that fit with this level of performance. These stereotypic attributions would then be used to make self reports. A laboratory experiment was conducted in which 60 undergraduate students were asked to split into groups of thirty to complete a “Financial Puzzle Task”. After 30 minutes of work on the task, groups were given bogus feedback on their performance, either that they had “done quite well” or that they had “not done too well”. Though this feedback had no link to their actual performance, group members evaluation of themselves and their group revealed an evaluation that matched the performance information that they had been given. As an extension to this laboratory experiment an interpersonal simulation was created in which 60 students were asked to listen to, and read, a description of how a group performed the “Financial Puzzel Task”. Though they themselves did not complete the task the evaluations completed by these participants mirrored those of the laboratory experiment.
such that those who were asked to rate a high performing group completed evaluations that were significantly more favorable when compared to those who were asked to rate a low performing group. These results supported Staw’s hypothesis, that “significant correlations between performance and self-report data may only be reflecting the respondent’s “theories” of organizational performance rather than actual events”, p.417. These findings not only supported Staw’s hypothesis but led organizational research in a new direction.

Since Staw set his challenge, researchers have gone on to recreate his study design (Bachrach, Bendoly, & Podsakoff, 2001; Binning & Lord, 1980; Denisi & Pritchard, 1978; Rush, Thomas, & Lord, 1977), and contribute their own hypothesis and unique conclusions. Research has also gone beyond ratings of just leaders, work groups, and self evaluations, and explored bias in organizations (Gardiner, Wright, & Gerhart, 2000; Lord & Maher, 1989). The possibility that Performance-Cue Bias goes beyond the laboratory and into field settings (Fichman, 1991; Gardiner et al., 2000) only strengthens Staw’s concern that this issue be given the attention it deserves. However, over the course of the past 35 years only one study has been published in an attempt to synthesize performance cue research into a narrative review (Lord & Maher, 1989). Though this qualitative effort has brought understanding of this phenomenon one step closer to Staw’s vision it lacks the quantitative analysis needed to fully comprehend the biasing effects of performance information. The current research is a meta-analytic review of performance-cue research, a quantitative attempt to provide an overall estimate of the magnitude of the Performance-Cue Bias, and an examination of the conditions in which bias is present. Understanding the overall impact that the Performance-Cue Bias can have on
performance assessments and the contextual factors that influence its impact can be vital to any organization. Based on the hypothesis that: *Raters given positive performance cues will rate the target more favorably than raters given negative performance cues*, the overall magnitude of the Performance-Cue Bias is investigated in this review. Numerous studies have been conducted based on the Performance-Cue Bias but discrepant findings call for an examination of the contextual factors that moderate the impact of performance-cues. The following research predictions address nine contextual factors that theory suggests may be moderators of the Performance-Cue Bias.

**Potential Moderators of the Performance-Cue Bias**

**Stimulus Modality**

The moderator coding process began with a look at what each study presented to their participants as stimulus materials. Studies were coded as presenting participants with videotapes, written scenarios, audiotapes, or creating a live work sample in the laboratory or an organization. Researchers often use videotapes to depict a leader or group working on an organizational task (Foti & Hauenstein 1993; Lord, Binning, Rush & Thomas 1978; Baltes & Parker, 2000), but seldomly use written scenarios. A flaw that must be considered with a written stimulus modality is that its impact may be short-lived and one of low-intensity, because of the distance between the rater and the target (Pillai, 1997). On the other end of the spectrum, however, some researchers use live work samples as a medium for providing performance information- a modality where there is limited distance between rater and target. In fact, some studies that employ a live work sample methodology ask their raters to rate themselves on their performance. The benefit
of creating a live work sample is that participants are not just passive observers but are engaged in concrete organizational tasks and behaviors. It can be argued that these participants are a more reliable reflection of actual organizational groups because they are sharing in the success or failure of the group (Nye, 2003). The possible difference in a performance cue’s impact given various modalities leads to the first research question. Research question one: To what extent does the magnitude of the performance-cue bias vary by stimulus modality?

**Timing of Performance Cue**

If there is a difference between presenting raters with performance information before they observe the target, after the observation, or within the stimulus presentation, avoiding and eliminating bias could prove relatively easy. An information processing perspective would suggest that expectations that develop prior to or within target observation may influence not just retrieval of behaviors but also attention and encoding, while expectations developed after observing a target are less susceptible to this influence. Research has been conducted that supports this perspective, with significant bias being shown in a preobservation condition and not in a postobservation condition (Mitchell, Larson, & Green, 1977). Other researcher suggests that performance information has the same impact regardless of when it is presented because a rater’s implicit theories exert their influence at the time of rating (Larson, 1982). This debate over when performance expectations influence raters leads to a second research question. Research Question 2: To what extent does the magnitude of the performance-cue bias
vary as a function of when the performance-cues are administered (viz. Prior to target observation versus during observation versus after target observation)?

**Delay**

Some research suggests a delay actually increases the influence of performance information, while other research suggests that there is no difference between immediate and delay conditions (Martell & Guzzo, 1991; Martell & Willis, 1993). This moderator is clearly a valuable one in the effort to simulate real world organizations in laboratory experiments. Researchers that have chosen a delay methodology have pointed to its ability to more accurately simulate the organizational climate, which it looks to better understand (Parker, 1999). This argument, that a delay may be more representative of true organizational practices is in contrast to the research that shows a delay may cause more biased ratings, creating a strong motivation to examine the following research question. Research Question 3: To what extent does the magnitude of the performance-cue bias vary by delay between stimulus presentation and rating?

**Rater Perspective**

It can not be assumed that the study of how observers make attributions and thus ratings is identical to the process that actual participants utilize (Nye, 2002). There is a call to examine actual groups and participants as opposed to “outsiders looking in” (Nye, 2003). Research that has directly manipulated a rater’s perspective has confirmed this difference and found a more significant biasing effect when looking at ratings made by observers, in contrast to those made by participants (Mitchell, Larson, & Green, 1977; McElroy & Downey, 1983). Findings based on this moderator will have implications for
whether organizations chose to use outside observers or intimate participants when designing assessment programs. The effect of rater perspective drives the following question. Research Question 4: To what extent does the magnitude of the performance-cue bias vary by rater perspective (viz. Observer, Participant)?

**Target**

Throughout the Performance-Cue Bias literature, researchers have varied the targets that raters observe and rate. Many ratings are that of leaders (Murphy & Jones, 1993; Doverspike & Cellar, 1987; Foti & Hauenstein, 1993; Butterfield & Powell, 1981), groups (Guzzo, Wagner, McGuire, & Herr, 1986; Bachrach, Bendoly, & Podsakoff, 2001; Martell & Guzzo, 1991), and even self evaluations (Binning & Lord, 1980; Denisi & Pritchard, 1978; McElroy & Downing, 1982; Staw, 1975). While studies vary in the target they choose there has been little discussion as to whether this variance in target plays a role in the influence performance information may have on a rater. This question’s applicability is apparent if performance cue research were to generalize to all types of rating, whether it be that of a leader, group, or individual. This scarcity of inquiry leads to the following research question. Research Question 5: To what extent does the magnitude of the performance cue bias vary by rating target (viz. Leader, Group, Self)?

**Rating Category**

In addition to the target of rating there are other specific details about the rating measures that need to be considered. The first is the nature of rating instruments, is it a subjective evaluation based on a Likert type scale, or a more objective behavioral
measurement? Research seems to suggest that more objective, behaviorally based, measures may be more resistant to bias (Binning, Zaba, and Whattam, 1986; McElroy & Downey, 1982).

Significant bias has been shown when participants were asked to complete ratings after no actual behavior had been observed (Rush, Thomas, & Lord, 1977). In contrast, when raters were presented with stimulus materials filled with relevant behaviors the effect of performance information was very small (Lord, Binning, Rush & Thomas, 1978; Gioia & Sims, 1985). Should these observations continue to be supported, attention should be focused on the value of using measurements of specific behaviors to combat the Performance-Cue Bias (Parker, 1999). This leads to the following research question.

Research Question 6: To what extent do performance cues interact with the subjective-objective nature of the rating instrument? Specifically, is the performance-cue bias greater for subjective evaluations than objective-behavioral measures?

Behavior Rating

Obviously the subjective vs. objective rating argument is one of interest, but what if there are differences within these broad rating categories? Though researchers have not dealt with this issue specifically, the variability amongst rating measures is an issue that deserves consideration. Focusing on objective ratings alone, four different measurements surface throughout the literature. Researchers have measured behavior frequency (Mitchell, Larson, & Green, 1977), hit rates (Martell & Guzzo, 1991), false alarm rates (Phillips, 1984), and free recall of relevant behaviors (Baltes & Parker, 2000). The possible variability between these measurements has led to the following question.
Research Question 7: To what extent does the magnitude of the performance-cue bias vary by behavioral rating type (viz., Behavior Frequency, Hit Rates, False Alarm Rates, Free Recall)?

**Behavior**

The actual nature of the behaviors being observed deserves attention as relevant behaviors that raters are presented with can vary in effectiveness. Some behaviors are positive in nature and characteristic of effective leadership or group membership while others are more negative, signaling an ineffective behavior. Researchers have used stimulus materials that include both effective and ineffective behaviors (Martell & Willis, 1993; Foti & Hauenstein, 1993; Doverspike, 1987; Phillips & Lord, 1982; Phillips, 1984; Baltes & Parker, 2000; Murphy & Jones, 1993). Our following research question deals with the potential difference in rating these behaviors. Research Question 8: To what extent does the magnitude of the performance-cue bias vary by behavior type (viz. Effective, Ineffective)?

**Training**

The previous 9 questions have dealt with moderators that could potentially influence the effect of a performance cue. From stimulus materials to rater perspective, evaluative to behavioral ratings, the literature has been broad and varied in its view of the Performance-Cue Bias and how it is created and intensified. In contrast, there has been limited focus on creating trainings and interventions to reduce this inevitable rating bias. The first intervention suggested to help reduce the Performance-Cue Bias was a free recall intervention (Baltes & Parker, 2000) that asked raters to spend a few moments,
before rating, recalling as many leader behaviors as they could generate in regards to a video they had viewed. This intervention was hypothesized to make all observed behaviors more readily accessible in the rater’s memory, not just those that fit with the performance information they had received. Though there was no difference between raters that had received the free recall intervention and those that had not, training research continued to be investigated. Using the free recall intervention as a foundation, a second intervention was developed that lead raters through a structured recall task before they completed their ratings (Baltes & Parker, 2000). Similar to the free recall intervention, the structured recall intervention was developed to influence the accessibility of observed, relevant, behaviors. By having participants in the intervention condition view performance dimensions that formed the basis of the behavioral ratings they would complete, and then list behaviors they had observed that matched these dimensions, the structured recall participants showed reduced reliance on performance cue information. This promising intervention was tested alongside halo error training that instructed participants about the phenomenon of halo error rating, presented them with an example of ratings that were plagued by error and also a rating that was free from bias. Participants in this halo error condition were also encouraged to consider strengths and weaknesses of the leader they were rating and not to rely on overall judgments. Results were not as promising as those found in the analysis of the structured recall condition. Halo error participants still showed lower rating accuracy when compared to participants in the control condition. Though these interventions could not completely eliminate the effects of the Performance-Cue Bias they inspired further research that investigated other possible interventions. One such intervention used groups to rate
performance as opposed to individual raters (Martell & Leavitt, 2002). Based upon group memory research, this intervention asked participants to engage in a group discussion about what they had observed in their viewing of a videotaped work group and to reach a consensus prior to completing their rating. This research confirmed the hypothesis that group discussion would eliminate the effects of performance feedback.

The most recent attempt to reduce the performance cue bias was made using a source-monitoring procedure (Martell & Evans, 2005). Raters who were a part of the source-monitoring intervention were encouraged to make judgments based only on behaviors that they could vividly recall in detail as opposed to reporting behaviors that felt familiar to them. This training technique was designed to help increase the rate of confident and correct responses and decrease the amount of biased judgments that are made based on a feeling of familiarity. Results indicated that source monitoring participants were less biased in their responses and, in some cases, that they reflected completely unbiased ratings. These findings are encouraging and provide a future direction in the effort to eliminate the biasing effects of performance information. This leads to a final research question. Research Question 9: To what extent is the magnitude of the performance-cue bias influenced by the implementation of a training procedure?
METHOD

Sample of Studies

To identify all experimental studies of the Performance-Cue Bias, we proceeded as follows, focuses our search on research conducted from 1975 to 2005. First, we conducted a literature search of the Journal of Applied Psychology, Organizational Behavior and Human Decision Processes, Personnel Psychology, Academy of Management Journal, Academy of Management Review, Journal of Applied Social Psychology, Journal of Management, Journal of Business and Psychology, Education and Psychological Measurement, Advances in Information Processing in Organization, Small Group Research, Journal of Social Behavior and Personality, and Journal of Management Systems. Second, numerous review articles, books, and the reference sections of all identified studies were reviewed. Third, a computer search was conducted using the following data bases; Psychological Abstracts (PsycINFO, 1975 to 2003), Educational Resources Information Center (ERIC, 1975 to 2003), a large-scale management data base (ABI/INFORM, 1975 to 2003), and Dissertation Abstracts International (DISS, 1975 to 2003). The Social SciSearch data base was also used to locate articles that cited the Staw (1975) article. Finally, we contacted researchers who had published previously on the performance-cue bias to request unpublished studies.

Criteria for Inclusion

We included studies in our meta-analysis that (1) explicitly provided subjects with positive and negative feedback regarding the work performance of an individual (either
themselves, or another person), or a group (one’s own group, or another group), and (2) included a target rating that was either of a subjective-evaluative or objective-behavioral nature. Any additional information other than performance feedback must have been held constant or varied identically in all feedback conditions. The above selection criteria produced 37 primary studies, 28 of which (which included 32 experimental manipulations) appear in this review. There were 9 studies that fit the above criteria but are not present in the subsequent analysis. For most of these 9 studies, coding was incomplete because necessary statistical information was absent from the manuscript. In each of these cases an effort was made to contact the author asking for the missing information. If the information was not retrieved then studies were removed before the analysis began.

Coding Scheme

The following information was coded for each study: a) rater gender, b) publication year, c) rater population (i.e. undergraduate students, graduate students, managers, employees, or a mixed sample), d) form of publication (i.e. journal, book, technical report, dissertation, thesis, or unpublished manuscript), e) stimulus modality (i.e. written, videotape, audiotape, live work sample), f) research site (i.e. laboratory or field), g) manipulation check on feedback effect (i.e. yes or no), h) perspective of rater (i.e. participant or observer), i) timing of feedback manipulation (i.e. prior to target observation, after observation, or sometime within the observation period), j) familiarity with the target of rating (i.e. no prior history with the target or previous working history with the target), k) amount of ratee information (less than 10 behaviors or greater than 10
behaviors, l) length of stimulus task (in minutes), m) target of rating (i.e. leader, group, self), n) rating category (i.e. subjective-evaluative, which required evaluative scale ratings or objective-behavioral, which required recognition of specific rate behavior)?

These variables were coded for each study by two independent raters, Julie Maertens and Ann Moody. Inter-rater reliability (using Cohen’s Kappa) was found to be quite high, with only one potential moderator (i.e, familiarity) below .99. The few discrepancies in coding were resolved through discussion.

Calculation of Effect Sizes

All reported effect sizes were derived from between-subjects designs. These effect sizes were calculated as the mean bias of the positive feedback group minus the mean bias of the negative feedback group divided by the pooled standard deviation (Rosenthal, 1994). The final overall data set yielded 43 effect sizes to be used for analysis. Effect size coding was conducted by computing g, the standard effect size estimate (Hedges & Olkin, 1985). The g index tends to overestimate the magnitude of population effect size, especially when the sample size is small (Hedges & Olkin, 1985). Thus, g’s were converted to Cohen’s d. Various statistical values were used to calculate the g’s. The use of means and standard deviations was the most common method of calculating effect sizes, yielding 34 effects in the overall data set. F and t values were also used. There were 6 effect sizes derived from F values and just 3 from t values. The sign of the effect is important to clarify so as to understand the direction and magnitude of the Performance-Cue Bias. Positive effect sizes show bias in the predicted direction such that raters given positive performance information rate their target more favorably
and raters who are given negative performance information rate their target less favorably. Negative effect sizes indicate reverse bias such that those given positive information rate the target less favorably than raters given negative feedback, or conversely, that raters given negative information rate their targets more favorably than raters given positive information. If the effect sizes are zero, there is no sign of bias in either direction and therefore there is no effect. To determine whether aggregated effect sizes differed between moderator categories, the $\chi^2$ distributed $QB$ statistic, was used.

**Analytic Issues**

A fundamental analytic issue that any researcher conducting a meta-analysis must deal with is the choice between two statistical models: random or fixed effects (Field, 2001; Hedges & Vevea, 1998; Hunter & Schmidt, 2000). When the homogeneity assumption is satisfied, fixed-effects can prove more powerful. However, when all the effect sizes being analyzed are not derived from a common population (i.e., the homogeneity assumption is not satisfied), then a fixed-effects model will inflate the Type I error rate, thus underestimating confidence intervals, as well as underestimating standard errors of parameter estimates. For example, Monte Carlo simulations suggest that the Type I error rate in heterogeneous fixed-effects models range between .43 and .80, dramatically higher than the .05 level (Field, 2003). Each moderator analysis began with an initial test of the homogeneity assumption of the fixed effects model. This test is equivalent to the test of random-effects variance. In all but one instance, the random effects tests were equal to zero, thus confirming the homogeneity assumption. Consequently, random-effects analysis was employed except where needed. The
moderator that violated the assumption (as subsequently noted) was analyzed using the more appropriate fixed-effects model.

An overall data set was created in which each study contributed one effect size per independent sample for a total of 43 effect sizes. To complete subsequent analysis, additional data sets were created in which some studies contributed more than one effect size. For example, a target data set was created to examine the effect that rating target may have on Performance-Cue Bias. In this target data set some studies may have contributed two to three effect sizes if participants rated themselves in addition to rating their group and its leader. A total of 4 additional data sets were created to analyze the effect of target of rating, behavior type, rating type, and behavioral rating type. In these additional data sets each study contributed only 1 effect size per level of the appropriate moderator.
RESULTS

Characteristics of Studies

Table 1 presents the variables that were coded for each study and the number of studies that contributed effect sizes that were used in the subsequent analysis.

Table 1
Summary of Study Characteristics

<table>
<thead>
<tr>
<th>Variable and Class</th>
<th>Value</th>
<th>Variable and Class</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median date of publication</td>
<td>1985</td>
<td>Perspective of rater</td>
<td></td>
</tr>
<tr>
<td>Publication form</td>
<td></td>
<td>Participant</td>
<td>9</td>
</tr>
<tr>
<td>Journal article</td>
<td></td>
<td>Observer</td>
<td>23</td>
</tr>
<tr>
<td>Unpublished manuscript</td>
<td></td>
<td>Familiarity with target of rating</td>
<td></td>
</tr>
<tr>
<td>Median number of subjects</td>
<td>114</td>
<td>No prior history</td>
<td>30</td>
</tr>
<tr>
<td>Median percentage of female subjects</td>
<td>50%</td>
<td>Prior History</td>
<td>2</td>
</tr>
<tr>
<td>Subject Population</td>
<td></td>
<td>Target</td>
<td></td>
</tr>
<tr>
<td>Undergraduate students</td>
<td>32</td>
<td>Leader</td>
<td>20</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>Group</td>
<td>15</td>
</tr>
<tr>
<td>Research site</td>
<td></td>
<td>Self</td>
<td>4</td>
</tr>
<tr>
<td>Laboratory</td>
<td>32</td>
<td>Delay between stimulus and rating</td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td>0</td>
<td>No delay</td>
<td>29</td>
</tr>
<tr>
<td>Stimulus modality</td>
<td></td>
<td>Delay</td>
<td>5</td>
</tr>
<tr>
<td>Written scenario</td>
<td>4</td>
<td>Rating Category</td>
<td></td>
</tr>
<tr>
<td>Videotape</td>
<td>18</td>
<td>Evaluative</td>
<td>19</td>
</tr>
<tr>
<td>Audiotape</td>
<td>1</td>
<td>Behavioral</td>
<td>23</td>
</tr>
<tr>
<td>Live work sample in lab</td>
<td>9</td>
<td>Behavior Type</td>
<td></td>
</tr>
<tr>
<td>Timing of performance information</td>
<td></td>
<td>Effective</td>
<td>12</td>
</tr>
<tr>
<td>Prior to observation of the target</td>
<td>13</td>
<td>Ineffective</td>
<td>12</td>
</tr>
<tr>
<td>After observation of the target</td>
<td>19</td>
<td>Behavioral Rating</td>
<td></td>
</tr>
<tr>
<td>Within observation of the target</td>
<td>2</td>
<td>Free recall of relevant behaviors</td>
<td>1</td>
</tr>
<tr>
<td>Manipulation check on perf. Information</td>
<td></td>
<td>Behavior frequency</td>
<td>14</td>
</tr>
<tr>
<td>Yes</td>
<td>23</td>
<td>Hit rate</td>
<td>10</td>
</tr>
<tr>
<td>No</td>
<td>9</td>
<td>False alarm rate</td>
<td>8</td>
</tr>
</tbody>
</table>

Note. Study characteristics are based on the overall data set, compiled from 28 studies and 32 experimental manipulations.

These characteristics show that, in general, studies (a) were published recently, (b) appeared in prominent journals, (c) used undergraduate students as raters, (d) and were conducted in laboratory settings. Concerning stimulus presentation, studies (a) used videotapes most often, (b) presented performance information after target observation, (c)
and completed a manipulation check on this information. The relationship between raters and their targets was generally (a) one in which the rater was an observer, (b) who was unfamiliar with the target, (c) who was most often a leader.

Regarding the ratings that were used in these studies, (a) most were presented immediately after stimulus presentation, (b) and were behavioral in nature, (c) sometimes asking raters to report both effective and ineffective behaviors, (d) but most often asking raters to report behavioral frequency.

**Overall effect of Performance-Cue Bias**

The overall effect of Performance-Cue Bias was calculated using 43 effect sizes derived from 28 primary studies. Averaging across all moderators, $d = .65$, (95% CI .53/.77), $QW(42) = 70.73, p<.01$, an effect that is both statistically significant and moderately strong. Table 2 breaks down this overall effect by potential moderators.

**Moderators of the Performance-Cue Bias**

**Stimulus Modality**

Various modalities were used to present participants with stimulus information. Most studies used a videotape as a stimulus source, while others used written, audio taped, or live work sample sources, there was no significant difference between these four levels, $QB(3) = .36, ns$. There was significant bias shown in the video tape, $d = .69$, (95% CI .52/.86), written, $d = .63$, (95% CI .27/1.00), and live work sample, $d = .59$, (95% CI .33/.86), modalities. Interestingly, there was no significant bias shown in the audio tape conditions, $d = .61$, (95% CI -.86/2.07). It should be noted that because there
was only one effect size that came from a study that used an audiotape the analysis was conducted without this level of the stimulus modality. This exploratory analysis did not differ from the analysis presented above.

Timing of Performance-cue

When the performance-cue is introduced has an effect on biased ratings. A significant difference was shown between the three levels $QB(2) = 7.84$, $p<.05$ of timing. Post hoc tests, using Scheffe adjustments, revealed that this difference was between after observation performance cues and within observation performance cues, $d = .59$, (95% CI .07/1.10). Receiving a performance cue prior to observation was shown to produce biased ratings, $d = .55$, (95% CI .35/.75), as was receiving the performance feedback after observing the target, $d = .77$, (95% CI .62/.92). There was no bias found across studies where the performance information was given during or within the presentation of the stimulus materials, $d = .18$, (95% CI -.25/.61). Additional analyses were conducted which took the within level of timing out of the analysis because only two effect sizes contributed to this level of timing moderator, removing this level had no impact on the results so the complete analysis was presented.

Delay

Some studies used a design that included a delay between presentation of the stimulus and completion of the rating instrument. There was no significant difference between the two times of rating, $QB(1) = 1.94$, $ns$. There was significant bias found, $d = .40$, (95% CI .01/.78) when looking at the delay studies. The remainder of the studies used a design where the rating instrument was given immediately after presentation of the
stimulus. There was significant bias in these immediate conditions as well, $d = .68$, (95% CI .55/.82). It must be noted that time of rating could not be coded for one study (Binning & Lord, 1980).

**Rater Perspective**

Rater’s, regardless of their role as participants or observers, showed significant bias in their ratings of the target. No difference was shown between the two rater levels $QB(1) = .28$, *ns*. With a $d = .59$, (95% CI .33/.85) participant raters and observer raters, $d = .68$, (95% CI .53/.83) biased their ratings in the predicted direction.

**Target of rating**

There was a significant difference between the three target types, leader, group, and self, $QB(2) = 8.30$, *p<.05*. Given the fact that target of rating had three moderator levels a Post hoc test was necessary. Pair wise comparisons using Scheffe adjustments revealed that there was a significant difference when looking at comparisons of leader vs. group, $d = -.34$, (95% CI -.61/-07). Bias was significant regardless of the target of rating. The three levels of the target moderator; leader ($d = .52$, (95% CI .37/.66)), group ($d = .86$, (95% CI .67/1.04)), and self ($d = .59$, (95% CI .20/.97) all showed significant bias. It must be noted that some studies asked participants to rate more than one target so 50 effect sizes were used for this analysis.

**Rating Category**

Breaking this overall effect down by rating instrument; some studies used subjective, evaluative rating measures while others asked raters to complete a more
objective, behavioral based, rating instrument. The effect size for evaluative measures, $d = .98$, (95% CI .80/1.17), significantly differed from that of behavior measures, $d = .52$, (95% CI .37/.66), $QB(1) = 14.94$, $p<.01$. Some studies asked participants to complete more than one rating measurement so 58 effect sizes were used for this analysis.

**Behavior Ratings**

Within the behavioral rating subset there were four types of rating measures; free recall, behavioral frequency, hit rates, and false alarm rates. There was no significant difference between the four types of ratings, $QB(3) = 0.09$, ns. Significant bias was shown in all behavioral rating types except for free recall; $d = .40$, (95% CI -.37/1.17), $d = .51$, (95% CI .29/.73), $d = .48$, (95% CI .27/.68), $d = .50$, (95% CI .25/.74), respectively. Some studies were able to contribute multiple behavior ratings from individual participants, accordingly, 56 effect sizes were used for this analysis. Of these 56 effect sizes only 2 came from free recall measurements, additional analysis were conducted to see if the effect of behavior rating would change if free recall measurements were excluded. This was not the case so you will see that they appear in our analysis.

**Behavior**

Behavior type was a moderator that was broken down into two levels based upon the behaviors participants were being asked to report. There was significant bias when reporting effective behaviors, $d = .62$, (95% CI .46/.77), which significantly differed, $QB(1) = 3.50$, $p<.06$, from the reporting of ineffective behaviors, $d = .40$, (95% CI .25/.56). For some studies behavior type was not specified so the behavior moderator analysis consisted of only 42 effect sizes.
Training

In hopes of reducing the biasing effects of performance cues some studies included training conditions. Training conditions differed significantly, $QB(1) = 3.86$, $p<.05$, from no training conditions, where bias was significant, $d = .70$, (95% CI .57/.83). For those studies that implemented a training of some type bias was not significant, $d = .30$, (95% CI -.07/.67).

Publication Type

The primary sample of studies came from both published and unpublished manuscripts. There was no significant difference between the two types of publication, $QB(1) = .40$, $p>.05$. The sample of studies that came from journal articles, 39 effect sizes, revealed a significant bias, $d = .67$, (95% CI .53/.81), as did studies that were unpublished (4 effect sizes) $d = .54$, (95% CI .17/.91).

Manipulation Check

Though most studies completed a manipulation check on the performance information participants received (33 effect sizes) others did not (10 effect sizes). There was no significant difference between samples that completed a check and those that did not, $QB(1) = 1.00$, ns, and both showed a significant bias, $d = .62$, (95% CI .47/.77), $d = .78$, (95% CI .51/1.05), respectively.
Table 2
Performance Cue Bias as a Function of Moderators

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Level</th>
<th>k^a</th>
<th>d</th>
<th>95% CI</th>
<th>QB^b</th>
<th>QW^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stimulus Modality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Written</td>
<td>4</td>
<td>.63</td>
<td>.27</td>
<td>/1.00</td>
<td>.36</td>
<td>70.06**</td>
</tr>
<tr>
<td>Video</td>
<td>29</td>
<td>.69</td>
<td>.52</td>
<td>/.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio</td>
<td>1</td>
<td>.61</td>
<td>-.86</td>
<td>/2.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td>9</td>
<td>.59</td>
<td>.33</td>
<td>/.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timing of PC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.84*</td>
<td>58.06*</td>
</tr>
<tr>
<td>Prior</td>
<td>20</td>
<td>.55</td>
<td>.35</td>
<td>/.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>After</td>
<td>21</td>
<td>.77</td>
<td>.62</td>
<td>/.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>2</td>
<td>.18</td>
<td>-.25</td>
<td>/.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time of Rating</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.94</td>
<td>67.04**</td>
</tr>
<tr>
<td>No Delay</td>
<td>36</td>
<td>.68</td>
<td>.55</td>
<td>/.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay</td>
<td>6</td>
<td>.40</td>
<td>.01</td>
<td>/.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rater Perspective</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.28</td>
<td>70.73**</td>
</tr>
<tr>
<td>Participant</td>
<td>9</td>
<td>.60</td>
<td>.33</td>
<td>/.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observer</td>
<td>34</td>
<td>.68</td>
<td>.53</td>
<td>/.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.30*</td>
<td>64.25*</td>
</tr>
<tr>
<td>Leader</td>
<td>27</td>
<td>.52</td>
<td>.37</td>
<td>/.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>19</td>
<td>.86</td>
<td>.67</td>
<td>/1.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>4</td>
<td>.59</td>
<td>.20</td>
<td>/.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rating Category</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14.94**</td>
<td>102.37**</td>
</tr>
<tr>
<td>Evaluative</td>
<td>22</td>
<td>.98</td>
<td>.80</td>
<td>/1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavioral</td>
<td>36</td>
<td>.52</td>
<td>.37</td>
<td>/.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavior Ratings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.09</td>
<td>96.43**</td>
</tr>
<tr>
<td>Free Recall</td>
<td>2</td>
<td>.40</td>
<td>-.40</td>
<td>/1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavior Frequency</td>
<td>16</td>
<td>.51</td>
<td>.29</td>
<td>/.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit Rate</td>
<td>22</td>
<td>.48</td>
<td>.27</td>
<td>/.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Alarm Rate</td>
<td>16</td>
<td>.50</td>
<td>.25</td>
<td>/.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavior Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.50*</td>
<td></td>
</tr>
<tr>
<td>Effective</td>
<td>21</td>
<td>.61</td>
<td>.46</td>
<td>/.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ineffective</td>
<td>21</td>
<td>.40</td>
<td>.25</td>
<td>/.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.86*</td>
<td>65.49**</td>
</tr>
<tr>
<td>yes</td>
<td>6</td>
<td>.30</td>
<td>-.07</td>
<td>/.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>37</td>
<td>.70</td>
<td>.57</td>
<td>/.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total k</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^a Indicates the number of effect sizes.
^b QB tests for differences across levels within a given moderator, with df equal to one less than the number of levels (e.g. df = 2 for target, which consists of leader, group, & self)
^c QW test of the homogeneity assumption.
^p < .06  * p < .05  ** p < .01
DISCUSSION

The findings of this meta-analysis are important in a number of ways. Support found for the initial hypothesis that *raters given positive performance cues will rate the target more favorably than raters given negative performance cue*, speaks to the magnitude of the Performance-Cue Bias. In statistical terms, $d=.65$, the overall effect of performance information can be considered a medium effect. In practical terms, the Performance-Cue Bias is a phenomenon with influence on organizational assessments that can not go unnoticed.

There are a number of key moderators that were identified and shown to affect the impact of the Performance-Cue Bias. Beginning with presentation of performance information, timing was discovered as a significant moderator such that bias was found when raters were given performance information prior to observing a target or after the observation but not when performance information was presented during stimulus presentation. This finding shows that presenting performance information within stimulus presentation results in less biased ratings, when compared to presenting this information before or after stimulus presentation. Organizations should consider this finding and think about presenting performance information within the actual observation of a leader, group, or company. Decision makers should also consider the finding that though bias was found regardless of whether the target was a leader, a group, or yourself, there was significantly more bias found when rating a group. The heightened susceptibility of being influenced by performance information when rating a group is a
finding that needs to be considered when choosing the most accurate way to conduct an assessment.

Another significant moderator that deserves attention is the difference in rating instruments. Though bias was found in both behavioral and evaluative ratings, the impact biasing information had on evaluative ratings was much more profound. This finding should encourage the use of assessment instruments that are based on concrete and objective behaviors as opposed to subjective, opinion based scales. It is important to realize that though behaviorally based ratings may be more resistant to Performance-Cue Bias, they are not free from error. When behavioral ratings were broken down, behavior itself was shown to be a significant moderator as effective behaviors were revealed as more prone to bias in comparison to ratings of ineffective behaviors. These findings call for a need to be cautious when making decisions that are based only on positives qualities and performance, for the ratings of effective behaviors are more likely to be biased. The above discussion of key organizational variables and their analysis can provide valuable insight to decision-makers as they design assessment procedures in their organizations.

Decision makers should also consider the significant impact that various trainings had on reducing and eliminating the performance-cue bias. The use of structured free recall, source-monitoring, and groups as raters have all been shown to reduce the Performance-Cue Bias. Organizations should consider the use of these trainings to combat the powerful impact that performance information can have and continue to research and implement other promising training interventions.

The identification of organizational factors that do not moderate the Performance-Cue Bias can be also be useful when examining rating bias so that time, money, energy,
and focus are not wasted on variables that do not have an impact. Regardless of choice in stimulus modality, time of rating, rater perspective, and type of behavior rating, the impact of the performance-cue bias was consistent. This is a limitation of the present research given the initial predictions made that these moderators would have significant impact on the Performance-Cue Bias. It should be noted that an attempt was made to investigate the impact that familiarity, between raters and ratees, would have on the Performance-Cue Bias. The coding for this moderator resulted in only two effect sizes that could be derived from groups who had had some type of previous history. The limited ability to compare these two effect sizes to the 41 that could be gleaned from studies that used unfamiliar raters led to this moderator being dropped from the results that have been presented. This limited look at familiarity that has been shown in the past offers a new direction in performance research. We need to look not only at the ratings made by unfamiliar raters but also explore Performance-Cue Bias and its magnitude amongst raters and ratees that have a prior history.

Though familiarity was dropped, there were other moderator levels that had very small numbers of effect sizes that could represent a unique level of the following moderators; stimulus modality, timing of performance-cue, and behavior ratings. Similar to the limited look at familiar raters and ratees, there were few studies that used audiotapes to present information to raters, presented performance information within stimulus presentation, and used free recall measurements. Though small, these cells should not be discounted from this review nor overlooked by future research. We should not view this as merely a limitation of the research, but see it as an opportunity to investigate these moderators using unconventional and novel experimental methods.
There are a number of other limitations that need to be discussed in evaluating the results of this review. The most basic limitation of this review is one common to the literature from which it drew from, a reliance on experimental studies. Though efforts were made to collect, review, and include field studies as well as experimental manipulations of the Performance-Cue Bias, only experimental studies provided the adequate information needed to statistically analyze their results. The 4013 participants that make up the sample population for this review were all undergraduates. Though there were studies that used managers and MBA students to complete ratings, none provided the statistical information that was needed. Though this restricts the conclusions that can be made about the Performance-Cue Bias, by limiting the external validity of this research, it also provides opportunities for future directions. This is not a new concern in organizational research; Staw (1975) addressed it in the study that forms the foundation for this analysis. Staw suggested the careful choice of variables and measurements, always considering the field when creating a laboratory study. He also encouraged conducting research in actual organizations by using strong methodology that is able to provide both internal and external validity. The rarity of these studies and the fact that none were able to be included in this analysis speaks to a need for action amongst organizational researchers.

The number of studies that were included in this review is a potential limitation for the generalizability of this research. There were 28 studies that our analysis was based on, a number some might consider too small to produce meaningful results. It should be noted, however, that many comprehensive meta-analyses have been conducted.
that include a similar number of studies and overall effect sizes (Hosoda, Stone-Romero, & Coats, 2003).

The potential threat of publication bias is a concern that was addressed and analyzed throughout the review process. The initial search for studies was not limited to research published in top tear journals. Though the effort was made to contact researchers and to search through dissertations to locate relevant unpublished studies, only 4 effect sizes came from this publication source. Though this may seem small compared to the 29 effect sizes that came from published manuscripts, there was no significant difference between the two types of publications, both showing significant levels of bias.

Taking together the moderators which have been shown to impact the influence of performance information and the limitations that have prevented complete understanding of this phenomenon, a discussion about future research directions is necessary. The call for solid experimental research in the field is one that can not go unnoticed. Organizational researchers need to work in cooperation with organizations if the Performance-Cue Bias is to be eliminated. These findings need to extend beyond the laboratory and the college classroom. Assessment programs need to be created that are behaviorally based and are sensitive to the difference between how bias may impact effective vs. ineffective behaviors. We need to discover if a Department Manager is rated differently alone then when his entire department is reviewed. We need to determine if it is even possible to present performance information only during assessment observation. Can assessors always conduct their work in groups? Can rater’s stifle familiar responses and respond only to items they can vividly recall? How much harm can this little bias do? Over time how would two employees differ if one was exposed to bias and the other
received an evaluation that was accurate and fair? Answering these questions is going to be challenging, and require ambitious and innovative researchers to take the necessary steps to eliminate the Performance-Cue Bias.
REFERENCES CITED

*References marked with an asterisk indicate that the studies were included in the meta-analysis


*Nye, J.L. (2003, March). Leadership perceptions in the Thick of it: Followers’ prototypes are important. Poster session presented at the annual meeting of the Eastern Psychological Association, Baltimore, MA.


