

**Effects of Dual Accountability and Purpose
of Appraisal on Accuracy**

Rachel L. Fredholm

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University in
partial fulfillment of the requirements for the degree of

Master of Science

in

Psychology

Neil M.A. Hauenstein, Chair

Jeffrey D. Facticeau

Roseanne J. Foti

November 2, 1998

Blacksburg, Virginia

Keywords: Accountability, Appraisal Purpose, Accuracy

Copyright 1998, Rachel L. Fredholm

Effects of Dual Accountability and Purpose
of Appraisal on Accuracy

Rachel L. Fredholm

(ABSTRACT)

This study investigated the effects of accountability and purpose of appraisal on rating and behavioral accuracy. Subjects viewed a videotape of a lecture and were asked to rate the lecturer's performance. Accountability to the ratee (the GTA on the videotape) was held constant. Accountability to a supervisor (a faculty representative) was manipulated such that subjects in the no (supervisor) accountability condition anticipated a meeting with the GTA only; subjects in the weak (supervisor) accountability condition anticipated a meeting with the GTA as well as a supervisory review of the ratings; while subjects in the strong (supervisor) accountability condition were led to believe that they would have to meet with both the GTA and the faculty representative to explain their ratings. Additionally, participants were led to believe that the purpose of this appraisal was either to provide feedback for development or to make administrative decisions.

Two-way ANOVAs were used to assess the effects of accountability and purpose of appraisal on rating accuracy (elevation accuracy, dimensional accuracy, leniency) and behavioral accuracy. Results indicated that (a) increased accountability to a supervisor led to greater elevation accuracy, (b) raters in the administrative purpose condition provided more lenient ratings than did raters in the developmental purpose condition, (c) behavioral accuracy increased with level of accountability (none, weak, strong) to a supervisor, (d) raters who believed that the purpose of appraisal was for development exhibited greater behavioral accuracy than did raters who believed that the purpose was to make administrative decisions.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Neil Hauenstein. Neil, I consider you both a mentor and a friend. Without your guidance, this may have taken a lot longer (can you imagine??). Seriously, I have a great deal of respect for you as a researcher and as a person. Thank you so much for your contributions to this project.... for listening.....and for keeping slam-dancing alive at Virginia Tech.

Next, I would like to thank Jeff Facticeau and Roseanne Foti. Jeff, somehow you made research more fun. Your enthusiasm and jokes allowed all of us to become part of the family, especially around those September SIOP deadlines. I owe you a special thanks for your contributions to this project, and I am particularly grateful that you stayed on my committee.

Roseanne, since the first day I came to Tech, I have admired and looked up to you as a role-model in this department. It is no secret that you are the reason that so many students in our department chose this program. Thank you also for the insight that you contributed to this project.

Although each person that I have met along this journey has influenced me in some way, there are a few people that I must acknowledge. First, David, you were an inspiration to me in this process. You pushed me when I didn't want to be pushed and always had a shoulder there for me when I needed one. More importantly, you have stood by me through a seemingly endless period of soul-searching. You have been and always will be my rock.

Next, Wendy, I am not sure where to begin. I know I said I thought that acknowledgments were cheesy, but it is an opportunity for me to tell you what an amazing friend you have been and always will be. In all my life, I don't think I have ever known anyone quite like you. You are the most unconditional, true friend that anyone could ask for. I will never, ever forget all of the laughs and secrets that we have shared. I look forward to all of the ones to come.

I can't forget my other closest compadres, Trina, Niki, Tim, Kevin Bradley, Kevin Basik, Theresa, and Christi. I will forever treasure the times we shared in this town and the many others we traveled to together.

Finally, a very special thank you goes out to my family. You have stood by me through all of my endeavors and loved me unconditionally. For this I will be forever grateful.

TABLE OF CONTENTS

Chapter 1. Introduction	1
Chapter 2. Accountability.....	6
2.1. Accountability in a Performance Appraisal Context.....	9
2.2. Accountability to Whom?	11
2.3. Dual Accountability	13
Chapter 3. Purpose of Appraisal	14
Chapter 4. Overview of Study.....	15
4.1. Hypotheses	17
Chapter 5. Method.....	17
5.1. Participants.....	17
5.2. Stimulus Materials	18
5.3. Procedure.....	19
5.4. Independent Variables.....	20
5.5. Dependent Variables	22
5.6. Pilot Test	24
Chapter 6. Results	25
6.1. Pilot Test	25
6.2. Rating Accuracy.....	26
6.3. Behavioral Accuracy	28
Chapter 7. Discussion.....	29
Chapter 8. References.....	35

Appendix A	46
Appendix B	47
Appendix C	48
Appendix D	49
Chapter 9. Vita	50

LIST OF TABLES

1. Target Scores for Rating Form.....	41
2. Means and Standard Deviations for Pilot Test Questionnaire	42
3. ANOVA Table for Pilot Test Questionnaire.....	43
4. Means and Standard Deviation for Accuracy Measures	44
5. ANOVA Table for Accuracy Measures	45

Effects of Dual Accountability and Purpose of Appraisal on Accuracy

INTRODUCTION

Performance appraisal research has received a great deal of attention over the past few decades (Harris, Smith, & Champagne, 1995). This interest has likely been inspired by the prevalence of performance appraisals in organizations. Survey research in the 1970's and 1980's suggested that between 74% and 89% of business organizations had a formal appraisal system and that state governments and larger organizations were even more likely to employ formal appraisal systems (Murphy & Cleveland, 1995). As a result, it has been concluded that performance appraisal seems to be nearly universal in organizations (Cleveland, Murphy, & Williams, 1989; Murphy & Cleveland, 1995). In the past 30 years, the uses of performance appraisal have evolved a great deal. The earliest use of performance appraisal was primarily to make administrative decisions such as promotions, salary increases, and terminations. However, since the 1960's, performance appraisal has been increasingly used for employee development and feedback, corporate planning, legal documentation, systems maintenance, and research (Murphy & Cleveland, 1995).

Due to the widespread use of performance appraisal in organizations, much of the empirical research on performance appraisal is dominated by concerns about accuracy as the primary criterion of interest (Illgen, Barnes-Farrell, & McKellin, 1993; Murphy & Cleveland, 1995). Although a great deal of research has examined factors that influence rating accuracy, limited progress has been made in improving rating accuracy (Mero & Motowidlo, 1995). Contributions to research on improving rating accuracy fall within three general categories: rating scale formats, rater cognitive processes, and rater motivation.

Rating format research focused on the structure of appraisal by influencing the type and number of dimensions assessed, the types of judgments made, appraisal length, and comprehensiveness. Researchers developed numerous formats such as graphic rating

scales, behaviorally anchored rating scales (BARS), mixed standard scales, and forced-choice scales (Borman, 1979).

Borman (1979) examined the effects of rating format on performance rating accuracy. He found significant effects of format on rating accuracy, but a strong Job X Format interaction led to the conclusion that no single format was better or worse overall than the other formats. Rather, each format may have strengths and weaknesses in relation to specific jobs.

Landy & Farr (1980) reviewed the effectiveness of the graphic rating scale and other rating formats developed as alternatives to the graphic rating scale. They found that none of the formats were more efficient or psychometrically sound than the graphic rating scale. Additionally, they stated that format research provides limited contributions to the study of performance appraisal. From this research, they concluded that the rater should have a clear understanding of the rating process, the number of response categories should be limited, the anchors on the scale should be rigorously developed, and the labels should be more than simple, descriptive labels. However, even considering these factors, they found that rating format only explained 4% - 8% of the variance in performance ratings. As a result, they called for a moratorium on rating format research. They suggested that future research should focus on understanding the cognitive processes involved in performance rating.

As a result, much of the appraisal research since the 1980's has focused on the cognitive processes involved in performance appraisal (e.g., Feldman, 1981; Foti & Hauenstein, 1993; Hauenstein & Alexander, 1991; Lord, 1985; Murphy, Balzer, Lockhart & Eisenman, 1985; Nathan & Lord, 1983). This approach focuses on how judgments are made and retained for use in performance appraisal. Specifically, the cognitive process research in performance appraisal has focused on the acquisition, encoding, storage, and retrieval processes involved in making performance judgments.

Rater training (e.g., Bernardin & Buckley, 1981; Bernardin & Pence, 1980; Hauenstein & Foti, 1989; Sulsky & Day, 1992) has been another cognitive approach to improving the accuracy of ratings. The goals of rater training research are to promote proper utilization of appraisal systems and to improve rating skills. Frame-of-reference

(FOR) training, which provides the rater with normative standards and behavioral examples, has been effective at improving the accuracy of performance ratings (e.g., McIntyre, Smith, & Hassett, 1984; Sulsky & Day, 1992). However, Stamoulis and Hauenstein (1993) found that FOR training improves the accuracy of between-dimension discrimination but does not necessarily improve the accuracy of between-individual discrimination.

In a review of performance appraisal research and rating accuracy, Illgen et al (1993) concluded that cognitive process research has reached a point of diminishing returns. Additionally, they stated that the amount of variance that cognitive processes account for in appraisal ratings is limited and that expanding our rating models beyond cognitive processing variables may advance the field more quickly.

Research on both rating formats and rater cognitive processes has failed to address the difference between ability and motivation to rate accurately. Banks and Murphy (1985) argue that most research to date has focused on the ability to rate accurately, rather than motivation to make accurate ratings. Further, most laboratory studies on performance appraisal have only examined the judgment aspect of the process, whereas appraisals in organizations also include a rendering process that takes place before the marking of the appraisal form. Laboratory studies typically fail to consider motivational factors involved in the rendering process such as conflict avoidance, personal and political agendas, and financial need (Banks & Murphy, 1985; Murphy & Cleveland, 1995). As a result, the recorded ratings in lab studies are often equivalent to the raters' evaluations, and the rendering process is eliminated. Banks and Murphy (1985) suggested that laboratory research should consider the motivational context of performance appraisal involved in the rendering process.

Some research has focused on the motivational context of appraisal. In one such study, Salvemini, Reilly, and Smither (1993) examined the influence of motivation on rating accuracy by providing raters with incentives to be accurate. Raters were told that 1) the ratees had received below average ratings in a previous position, 2) above average ratings in a previous position, or 3) no previous rating information. They found that the expectation of incentives (monetary rewards) reduced bias and increased rating accuracy.

However, this study focused solely on the judgmental aspect of the appraisal process. Raters knew that their ratings had no consequences for the ratees and that they would not have to interact with the ratees in the future. As a result, there was no reason for raters to distort their ratings.

In another study that explored the motivational aspects of appraisal, Longenecker, Sims, & Gioia (1987) found that executives believed there was actually a justifiable reason for generating appraisal ratings that were inflated or deflated. Generally, the executives felt that it was within their discretion to knowingly distort ratings. They did not mind distorting because they did not see these distortions as errors. They saw the distortions as "discretionary actions that help them manage people more effectively" (p.190). Further, Longenecker, et al. (1987) concluded that the formal appraisal process is a political process and that few ratings are made without political consideration. One of the primary reasons given for providing distorted ratings was avoidance of unnecessary conflict. Specifically, many executives said that they inflated ratings to avoid confrontation with a subordinate with whom they had recently had difficulties.

Since rater motivation is an important aspect of the appraisal process, it is important to consider the factors that influence this motivation. The three determinants of rater motivation, as outlined by Harris (1994), are perceived rewards, perceived negative consequences, and impression management. Certain situational variables are suggested to influence these determinants of rater motivation. One of these situational variables is accountability. According to Harris (1994), accountability is the most complex of these situational variables and its specific effects depend on a host of other factors. Additionally, most organizations that use performance appraisal make raters accountable for their ratings in some fashion (Murphy & Cleveland, 1995).

Accountability

Accountability is the pressure to justify one's ratings to another (Tetlock, 1983a). Typically, the concept of accountability has been researched in the context of attitudes towards social issues (Tetlock, 1983a, Tetlock, 1983b). Tetlock (1983a) examined the impact of accountability on the complexity of people's appraisal of controversial social

issues. Complexity consisted of two parts, differentiation and integration. Differentiation referred to the number of characteristics or dimensions of a problem that an individual took into account. Integration involved recognizing interrelationships among issues. Therefore, differentiation was a necessary condition for integration. Subjects were asked to report their thoughts on three issues and respond to attitude scales relevant to each topic. They were assigned to one of four accountability conditions: expecting their attitudes to be anonymous or expecting to justify their views face-to-face to a person with liberal, conservative, or unknown views.

Tetlock (1983a) found that accountability led people to engage in more complex information processing only if they did not know the views of the person to whom they were accountable. Otherwise, they simply took the "lazy" option of expressing views congruent with those of the person to whom they were accountable.

Tetlock (1983b) later explored the hypothesis that accountability leads to more vigilant information processing strategies. Subjects read a brief description of a murder trial, followed by evidence about the defendant's guilt. The evidence was presented in one of three ways. One group viewed evidence of guilt followed by evidence of innocence. The second group received the evidence in reverse order. A third group read the evidence in random order. This manipulation produced either a guilt, an innocence, or no primacy effect, respectively. Additionally, one of the three groups was not held accountable for their "likelihood of guilt" rating, while the other two groups were. To examine the effects of accountability on encoding, information recall and response biases in guilt perceptions, accountable subjects were either informed of their accountability before (pre-exposure accountability) or after (post-exposure accountability) reading the evidence.

The results of the study suggested that the order of presentation of the evidence (guilt vs. innocence) only influenced perceptions of guilt in non-accountable subjects and post-exposure accountable subjects. Additionally, presentation order did not influence the number of facts about the case recalled. However, pre-exposure accountable subjects recalled more specific information about the case than did non-accountable or post-exposure accountable subjects. Specifically, the subjects in these conditions may have

recalled the same number of facts, but subjects in the pre-exposure accountable condition included more detail in their stated facts than did non-accountable or post-exposure accountable subjects. Therefore, it appears that accountability is effective at producing more vigilant information processing by affecting the encoding process. As a result of this vigilant information processing, early-formed impressions have less influence on final judgments.

Tetlock (1985) proposed a social contingency model to examine the effects of accountability on choice. The model suggests that people are "cognitive misers" (Fiske & Taylor, 1984) who will exert the minimum amount of effort necessary to reach a decision. According to Tetlock (1985), people often avoid unnecessary cognitive work by relying on the acceptability heuristic, the salient "acceptable" option.

Tetlock's (1985) model applies to four accountability situations. First, in situations where an individual is not accountable for his or her position, the model predicts that the individual will put little effort into developing a justification for his or her decision. Second, in situations where the subject knows the opinion of his or her audience and has not made a previous commitment to a position, he or she will also put minimal effort into justification. Instead, these individuals will rely on the acceptability heuristic, comparing the pros and cons of each position, and choose the socially acceptable, or audience, position. This approach makes justification less cognitively demanding. Therefore, both of these strategies avoid unnecessary cognitive effort.

However, we are not always aware of the attitudes or opinions of our audience. As a result, when the audience's opinion is unknown and individuals have not previously committed to a position, they often engage in preemptive self-criticism. In this situation, individuals try to anticipate potential criticisms of their decision in order to defend it to the audience. While this would be an effective strategy in the previous situations as well, it requires a great deal more cognitive effort and is generally only used when people feel it is necessary to defend their views.

Other times, individuals must defend their previous decisions to an audience. Because we wish to appear consistent (Cialdini, 1995), the task becomes one of justification rather than decision making. Therefore, individuals must engage in

retrospective rationality, where they attempt to develop ideas that justify their position. However, this process is also cognitively demanding, and does not apply to the previous situations.

Other research has examined the effects of accountability on the judgment process. For example, Hagafors and Brehmer (1983) examined the effects of accountability on the application of judgment policies under conditions of high and low task predictability and provision versus no provision of feedback. The authors found that when the task was not predictable and feedback was not provided, accountability led to greater consistency in judgment policies. The explanation suggested for these findings was that accountability leads to a more analytical judgment process.

Accountability in a Performance Appraisal Context

For the most part, the social contingency model of judgment has been tested in relation to attitudes and thoughts towards social issues. However, recently the concept of accountability has been applied to performance appraisal situations in an organizational context (e.g., Klimoski & Inks, 1990; Simonson & Nye, 1992). From a social-cognitive perspective, it is almost impossible to find context-free cognitions (Tetlock, 1985). Contextualists argue that research should involve more ecologically representative situations. Focusing on the situational characteristics that motivate raters during performance appraisal may improve rating accuracy research.

For example, Klimoski and Inks (1990) examined the effects of accountability and knowledge of subordinate self-appraisal on performance rating quality. They suggested that situations in which a rater anticipates face-to-face feedback sharing will promote greater accountability than situations in which performance feedback is not to be given in a face-to-face meeting. Subjects received either favorable, unfavorable, or no self-appraisal information and expected either face-to-face (strong), written (weak), or no accountability with a subordinate. The authors found main effects for accountability type and level of subordinate self-appraisal. Performance ratings of supervisors who were to be held accountable in a face-to-face context rated their subordinates significantly more positively than did supervisors who were not held accountable for their ratings. Additionally, supervisors who received positive self-appraisal information rated their

subordinates the most positively, and raters who received negative self-appraisal information rated their subordinates the most negatively.

Simonson and Nye (1992) further presented evidence that accountability effects may increase pressure to maintain a positive social image. In a decision making context, it was shown that accountable subjects were less likely to exhibit the sunk cost effect, a common judgment error in management. However, accountable subjects were not more likely than non-accountable subjects to consider all the available and relevant information in making their decision, as would be expected. Additionally, regardless of accountability condition, more thorough information processing did not lead to better decisions. Simonson and Nye concluded that the variable underlying these results was the decision maker's desire to avoid criticism. When the socially desirable decision was also the most accurate decision, it was impossible to separate the effects of impression management from those of information processing effort.

In another recent study, Mero and Motowidlo (1995) developed videotaped vignettes of good and poor subordinate performance and asked subjects to evaluate the performances as part of an in-basket exercise. Subjects were either given no previous performance information, or were told that all previous ratings had been too low, too high to discern performance differences, or that women had consistently been rated lower than men. Thus, subjects were instructed to correct for these errors in the previous ratings by inflating all of their ratings, by being as accurate as possible, or by inflating only women's ratings, respectively. Additionally, subjects were informed either that they would have to justify their ratings to the experimenter or that their ratings would be anonymous. Results indicated that accountable subjects were more careful in making their ratings. Accountable subjects were more attentive to the task, more engaged in the task, and took more and better notes.

More importantly, however, the applicability of Tetlock's (1985) social contingency model of judgment to performance ratings was supported. Accountable subjects who received no information about their "supervisor's" desired rating produced the most accurate performance ratings. Additionally, with the exception of the equitable treatment context, accountable raters were more likely to rate the vignettes (to distort or

make accurate ratings) in accordance with the social pressure they experienced. The lack of an accountability effect in the equitable treatment context was attributed to the difficulty in justifying such ratings.

Accountability to Whom?

It could be argued that being held accountable for ratings to a subordinate ("downward accountability") is akin to being accountable to someone with known views. Ratees think highly of themselves and want to receive good ratings. This can be substantiated by findings in the self-appraisal literature. Harris and Schaubroeck (1988) conducted a meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. They found that self-ratings were over a half standard deviation higher than supervisor ratings and one-quarter of a standard deviation above peer ratings.

Many researchers have found that "downward accountability" leads raters to inflate performance evaluations (e.g., Fisher, 1974; Hauenstein, 1992; Klimoski & Inks, 1990). Fisher (1974) was one of the first researchers to demonstrate this phenomenon. Her study revealed that when supervisors were held accountable to their subordinates for their ratings and subordinate performance was poor, ratings were more lenient than ratings of the same subordinates in a non-accountable situation. However, no distortion effects were found when subordinate performance was good.

As research has shown, inflation is often used to avoid conflict when employee performance is poor (Klimoski & Inks, 1990; Longenecker, et al, 1987). Pearce and Porter (1986) have demonstrated that "satisfactory" ratings also lead to negative employee reactions. Specifically, Pearce & Porter (1986) found that "satisfactory" employees experienced a significant and stable decrease in organizational commitment after the implementation of a formal appraisal procedure. As a result, managers may also inflate "satisfactory" ratings in order to avoid conflict.

Expectations of justifying appraisals to a supervisor ("upward accountability") might have a different effect on performance ratings. Recently, Longenecker (1992) suggested that raters' supervisors possess great influence on the quality of performance ratings. The author explained that accountability to the raters' supervisors may lead to decreased rater biases.

Corrigan (1994) investigated the effects of "upward accountability" on performance ratings. They found that subjects who were accountable to their supervisor for their ratings and were told that the supervisor believed the subordinate's performance was "poor and ineffective" were less accurate than the no accountability and/or no view raters. However, no differences were found between the accountability/no view and no accountability/no view conditions. This may have been the result of a weak manipulation of accountability. The subjects were held accountable to their instructor who supervised a class of 1200 students. Also, subjects were not told that their ratings would be challenged. As suggested by Harris (1994), expecting the supervisor to challenge ratings is one way to increase accountability to the supervisor.

Further, Harris (1994) suggested that increased accountability to subordinates will decrease rater motivation to make accurate ratings, but accountability to one's supervisor will increase rater motivation to increase processing and make more accurate ratings. Harris suggests that there are two factors that are likely to increase the accountability a rater has to his or her supervisor: 1) the extent to which the rater perceives that the supervisor will carefully review the ratings 2) the likelihood that that the supervisor will challenge the ratings. When raters expect to justify their ratings to the supervisor and they are not aware of the supervisor's beliefs or preferences, motivated cognitive processing is more likely to occur so raters can prepare for criticism and appear more competent.

Dual Accountability

Some research has suggested that "downward" and "upward" accountability should be examined together. For example, Harris (1994) suggests that "how a rater deals with conflicting 'audiences' and justifies subsequent ratings and feedback is an interesting topic for future research" (p.745). Further, Longenecker, et al. (1987) found that when managers believed that appraisals would be seriously scrutinized and reviewed by their supervisors, the likelihood of intentionally distorting subordinate ratings was reduced.

Some researchers (Fandt & Ferris, 1990; Harris, 1994) suggest that increased accountability to one's supervisor should increase the accuracy of ratings. However,

because raters are typically held accountable to their subordinates, this would create a situation of "dual accountability". Due to the conflict that this situation creates, it is likely that raters will rely on other factors when making ratings in a dual accountability context. One factor that is typically salient and influences rater motivation is the purpose of the appraisal.

Purpose of Appraisal

Another factor that has been shown to affect rating accuracy is the purpose of the appraisal. Cleveland and Murphy (1992) suggested that ratings that are "used for one purpose may not (under similar circumstances) yield the same outcome when the appraisal system is used for a different purpose" (p. 138). Research has suggested that appraisal purpose may affect the standards used in making ratings, the cognitive processes involved in making ratings (e.g., observation, recall), and the integration of information in making ratings (e.g., Murphy, Balzer, Kellam, & Armstrong, 1984; Reilly & Balzer, 1988). In fact, it has been suggested that purpose of appraisal may be the single most important contextual factor for understanding performance appraisal processes and outcomes (Jawahar & Stone, 1997).

Several studies have shown that ratings collected for administrative purposes, such as salary administration, are significantly higher than ratings of the same individuals collected only for feedback or research purposes (e.g., Harris, et al, 1995; Reilly & Balzer, 1988; Zedeck & Cascio, 1982). One possibility for this difference is that raters feel more concern for the ratee if the consequences of the ratings are more severe (Murphy et al, 1984; Reilly & Balzer, 1988). If a performance rating is used to make decisions about salary increases or promotions, for example, the rater may feel more pressure to inflate. Conversely, if the rater is told that the ratings will only be used to give the employee useful feedback, the rater may not feel as much pressure to distort their ratings. Murphy and Cleveland (1995) state that when raters are told that the purpose of the appraisal is for feedback, inflation and range restriction are not serious problems. They suggest that this is because a supervisor does not experience much conflict when they believe that information from performance appraisals will only be used for providing feedback to employees. Further, they suggest that raters whose appraisals are used in

making administrative decisions are equally capable of making accurate ratings but refuse to do so.

Although rating purpose appears to influence rating accuracy, it is important to note that this effect has been stronger in field studies than in laboratory studies (Murphy & Cleveland, 1995). Typically, lab research produces larger effect sizes than field study research due to the highly controlled environments in lab research. However, this is not true for research on the effects of rating purpose on accuracy. One possible explanation for this may be that it is difficult to adequately simulate the pressures involved with rating for different purposes in an organization.

Overview of Study

The purpose of the present study was to examine the effects of dual accountability and purpose of appraisal on rating accuracy and behavioral accuracy. Since raters are typically held accountable to subordinates for their ratings (Murphy & Cleveland, 1995), subordinate accountability was held constant in this study. Of greater concern, was how the addition of accountability to a supervisor would effect performance appraisal accuracy. It was expected that subjects who experienced strong accountability (face-to-face feedback) to a supervisor would exhibit the highest levels of accuracy while those who were not held accountable to their supervisor at all should have provided the least accurate ratings (Klimoski & Inks, 1990).

In addition, ratings made for the purpose of development should have been more accurate than those made for the purpose of making administrative decisions because the subjects should not have perceived negative consequences for the ratee in the former situation. Further, it was expected that there would be a significant increase in accuracy in the developmental purpose condition when raters experienced even weak accountability to a supervisor. Conversely, a significant increase in accuracy for subjects in the administrative purpose condition was not expected until they experienced strong accountability to a supervisor. As a result, the weak accountability condition is where the greatest differences in accuracy were expected to be found. The reason for this distinction was that subjects in the administrative purpose conditions were expected to perceive more negative consequences for the ratee than were subjects in the

developmental purpose condition. Consequently, subjects in the administrative purpose condition would need more pressure from another source to motivate them to provide more accurate ratings.

Finally, raters are often asked to rely on their memories of behaviors because rating accuracy is thought to be a function of the accuracy of behavioral memories (Sanchez & De La Torre, 1996). Rating accuracy and behavioral accuracy were examined separately in this study to ensure that performance ratings were not being made based on false memories. It was hypothesized that the patterns of behavioral accuracy should be similar to those of rating accuracy, based on the assumption above.

Hypotheses

H1. Raters in the strong accountability condition will exhibit the greatest level of rating accuracy while those in the no (supervisor) accountability condition will exhibit the lowest level of rating accuracy.

H2. Rating accuracy will be higher when raters believe the purpose of appraisal is for employee development than when they believe the purpose of appraisal is to make administrative decisions.

H3. A two-way interaction will occur between accountability and purpose of appraisal. The difference in rating accuracy between the administrative and developmental purpose conditions will be greatest when subjects experience weak accountability.

H4. Raters in the strong accountability condition will exhibit the greatest level of behavioral accuracy while those in the no (supervisor) accountability condition will exhibit the lowest level of behavioral accuracy.

H5. Behavioral accuracy will be higher when raters believe the purpose of appraisal is for employee development than when they believe the purpose of appraisal is to make administrative decisions.

H6. A two-way interaction will occur between accountability and purpose of appraisal. The difference in behavioral accuracy between the administrative and developmental purpose conditions will be greatest when subjects experience weak accountability.

Method

Participants

At least 120 female students, 20 per cell, enrolled in an introductory psychology course at Virginia Polytechnic Institute and State University were randomly assigned to one of six conditions in a 2 (purpose of appraisal) X 3 (accountability) design. Female subjects alone were used in the present study because research has shown that females exhibit greater accountability effects (Stamoulis, 1993). In exchange for participation, subjects received one extra credit point in their psychology classes.

Stimulus Materials

A 15 minute videotape depicting a graduate teaching assistant (GTA) lecturing on the topic of consumer psychology served as the stimulus in the current study. The tape consists of 16 behavioral incidents (see Appendix A) corresponding to four performance dimensions: depth of knowledge, organization, delivery, and relevance. Each dimension is represented by four behavioral incidents. The quality of the GTA's performance on each dimension is determined by the number of good and poor behavioral incidents representing the dimension. Four good behaviors correspond to the depth of knowledge and organization dimensions, and two good and two poor incidents correspond to the delivery and relevance dimensions. The videotape is based on a videotape originally developed by Hauenstein and Alexander (1991). A different tape was made in order to update the clothing of the GTA so subjects will be more likely to believe that the GTA in the video is currently a GTA at Virginia Tech.

Additionally, new target scores were developed by 12 graduate students in Industrial/Organizational Psychology at the university. These raters served as expert raters due to their knowledge of the psychometric criteria used in performance appraisal studies. Mean ratings for each performance dimension were obtained from the graduate students and represent the target performance scores in this study (see Table 1). The quality of the target scores were assessed by examining the convergent validity of the scores (Kavanagh, MacKinney, & Wolins, 1971). The resulting intraclass index was .61, indicating that there was sufficient agreement (convergent validity) of raters across dimensions.

As stated earlier, four good behaviors correspond to the depth of knowledge and organization dimensions, and two good and two poor incidents correspond to the delivery and relevance dimensions. As a result, ratings were higher for the depth of knowledge (5.0) and organization (5.25) dimensions than for the delivery (3.42) and relevance (4.17) dimensions. The means from the Hauenstein and Alexander (1991) tape were very similar: 4.88, 5.16, 3.48, and 4.20 respectively.

The behaviors on the Hauenstein and Alexander (1991) tape were previously allocated to the appropriate dimensions through a behaviorally anchored rating scale procedure (Nathan & Lord, 1983). This videotape was chosen because it has been found to depict average performance (Nathan & Lord, 1983; Hauenstein & Alexander, 1991; Hauenstein, 1992; Foti & Hauenstein, 1993). Average performance was chosen to enable raters to distort ratings in both directions and avoid ceiling or floor effects in subjects' ratings.

Procedure

All subjects were told that the purpose of this study was to obtain evaluations of the GTAs in the psychology department at this university. It was explained that they were chosen to participate because in-class student evaluations are often biased. Further, subjects were provided with information regarding the specific purpose of the evaluations (for developmental purposes, or to assist in making administrative decisions, based on the assigned condition). Additionally, subjects were told before viewing the video that they would have to attend feedback meetings with the GTA (in all conditions) and a faculty

representative (in the strong accountability condition). They were informed that they must attend this meeting in order to receive full credit for participation in this study. The available times were approximately one week from the day of the experiment. Subjects were asked to schedule appointments for the face-to-face feedback meetings at this time.

Next, subjects were told that they would be rating one of the GTAs at random based on their impressions of a short lecture that the GTA was required to write and present for the purpose of these evaluations. Although subjects were told that they would be randomly assigned to rate one of the GTAs, all subjects viewed the same presentation.

Then, subjects viewed the videotape. Afterwards, they were asked to fill out their opscans for extra-credit. They were told that the experimenter forgot to have them fill these forms out initially. This procedure was used to control for the effects of short term memory. Finally, they filled out the performance rating form (see Appendix B), and a recognition-memory questionnaire (see Appendix C). The recognition-memory questionnaire was used to assess the degree to which subjects experienced false memories or response bias. Response bias refers to one's tendency to over or underattribute behaviors to the target person. Recent research has suggested that recognition memory tests are meaningful because of the usage of behavioral anchors and behavioral incidents in performance-rating tasks (Sulsky & Day, 1992; Sanchez & De La Torre, 1996).

Approximately one week after the subjects have made their performance ratings, they were called and told that there would not be a feedback meeting. The purpose of delaying this information was to minimize the possibility of subjects telling future participants that these meetings would not be held. Additionally, subjects were debriefed about the true purpose of the study, and they had an opportunity to ask questions.

Independent Variables

The presentation order of the two independent variables in this study was counterbalanced. This procedure was followed to ensure that the order of presentation of the independent variables did not strengthen the manipulation of one independent variable more than the other.

Accountability. Subjects were informed that they would be required to meet with the GTA (in all conditions) and a faculty member representing the department (in the

strong accountability condition), for 10 minutes each, to explain and justify their ratings. They were told that the GTA and, in the strong accountability condition, the faculty member would carefully review the ratings. They were also told to be prepared to defend their ratings in case the ratings were challenged. These were the two suggestions given by Harris (1994) to increase the accountability a rater experiences. Subjects in the weak accountability condition were told that their ratings (with their names on them) would be presented to a faculty representative for review. In the no accountability condition, subjects were not told anything about a supervisory review of ratings.

Purpose of Appraisal. Subjects received specific information about the purpose of the evaluations they were asked to complete (based on their assigned condition). Subjects in the developmental purpose condition were told that the purpose was to provide GTAs with feedback that they could use to improve their teaching skills in the future. However, subjects in the administrative purpose condition were told that their evaluations would be used to make decisions about assistantships for the upcoming school year. Specifically, they were told that these evaluations would be used to determine whether GTAs would be assigned a teaching or non-teaching assistantship. Further, subjects were told that the teaching assistantships are more desirable to the GTAs than the non-teaching assistantships and that the teaching assistantships typically involve a larger stipend.

Dependent Variables

Rating accuracy. All subjects completed a graphic rating scale (see Appendix B) developed by Hauenstein (1992). This measure consists of six 7-point scales ranging from 1 (poor) to 7 (excellent). The first four scales require the rater to assess the ratee's performance on the following dimensions: depth of knowledge, delivery, relevance, and organization. The last two scales ask the rater to make a rating based on their overall assessment of the ratee's performance. First, rating accuracy scores were computed for subjects in each condition. Sulsky and Balzer (1988) stated that the primary advantage of using accuracy scores is that they provide a direct, rather than indirect, measure of accuracy. Further, unlike traditional rater error measures, accuracy measures make no assumptions regarding the actual distribution of ratee performance.

The present study will employ multiple accuracy measures. Murphy and Cleveland (1995) discussed the importance of using multiple accuracy measures because "it is disconcerting to note that one's results may depend more on the choice of accuracy measures than on the phenomenon being studied" (p. 289).

Accuracy measures were obtained through the employment of the single ratee accuracy measure developed by Hauenstein and Alexander (1991). This measure consists of two components. The first is labeled elevation because it is analogous to Cronbach's (1955) elevation component of accuracy. Perfect elevation accuracy requires a rater's average observed rating to equal the average of the true scores.

$$E^2 = (x.-t.)^2$$

The second component of this measure is dimensional accuracy. Perfect dimensional accuracy demands both a correlation of positive one between a rater's observed ratings and the true scores and that a rater's variance equals the variance of the true scores. Dimensional accuracy is conceptually similar to Cronbach's (1955) differential accuracy as it measures the accuracy with which each rater evaluated a single ratee on each dimension. However, in the case of a single ratee, "differential" accuracy does not apply.

$$DA^2 = 1/n \sum [(x_j-x.)-(t_j-t.)]^2$$

In addition to the single-ratee accuracy measure, a leniency/severity measure developed by McIntyre, Smith, & Hassett (1984) will be employed. Although previous research has often measured leniency/severity by accounting for deviations of actual ratings from the midpoint of the scale, this measure accounts for deviation of actual ratings from true score ratings. This accuracy measure was used to determine the direction of distortion among inaccurate ratings.

$$L = \frac{\sum (T_{ij}-R_{ijk})}{d}$$

Behavioral accuracy. Subjects completed a 32-item recognition memory questionnaire (see Appendix C) immediately after they completed the performance ratings. Eight behaviors were provided for each of the four dimensions. Half of the items were behaviors that did not occur in the presentation, and the remaining behaviors were those that were present in the lecture. The questionnaire consists of 24 good behaviors and 8 poor behaviors because more good behaviors than poor behaviors occurred in the presentation. Subjects were instructed to read each behavior and respond yes or no to whether they believed that behavior occurred in the lecture.

Behavioral accuracy was then assessed based upon subjects' responses to the recognition memory questionnaire. Each subject's responses were analyzed through his or her false-positive rate and true hit rate. The false-positive rate is the percentage of nonoccurring behaviors incorrectly identified. The true hit rate is the percentage of correctly identified occurring behaviors. To determine a rater's behavioral accuracy, their false positive rate was simply subtracted from their true hit rate (Sulsky & Day, 1992).

Pilot test

Prior to running the actual study, pilot testing was completed to ensure that the subjects believed that their ratings would be used to evaluate the GTAs and that they believed that they would actually be meeting with the ratee and a faculty representative. Subjects viewed the videotape stimulus used in the actual study. They followed the procedures as outlined for the actual study except that they completed a manipulation check questionnaire after completing the rating form, and they did not complete the recognition memory questionnaire. On the manipulation check questionnaire (see Appendix D), subjects were asked a series of questions such as "To what extent did you feel that the feedback meeting(s) would take place?" and "To what extent did you feel accountable to the GTA?" and "To what extent did you feel accountable to a faculty representative from the Department of Psychology?" as well as "To what extent did you consider the purpose of these evaluations?". Subjects responded using a 7-point likert scale ranging from "to no extent" to "to a great extent". Additionally, subjects were asked "Did you feel that the pressures created by the accountability you experienced and the

purpose of the evaluation were equal?". They were asked to explain their answer to this question.

Further, brief meetings were held with subjects in the pilot test after they completed the questionnaires. In this meeting, subjects were asked to provide feedback regarding suggestions for improving the believability of the deceptions in this study. Also, subjects were asked how they resolved the conflicting pressures created by accountability and the purpose of the evaluations.

Finally, a manipulation check for purpose of appraisal was not included in the pilot study because the difference in accuracy due to purpose is one of the most robust findings in the performance appraisal literature (Jawahar & Stone, 1997; Murphy & Cleveland, 1995). Of greater concern was the strength of the accountability manipulation.

Results

Pilot test

Four ANOVAs were used to assess the impact of the experimental manipulations (accountability and purpose of appraisal) on the participants' questionnaire responses (see Table 3). For the first question, ("To what extent did you feel that the feedback meetings would take place?"), subject responses indicated that they did believe that the meetings would take place ($M=6.07$). Additionally, there were no effects found for this dependent variable, suggesting that the manipulation was not significantly more believable for subjects in any particular condition.

In response to the second question ("To what extent did you feel accountable to the GTA?"), subjects indicated that they did feel accountable to the GTA ($M=5.72$). As expected, there were no effects found for accountability to the GTA, indicating that subjects did not feel significantly more or less accountable to the GTA depending on their experimental condition.

Next, a main effect of accountability was found for the following question ("To what extent did you feel accountable to a faculty representative for the Department of Psychology?") $F(2, 57) = 17.39$ $p < .001$. Although subjects in the no supervisor

accountability condition expressed the lowest level of accountability to a faculty representative ($M=3.0$), accountability to a faculty representative increased with level of accountability (weak $M=3.7$, strong $M=5.45$).

Finally, there were no effects found for the last question ("To what extent did you consider the purpose of these evaluations?"). This indicated that subjects did not consider the purpose of the evaluations significantly more or less based on their condition. The overall mean for this variable was 5.62.

Rating Accuracy

First, the hypotheses for rating accuracy were tested with three two-way ANOVAs (see Table 5). Although evidence was found to support Hypotheses 1 and 2 (main effects), no evidence was found to support Hypothesis 3 (interaction).

Elevation Accuracy.

Elevation accuracy was scored so that lower numbers indicate greater elevation accuracy. First, in support of Hypothesis 1, a main effect of accountability was found for elevation accuracy $F(2, 117) = 3.824$ $p < .05$ (see Table 5). Subjects in the no supervisor accountability condition exhibited the least accurate ratings ($M=1.1397$), while subjects in the weak accountability ($M=.7688$) and strong accountability ($M=.7771$) conditions provided more accurate ratings. Although there was a main effect for accountability, it was not expected that ratings in the weak accountability condition would be higher (if only slightly) than those in the strong accountability condition.

Since an effect was found for accountability, a post-hoc test was used to examine where the differences were. Results of the Tukey's LSD revealed that elevation accuracy for the no supervisor accountability condition was significantly different from both the weak and strong supervisor accountability conditions. However, the weak and strong accountability conditions did not differ significantly from one another.

Unfortunately, no support was found for Hypotheses 2 or 3 with respect to elevation accuracy. However, from looking at Figure 1, it appears as though an interaction occurred. The greatest difference in accuracy did occur in the weak accountability condition. As a result, a t-test was performed to determine if there was a significant difference in elevation accuracy between the administrative purpose and

developmental purpose conditions for weak accountability subjects. Indeed, there was a significant difference between these two conditions.

It appears that the reason for the lack of an interaction effect was due to the variance in the cell means. Levene's test for Homogeneity of Variance revealed that the error variance for elevation accuracy was not equal across groups $F(5, 114) = 2.854$ $p < .05$. Consequently, the magnitude of the variance most likely prevented an interaction effect from occurring.

Dimensional Accuracy. Dimensional accuracy was scored so that higher numbers reflect greater dimensional accuracy. ANOVA results for dimensional accuracy revealed no effects (see Table 5). This is not surprising; however, because while perfect elevation accuracy requires only that a rater's average observed rating equal the average of the true scores, perfect dimensional accuracy demands both a correlation of positive one between a rater's observed ratings and the true scores and that a rater's variance equals the variance of the true scores. Dimensional accuracy examines a rater's tendencies by dimension, holding their overall tendencies (leniency and elevation accuracy) constant. It was not expected that raters would differ significantly, by condition, on their specific dimensional ratings.

Leniency. Leniency was scored so that higher positive numbers reflect greater leniency and negative numbers reflect rating severity. Finally, in support of Hypothesis 2, there was a main effect of purpose on the leniency of ratings $F(1, 118) = 27.019$ $p < .001$ (see Table 5). As a whole, subjects in the administrative decisions purpose condition provided lenient ratings ($M = .70$) while subjects in the developmental purpose condition provided somewhat severe ratings ($M = -.26$). Additionally, an effect of accountability (Hypothesis 1) was approaching significance $F(2, 117) = 2.467$ $p = .09$. However, no effect was found for Hypothesis 3.

Behavioral Accuracy

Behavioral accuracy was scored so that higher numbers reflect greater behavioral accuracy. A two-way ANOVA was used to assess the effects of the independent variables on behavioral accuracy (see Table 5). In support of Hypothesis 4, a main effect was found for accountability $F(2, 117) = 3.520$ $p < .05$. Subjects in the no supervisor

accountability condition demonstrated the lowest levels of behavioral accuracy ($M=.5879$), whereas subjects in the weak accountability condition exhibited even higher levels of behavioral accuracy ($M=.6875$), and subjects in the strong accountability condition demonstrated the highest levels of behavioral accuracy ($M=.6999$).

Since significant differences were found due to accountability, a post-hoc test was conducted to examine where the cell differences were. Results of the post-hoc test (Tukey's LSD) revealed that behavioral accuracy for the no supervisor accountability condition was significantly different from that of both the weak and strong supervisor accountability conditions. However, the weak and strong accountability conditions did not differ significantly from one another.

Additionally, in support of Hypothesis 5, a main effect of purpose was found on behavioral accuracy $F(1, 118) = 4.372$ $p < .05$. Overall, subjects in the developmental purpose condition demonstrated higher levels of behavioral accuracy ($M=.6980$) than did subjects in the administrative purpose condition ($M=.6189$).

Unfortunately, no support was found for Hypothesis 6 which predicted a two-way interaction. Although the greatest difference in behavioral accuracy between the administrative and developmental purpose conditions occurred when subjects experienced weak accountability to a supervisor, there was not a significant interaction.

Discussion

The results of the current study indicate that raters who believed that the purpose of the ratings was to make administrative decisions provided more lenient ratings than raters who believed that the purpose was solely for development. This result supports previous research findings that ratings made for administrative purposes were more lenient than those made for research or developmental purposes (e.g. Harris et al, 1995; Reilly & Balzer, 1988; Sharon & Bartlett, 1969; Zedeck & Cascio, 1982). As a whole, raters in the administrative purpose condition inflated ratings ($M=.70$) while raters in the developmental purpose condition provided more severe ratings ($M= -.26$).

Given this inflation problem, one possible solution may be to hold raters accountable for their ratings to a supervisor. The present study found that increased accountability to a supervisor led to more accurate overall subordinate ratings (greater

elevation accuracy). Elevation accuracy increased in a linear fashion with level of accountability to a supervisor (none, weak, strong) for raters in the administrative condition. However, elevation accuracy for subjects in the developmental condition greatly increased with weak accountability to a supervisor but decreased with strong accountability to a supervisor.

One possible explanation for this finding may be that in the developmental purpose condition, there was less pressure felt to inflate subordinate ratings because raters did not perceive negative consequences for the ratee. However, when the anticipation of a meeting with a supervisor was added, raters may have felt pressure to provide ratings that would be acceptable to the supervisor. Murphy and Cleveland (1995) refer to this phenomenon "when operating in an organizational context that requests appraisal information as a basis for multiple and possibly conflicting uses, the rater may attempt to balance these purposes, frequently resulting in ratings that reflect political judgments rather than judgments about performance. There is also evidence that in some settings the rater will select one purpose or goal and then complete the appraisal with that purpose in mind" (p.108).

For raters in the administrative purpose condition, the addition of an expected meeting with a supervisor (strong accountability) may have balanced the pressure that they were already feeling towards the GTA, leading to more accurate ratings. However, in the developmental purpose condition, the expectation of a meeting with a supervisor (strong accountability) may have led to an unbalanced pressure where raters were primarily concerned with explaining their ratings to a supervisor. It is likely that accountability to a supervisor was much more salient to subjects in the strong accountability condition because the manipulation was much stronger. As suggested by Harris (1994), a strong manipulation of accountability involves the expectation of explaining and justifying ratings as well as being prepared to defend them.

If subjects in the developmental/strong condition were influenced by the expectation of a meeting with a supervisor but they were not aware of the view of the supervisor, a high rate of guessing may result. Further, a high rate of guessing would lead to less accurate ratings. In addition, it would explain why raters in the

developmental/strong accountability context were not overly severe or lenient as a whole but provided less accurate ratings than did subjects in the developmental/weak accountability condition. In support of this theory, ratings in the developmental/no supervisor accountability condition had the highest standard deviations.

Although strong accountability did not appear to lead to greater elevation accuracy for raters in the developmental purpose condition, there is not as great of a need for increased accuracy of ratings in this context. The problem of rating distortion appears to be more of a problem for raters when the ratings are to be used to make administrative decisions.

Next, raters who believed that the purpose of appraisal was to provide feedback for development exhibited greater behavioral accuracy than did raters who believed that the purpose was to make administrative decisions. One explanation for this finding may be the "acceptability heuristic" (Tetlock, 1985). Raters who believed that the ratings would be used to make administrative decisions were probably more likely to endorse false positive responses in favor of the ratee in order to support a positive evaluation of the ratee. Subjects in the developmental purpose condition, however, were probably more likely to evaluate the performance of the GTA without a predetermined bias. This lack of bias involves more cognitive work, so raters in this condition may have paid closer attention to the lecture in order to make their evaluations.

Additionally, results of the current study indicated that increased accountability to a supervisor led to greater behavioral accuracy, indicating that raters who were to be held accountable to a supervisor tended to recall specific behaviors of the GTA more accurately than those who were not to be held accountable to a supervisor. This finding complements the Tetlock (1985) finding that people are "cognitive misers" who will exert the minimum amount of effort necessary to reach a decision. This suggests that the anticipation of a review of their ratings or a meeting with a supervisor may have led raters to exert more effort by paying closer attention to the lecture.

Although Hypotheses 1,2,4 and 5 received support, Hypotheses 3 and 6 did not. No interactions were found in the analyses of this study. Overall, observed ratings in the administrative conditions were fairly consistent with the predictions of the study;

however, ratings in the developmental condition appeared to become less accurate in a strong accountability context.

Finally, one advantage of examining patterns of both rating accuracy and behavioral accuracy in this study is that these patterns can be compared. Similar patterns emerged for accountability and appraisal purpose on rating accuracy (elevation accuracy) and behavioral accuracy. The Pearson correlation for these two variables was $-.692$ ($p < .01$). The correlation is negative because the elevation accuracy measure is scaled so that lower numbers reflect greater accuracy, whereas the behavioral accuracy measure is scaled so that higher numbers reflect greater accuracy. In both cases, increased accountability to a supervisor led to increased accuracy when the purpose was to make administrative decisions. Also, in both cases, weak accountability to a supervisor led to increased accuracy but strong accountability to a supervisor led to decreased accuracy when the purpose was for feedback and development. These findings suggest that strong accountability (a face-to-face meeting) to a supervisor may only be beneficial when the purpose of the appraisal is to make administrative decisions. When ratings are being made for the purpose of feedback and development, it appears that weak accountability (a supervisor review of ratings) leads to the most accurate ratings.

One possible limitation of this study was the conceptualization of "downward" accountability. Specifically, subjects may not have viewed the GTA as a subordinate. For the purpose of this study, it may make more sense to conceptualize "downward" accountability as accountability to a vested source. One explanation for rating inflation in previous research has been that subordinates have vested interests in their performance appraisals (i.e., raises, promotions, wanting to be rated favorably) (Longenecker et al., 1987). Given this finding, the results of the current study still make intuitive sense when downward accountability is conceptualized as accountability to a vested source.

Additionally, when considering the results of this study, it is important to consider that this was a study conducted in a lab setting involving teaching evaluations. The results of this study should not be generalized to an organizational context. The primary goal of this study was to gain insight into the motivational determinants of accuracy in a performance appraisal context.

Future research should also attempt to duplicate these findings in an organizational context. In addition, an examination of the cognitive processes that are involved in different motivational contexts of performance appraisal would provide an opportunity to integrate these two domains of research.

Finally, it would be useful to further examine the relationship between rating accuracy and behavioral accuracy. There have been only two studies on this topic in the context of performance appraisal (Murphy, Garcia, Kerkar, Martin, and Balzer, 1982; Sanchez & De La Torre, 1996). Both studies found a relationship between rating accuracy and behavioral accuracy. However, the usage of a common method to index both rating and behavioral accuracy in the first study may have inflated the intercorrelation, and only small, limited relationships were found in the latter study. Sanchez and De La Torre (1996) found that the relationship between rating and behavioral accuracy was confined to stereotype and differential accuracy, and the effect sizes for these relationships were small. Further research on this topic will help determine the nature of the relationship between these two variables.

References

- Banks, C.G. & Murphy, K.R. (1985). Toward or narrowing the research-practice gap in performance appraisal. Personnel Psychology, 38, 335-345.
- Bernardin, H.J., & Buckley, M.R. (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.
- Bernardin, H.J., & Pence, E.C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.
- Borman, W.C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.
- Cleveland, J.N. & Murphy, K.R. (1992). Analyzing performance appraisal as goal-directed behavior. In G. Ferris & K. Rowland (Eds.), *Research in personnel and human resources management* (Vol. 10, pp. 121-185). Greenwich, CT: JAI Press.
- Cleveland, J.N., Murphy, K.R., & Williams, R.E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. Journal of Applied Psychology, 74, 130-135.
- Corrigan, D.K. (1994). Accountability effects on cognitive complexity and accuracy of performance ratings. Unpublished master's thesis, Virginia Polytechnic Institute and State University, Blacksburg.
- Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". Psychological Bulletin, 90, 218-244.
- Fandt, P.M., & Ferris, G.R. (1990). The management of information and impressions: When employees behave opportunistically. Organizational Behavior and Human Decision Processes, 45, 140-158.
- Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.
- Fisher, B.A. (1974). Transmission of positive and negative feedback to subordinates: A laboratory investigation. Journal of Applied Psychology, 64, 533-540.
- Fiske, S.T. & Taylor, S. (1984). *Social Cognition*. Reading, MA: Addison-Wesley.

- Foti, R.J. & Hauenstein, N.M.A. (1993). Processing demands and the effects of prior impressions on subsequent judgments: Clarifying the assimilation/contrast debate. Organizational Behavior and Human Decision Processes, 56, 167-189.
- Hagafors, R. & Brehmer B. (1983). Does having to justify one's judgments change the nature of the judgment process? Organizational Behavior and Human Performance, 31, 223-232.
- Harris, M.H., & Shaubroeck, J. (1988). A meta-analysis of self-supervisory, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-62.
- Harris, M.M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. Journal of Management, 20, (4), 737-756.
- Harris, M.M., Smith, D.E., Champagne, D. (1995). A field study of performance appraisal purpose: Research-versus administrative based ratings. Personnel Psychology, 48, 151-160.
- Hauenstein, N.M.A. (1992). An information-processing approach to leniency in performance judgments. Journal of Applied Psychology, 77, 485-493.
- Hauenstein, N.M.A., & Alexander, R.A. (1991). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. Organizational Behavior and Human Decision Processes, 50, 300-323.
- Hauenstein, N.M.A., & Foti, R.J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. Personnel Psychology, 42, 359-378.
- Illgen, D.R., Barnes-Farrell, J.L., & McKellin, D.B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? Organizational Behavior and Human Decision Processes, 54, 321-368.
- Jawahar, I.M., & Stone, T.H. (1997). Influence of raters' self-consciousness and appraisal purpose on leniency and accuracy of performance ratings. Psychological Reports, 80, 323-336.
- Kavanagh, M.J., MacKinney, A.C., and Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

- Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. Organizational Behavior and Human Decision Processes, 45, 194-208.
- Landy, F.J., & Farr, J.L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.
- Longenecker, C.O., & Gioia, D.A. (1992). The executive appraisal paradox. Academy of Management Executive, 6 (2), 18-28.
- Longenecker, C.O., Sims, H.P., & Gioia, D.A. (1987). Behind the mask: The politics of employee appraisal. Academy of Management Executive, 1, 183-193.
- Lord, R.G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. Journal of Applied Psychology, 70, 66-71.
- McIntyre, R.M., Smith, D., & Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.
- Mero, N.P., & Motowidlo, S.J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. Journal of Applied Psychology, 80 (4), 517-524.
- Murphy, K.R., Balzer, W.K., Kellam, K.L., & Armstrong, J. (1984). Effect of purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. Journal of Educational Psychology, 76, 45-54.
- Murphy, K.R., Balzer, W.K., Lockhart, M., & Eisenman, E. (1985). Effects of previous performance on evaluations of present performance. Journal of Applied Psychology, 70, 72-84.
- Murphy, K.R. & Cleveland, J.N. (1995). Understanding performance appraisal: Social, organizational, and goal based perspectives. Thousand Oaks, CA: Sage. Chapters 6,7.
- Nathan, B.R., & Lord, R.G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.
- Pearce, J.L., & Porter, L.W. (1986). Employee responses to formal appraisal feedback. Journal of Applied Psychology, 71, 211-218.

- Reilly, C.E., & Balzer, W.K. (1988). Effect of purpose on observation and evaluation of teaching performance. Unpublished manuscript, Bowling Green State University.
- Salvemini, N.J., Reilly, R.R., & Smither, J.W. (1993). The influence of rater motivation on assimilation effects and accuracy in performance ratings. Organizational Behavior and Human Decision Processes, 55, 41-60.
- Sanchez, J.I. & De La Torre, P. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. Journal of Applied Psychology, 81, 3-10.
- Sharon, A. & Bartlett, C. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. Personnel Psychology, 22, 252-263.
- Simonson, I., & Nye, P. (1992). The effect of accountability on susceptibility to decision errors. Organizational Behaviors and Human Decision Processes, 51, 416-446.
- Stamoulis, D. T. (1993). Making raters more accountable for their performance ratings: Effects of expecting a supervisory review of ratings. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg.
- Stamoulis, D.T., & Hauenstein, N.M.A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. Journal of Applied Psychology, 78, 42-54.
- Sulsky, L.M. & Balzer, W.K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 497-506.
- Sulsky, L.M. & Day, D.V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. Journal of Applied Psychology, 77(4), 501-510.
- Tetlock, P.E. (1983a). Accountability and complexity of thought. Journal of Personality and Social Psychology, 45, 74-93.
- Tetlock, P.E. (1983b). Accountability and the perseverance of first impressions. Social Psychology Quarterly, 46, 285-292.

- Tetlock, P.E. (1985). Accountability: The neglected social context of judgment and choice. In B. Staw & L. Cummings (Eds.), Research in Organizational Behavior. Greenwich, CT: JAI Press.
- Tetlock, P.E., Stitka, L. & Boettger, R. (1989). Social and cognitive strategies for coping with accountability: Conformity, complexity, and bolstering. Journal of Personality and Social Psychology, 57, 632-640.
- Zedeck, S., & Cascio, W.F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67, 752-758.

Table 1
Means and Standard Deviations for Rating Form Target Scores

Dimension	M	SD
Knowledge	5.00	1.28
Delivery	3.42	1.16
Relevance	4.17	1.27
Organization	5.25	1.06

Note: N= 12

Table 2

Means and Standard Deviations from Pilot Study Questionnaire

Supervisor	Purpose	Q # 1		Q # 2		Q # 3		Q # 4	
		M	SD	M	SD	M	SD	M	SD
Accountability	None								
	Adm	6.20	1.62	5.70	1.42	2.70	.95	5.90	.99
	Dev	5.90	1.37	5.40	1.43	3.30	1.49	5.30	1.57
	Total	6.05	1.47	5.55	1.39	3.00	1.26	5.60	1.31
Weak	Adm	6.10	1.91	6.10	1.37	4.10	1.66	5.70	1.34
	Dev	5.60	1.78	5.10	1.66	3.30	1.57	5.70	1.25
	Total	5.85	1.81	5.60	1.57	3.70	1.63	5.70	1.26
Strong	Adm	6.30	1.16	6.20	.79	5.70	1.16	6.00	1.25
	Dev	6.30	1.06	5.80	1.03	5.20	1.14	5.10	1.29
	Total	6.30	1.08	6.00	.92	5.45	1.15	5.55	1.32
Total	Adm	6.20	1.54	6.00	1.20	4.17	1.76	5.87	1.17
	Dev	5.93	1.41	5.43	1.38	3.93	1.64	5.37	1.35
	Total	6.07	1.47	5.72	1.32	4.05	1.69	5.62	1.28

Table 3

ANOVA Table for Pilot Study Questionnaire

Source	Sum of Squares	df	Mean Square	F-Ratio
Q#1				
Accountability	2.033	2	1.017	.443
Purpose	1.067	1	1.067	.465
Account*Purpose	.633	2	.317	.138
Residual	124.000	54	2.296	
Q#2				
Accountability	2.433	2	1.217	.703
Purpose	4.817	1	4.817	2.782
Account*Purpose	1.433	2	.717	.414
Residual	93.500	54	1.731	
Q#3				
Accountability	63.700	2	31.850	17.390**
Purpose	.817	1	.817	.446
Account*Purpose	5.433	2	2.717	1.483
Residual	98.900	54	1.831	
Q#4				
Accountability	.233	2	.117	.070
Purpose	3.750	1	3.750	2.248
Account*Purpose	2.100	2	1.050	.629
Residual	90.100		1.669	

Note: ** $p < .001$

Table 4
Means and Standard Deviations for Accuracy Measures

Supervisor Accountability	Purpose	Elevation		Dimensional		Leniency		Behavioral	
		M	SD	M	SD	M	SD	M	SD
None	Adm	1.2085	.6689	1.3097	.6211	1.1150	.8196	.5788	.1741
	Dev	1.0709	.9192	1.2913	.7178	-.2475	1.4102	.5970	.1992
	Total	1.1397	.7966	1.3005	.6626	.4338	1.3312	.5879	.1849
Weak	Adm	1.0251	.6632	.9632	.4444	.6900	1.0208	.6094	.2047
	Dev	.5125	.3938	1.2412	.8216	-.1225	.6452	.7656	.1824
	Total	.7688	.5977	1.1022	.6670	.2838	.9380	.6875	.2070
Strong	Adm	.7460	.5994	1.3237	.7045	.2900	.9248	.6686	.2569
	Dev	.8083	.7538	1.2107	.3856	-.3975	1.0446	.7313	.2147
	Total	.7771	.6729	1.2672	.5635	-.0054	1.0342	.6999	.2358
Total	Adm	.9932	.6620	1.1989	.6131	.6983	.9709	.6189	.2143
	Dev	.7972	.7470	1.2477	.6575	-.2558	1.0671	.6980	.2091
	Total	.8952	.7096	1.2233	.6335	.2213	1.1232	.6584	.2145

Table 5

ANOVA Table for Accuracy Measures

Source	Sum of Squares	df	Mean Square	F-Ratio
Elevation				
Accountability	3.588	2	1.794	3.824*
Purpose	1.152	1	1.152	2.456
Account*Purpose	1.704	2	.852	1.816
Residual	53.482	114	.469	
Dimensional				
Accountability	.902	2	.451	1.119
Purpose	.072	1	.007	.178
Account*Purpose	.833	2	.416	1.033
Residual	45.953	114	.403	
Leniency				
Accountability	4.988	2	2.494	2.467
Purpose	27.313	1	27.313	27.019**
Account*Purpose	2.579	2	1.290	1.276
Residual	115.241	114	1.011	
Behavioral				
Accountability	.302	2	.151	3.520*
Purpose	.187	1	.187	4.372*
Account*Purpose	.099	2	.050	1.158
Residual	4.887	114	.043	

Note: ** $p < .001$ Note: * $p < .05$

Appendix A

Please respond to the following questions by circling the number that most closely reflects your opinion.

1. To what extent did you feel that the feedback meetings would take place?

1 2 3 4 5 6 7
To no extent To some extent To a great extent

2. To what extent did you feel accountable to the GTA?

1 2 3 4 5 6 7
To no extent To some extent To a great extent

3. To what extent did you feel accountable to a faculty representative for the Department of Psychology?

1 2 3 4 5 6 7
To no extent To some extent To a great extent

4. To what extent did you consider the purpose of these evaluations?

1 2 3 4 5 6 7
To no extent To some extent To a great extent

5. Did you feel that the pressures created by the accountability you experienced and the purpose of the evaluation were equal? Please explain your answer.

Appendix B

Behavioral Incidents of the Videotaped Lecture

Dimension 1: Organization

1. Lecturer ties in the present day's lecture with the previous day's lecture.
2. Lecturer presents to the class a daily outline of the topics to be covered in the lecture.
3. Lecturer concludes by summarizing the day's and introducing the next day's topic.
4. Lecturer discusses each topic in the same order as was presented in the lecture outline.

Dimension 2: Depth of Knowledge

5. Lecturer is familiar with research critiquing Law of Demand.
6. Lecturer presents multiple citations concerning the effect of price on purchasing behavior.
7. Lecturer presents results of personal research.
8. Lecturer states that he possesses in depth knowledge of the Theory of Absolute Price Thresholds because of his future intentions of research in the area.

Dimension 3: Relevance

9. Lecturer explains the law of demand in an environment familiar to his audience (a supermarket).
10. Lecturer explains how price affects product choice using products familiar to his audience (household goods and products).
11. Lecturer explains the difference between lay person and expert shopper by using an example which confuses rather than elucidates his point (choosing a book by different authors based on the physical construction of the books).
12. Lecturer discusses upper and lower price thresholds using prices of a product meaningless to his audience (the price of a diamond ring in old French francs).

Dimension 4: Delivery

13. Initially, lecturer speaks from lectern or at the blackboard; refrains from pacing.
14. Lecturer writes legibly when using the blackboard to report the results of his single cue study.
15. Midway through the lecture, the lecturer begins pacing.
 16. Lecturer does not label or explain this graph of the Demand Curve.

Appendix C
Teaching Evaluation Form

Below you will find a list of dimensions that you will use to evaluate the GTA. Read the definition of the dimension carefully to be sure you understand exactly what you are evaluating, then rate the GTA on each dimension by circling the number that most closely reflects your evaluation.

Depth of Knowledge: The instructor's mastery of the subject matter; this includes how well he or she knows the literature and the research being discussed.

1	2	3	4	5	6	7
Low			Medium		High	

Delivery: The instructor's manner of speaking and the extent to which he or she uses the board to clarify and emphasize important points of the lecture.

1	2	3	4	5	6	7
Poor		Average			Excellent	

Relevance: The instructor's choice of examples used in conveying information; the extent to which examples are important and meaningful to the audience.

1	2	3	4	5	6	7
Poor		Average			Excellent	

Organization: The instructor's arrangement of the lecture; the extent to which the instructor leads the class through a logical and orderly sequence of the material.

1	2	3	4	5	6	7
Poor		Average			Excellent	

Appendix D

Behavioral Recognition Measure

The following is a list of behaviors that the lecturer in the videotape may or may not have performed. For each item, circle Y for “yes” if you remember the lecturer performing that behavior and N for “no” if you do not remember that behavior being performed.

- Y N 1. Instructor ended by summarizing current lecture and introducing the next topic.
- Y N 2. Instructor was very familiar with research on how brand names affect purchasing decisions.
- Y N 3. Instructor explained the Law of Demand in a familiar situation (i.e. buying meat at the supermarket).
- Y N 4. Instructor used overheads.
- Y N 5. Instructor drew diagrams on the blackboard before the lecture began.
- Y N 6. Instructor knew the author of an article critical of the Law of Demand.
- Y N 7. Instructor used a product popular with students (beer) to explain how price affects perceptions of quality.
- Y N 8. Instructor wrote legibly on the board.
- Y N 9. Instructor tied the current lecture into the previous lecture.
- Y N 10. Instructor clarified a confusing part in the textbook on supply and demand during his lecture.
- Y N 11. Instructor used an example of choosing a book by different authors based on the physical construction of the book.
- Y N 12. Instructor spoke too softly.

- Y N 13. Instructor brought chalk with him to the T.V. studio in case none was available.
- Y N 14. Instructor discussed his research in detail to illustrate a point in the lecture.
- Y N 15. Instructor used an example involving the purchase of drugs.
- Y N 16. The instructor paced during the lecture.
- Y N 17. Instructor put an outline of the day's lecture on the board.
- Y N 18. Instructor presented multiple examples of research studies illustrating the Theory of Absolute Price Threshold.
- Y N 19. Instructor illustrated how price affects perceptions of quality using various products (e.g. motor oil, shaving cream, razor blades) as examples.
- Y N 20. Instructor had handouts available for certain lecture topics.
- Y N 21. Instructor had handouts available outlining the material covered in lecture.
- Y N 22. Instructor mentioned many different articles concerning the effect of behavior.
- Y N 23. To show the effects of brand names on perception of quality, the instructor used designer jeans as an example.
- Y N 24. Instructor used chalkboard appropriately.
- Y N 25. Instructor discussed each topic in the same order as was presented in the lecture outline.
- Y N 26. Instructor was involved in a research project with an important figure in the area of consumer psychology.
- Y N 27. Instructor used an example of how many francs people in France were willing to spend on a diamond.
- Y N 28. Instructor had a distracting habit of removing his glasses and pinching the bridge of his nose.
- Y N 29. Prior to the lecture, instructor had set-up all necessary audio-

visual equipment for presenting the lecture.

Y N 30. Instructor was familiar with a body of research because of his intention to do future research in the area.

Y N 31. Instructor used unrealistic price levels when presenting examples of consumer purchasing decisions (i.e. buying a stereo for \$50.00).

Y N 32. Instructor did not label or explain his graph of the demand curve.

- Correct responses are underlined

Rachel L. Fredholm

EDUCATION:

M.S. Virginia Polytechnic Institute and State University
Industrial & Organizational Psychology

B.A. University of Texas at Austin
Psychology

RESEARCH AND WORK EXPERIENCE:

August 1996- Graduate Advisor & Research Assistant, Virginia Polytechnic
August 1998 Institute and State University, Blacksburg, VA

- Advised undergraduates about classes and job opportunities
- Attended regular meetings regarding university policies
- Developed surveys to assess effectiveness of advising center
- Analyzed survey results and prepared for presentation

February 1997- Assessor, Tennessee Valley Authority, Knoxville, Tennessee
July 1997

- Attended 40 hour assessor training program
- evaluated assessment center participants in multiple assessment center projects
- wrote feedback reports for presentation to participants
- participated in consensus meetings to determine overall ratings and rank order of job candidates

July 1997 Role-Player, Tennessee Valley Authority, Knoxville, Tennessee

- Assumed role of Regional Manager for role-play exercise in assessment center

Summer 1997 Instructor, Advanced Social Psychology Lab, Virginia Polytechnic
Institute and State University, Blacksburg, VA

- Developed lecture notes for senior-level lab
- Administered lab

September 1995-
August 1996 Research Assistant, Department of Psychology, Virginia Polytechnic
Institute and State University, Blacksburg, VA

- Organized database for large-scale leadership study
- Assisted with administration of study

August 1995-
May 1996 Graduate Teaching Assistant, Introductory Psychology, Virginia Polytechnic
Institute and State University, Blacksburg, VA

- Taught two lab classes
- Attended weekly planning sessions
- Proctored exams

December 1994-
August 1995 Human Resource Administrator, Foley's Department Store,
Highland Mall, Austin, TX

- Interviewed and hired all personnel (including executives)
- Conducted orientation training for new hires
- Input all employee information (from hire to termination)
- Assisted employees with any work-related problems
- Communicated urgent situations to corporate HR management
- Conducted performance evaluations
- Administered benefits to all employees
- Handled worker's compensation cases
- Terminated employees

July 1994-
December 1994 Human Resources Administrative Assistant, Barton
Creek Conference Resort and Country Clubs, Austin, TX

- Reviewed applications and resumes of prospective employees and selected qualified applicants to meet with the area managers

- Worked with employees and their managers on issues such as compensation, status, and vacation/sick leave
- Maintained and updated employee files
- Input new employee information into HR program, ABRA 2000
- Verified past and present employment

PUBLICATIONS:

Facteau, J.D., Facteau, C.L., McGonigle, T.P., and Fredholm, R.L. (1997). Characteristics of Ratings and Managers' Reactions to Multisource Performance Appraisal Feedback. Manuscript submitted for publication.

PRESENTATIONS:

Facteau, J.D., Facteau, C.L., McGonigle, T.P., and Fredholm, R.L. (April 1997). Characteristics of feedback and managers' reactions to multisource performance appraisal systems. Poster presented at the twelfth annual convention of the Society for Industrial and Organizational Psychology, St. Louis, MO

Facteau, J.D., Fredholm, R.L., Keller, K.D., McGonigle, T.P., & LeBreton, D.L. (April 1998). A further validation of the construct of goal orientation. Poster presented at the thirteenth annual convention of the Society for Industrial and Organizational Psychology, Dallas, TX.