EFFECTS OF PURPOSE OF APPRAISAL ON LENIENCY ERRORS: AN EXPLORATION OF SELF-EFFICACY AS A MEDIATING VARIABLE

By
Adam N. Prowker

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

Masters of Science
in
Psychology

1999

Neil M.A. Hauenstein, Chairperson
John J. Donovan
Roseanne J. Foti

12 May 1999
Blacksburg, Virginia

Keywords: Purpose of Appraisal, Leniency, Self-Efficacy, Performance Appraisal

EFFECTS OF PURPOSE OF APPRAISAL ON LENIENCY ERRORS: AN EXPLORATION OF SELF-EFFICACY AS A MEDIATING VARIABLE

Adam N. Prowker

(ABSTRACT)

The purpose of this study was to explore the causal relationship between purpose of appraisal and leniency in performance ratings. A model in which rater self-efficacy mediated the relationship between appraisal purpose and leniency was tested. In addition, behavioral recognition accuracy was hypothesized to affect rater leniency. In a laboratory setting, 109 undergraduate raters judged the videotaped performance of a graduate teaching assistant. Results of the study showed that (a) leniency was positively related to task-specific self-efficacy, and (b) behavioral recognition accuracy was positively related to general rating self-efficacy. A purpose of appraisal effect was not observed and the proposed mediational model was not supported. These results were discussed in relation to rater affect, accountability in performance ratings, and information processing.

CONTENTS

LIST OF TABLES

Effects of purpose of appraisal on leniency errors:
An exploration of self-efficacy as a mediating variable

Introduction

The errors associated with performance evaluations are a major concern in both academic and industrial organizations "because the actions based on the information from the rating process have critical implications for both the individual and the organization" (Lee, 1985, p. 322). Information provided by performance appraisals may be applied to many managerial functions such as: personnel training, wage and salary administration, placement, promotions, and personnel research (Murphy & Cleveland, 1995). Performance appraisal information also has uses associated with federal law on fair employment practices. The need for accurate performance measures is essential to the successful management of human labor (Bretz, Milkovich, & Read, 1990).

Despite the importance of performance measures, their validity is compromised by inaccuracies such as leniency. Research on performance appraisal has indicated that the purpose of rating has an effect on the severity of leniency errors (e.g., Taylor & Wherry, 1951). However, research on purpose of appraisal has not consistently yielded leniency effects (e.g., Murphy, Balzer, Kellam, & Armstrong, 1984), suggesting that perhaps the relationship between purpose of appraisal and leniency error is only part of a more complex model. Research has neglected to explore the possibility that the relationship between purpose of appraisal and rating accuracy may be mediated (Baron & Kenny, 1986) by a third variable.

Landy & Farr (1980) suggested that the purpose of rating is one of a number of situational or contextual variables that affects the evaluation of performance. Their review of performance rating research yielded inconsistent results surrounding the connection between demographic variables (i.e., rater sex) and rating effectiveness. Landy & Farr concluded that cognitive process research (i.e., cognitive complexity) held the most promise for understanding the performance appraisal process. However, 20 years of cognitive research has not shed much light on performance appraisal.

Cognitive processes in performance appraisal have been explored concerning rating format (e.g., Murphy & Constans, 1987) and rater training (e.g., Feldman, 1986). Results of these research foci have suggested that it may be possible to improve performance rating accuracy by tailoring rating formats and rater training to the individual cognitive styles of the raters (Murphy & Cleveland, 1995). More recently, cognitive variables such as rater intelligence, rater implicit theories (Hauenstein & Alexander, 1991), and rater information processing (Hauenstein, 1992) have demonstrated effects on rating accuracy. Although cognitive research has proved to be informative, research in this area has not been exhaustive. "The concepts that drive much of the current cognitive research…. are useful but they do not exhaust the possibilities" (Murphy & Cleveland, 1995, p. 209). Further research exploring individual differences in cognitive processing is needed to better understand the performance process.

The present study explored the relationship between rater confidence (operationalized as rater self-efficacy) and leniency in performance ratings. A model in which rater self-efficacy mediated the relationship between purpose of appraisal and leniency was tested. This model was based on the assumption that the potential negative consequences for ratees associated with low ratings are perceived by the rater as being

greater when the ratings are going to be used for administrative decisions than when the ratings are going to be used for developmental decisions. These perceptions of negative consequences were predicted to affect rater confidence and subsequently serve to augment the conflict-avoidance strategies that raters engage in.

Leniency

Saal & Landy (1977) defined leniency as a rater's tendency to assign higher ratings to an individual's performance than the actual performance demands. Similarly, Saal, Downey, & Lahey (1980) defined leniency as a shift in mean ratings away from a scale midpoint. Consistent among all definitions is that leniency is an inaccurate inflation of an individual's performance ratings.

The phenomenon of leniency has been examined from several different perspectives and its explanations have included motivational (e.g., Murphy & Cleveland, 1995), political (e.g., Longenecker, Sims, & Gioia, 1987), and cognitive (e.g., Hauenstein, 1992) processes. Murphy & Cleveland (1995) presented a leniency model that addressed the motivation to provide accurate versus distorted ratings. They believed that people are generally capable of accurately evaluating performance and those inaccuracies in evaluations are "more likely to be a result of the rater's willingness to provide accurate ratings than of his or her capacity to rate accurately" (p. 242).

Murphy & Cleveland's (1995) model of leniency considers both the rewards and the negative consequences of giving accurate ratings from an expectancy perspective. Organizations stress the importance of quality performance evaluations; however, there are generally no rewards for doing so. In addition, organizations typically do not monitor the quality of ratings and subsequently do not exercise sanctions for inaccurate ratings. The negative consequences for giving (low) accurate ratings are immediate and tangible (i.e., loss of rewards, negative interpersonal relationships). High ratings ensure ratees organizational rewards such as promotions and salary increases. Research suggests that ratings used to determine the allocation of rewards (i.e., administrative decisions) will be more biased by leniency than those used for other decisions (this literature will be discussed in a future section). Murphy & Cleveland (1995) suggest that raters are motivated to inflate ratings in order to (a) avoid negative interpersonal relationships with subordinates, (b) make themselves as managers look better, and (c) to conform to organizational norms of high ratings. Given that rewards for accurate ratings are scarce and consequences are numerous coupled with a low probability for receiving rewards and a high probability of experiencing consequences, raters are motivated not to accurately rate performance, but to distort ratings upward in order to avoid the negative consequences associated with low ratings.

Longenecker et al. (1987) examined performance appraisal inaccuracy from a political perspective. They defined political as "deliberate attempts by individuals to enhance or protect their self-interests when conflicting courses of actions are possible" (p. 184). Longenecker et al. (1987) believed that manipulations of appraisal information was systematic and intentional, and that executives felt it was within their discretion to distort ratings in order to more effectively manage their subordinates. Executives realized that they interact with their subordinates on a day-to-day basis and the permanence of written evaluation information. Consequently, deliberate inflation/deflation of ratings is a smart management strategy rather than bias or error. Reasons for inflating the appraisal include: (a) to maintain or increase performance, (b) maximize merit increases that a

subordinate would be eligible to receive, (c) avoid confrontation with subordinates, and (d) to promote poor performing subordinates "up and out" of the executive's department.

Although research seems to make the case that leniency in performance ratings is due primarily to motivated distortion, Hauenstein (1992) has offered evidence that leniency may be in part due to cognitive distortions associated with information-processing strategies. This approach to understanding leniency suggests that distortions in performance ratings are not necessarily an after-the-fact phenomenon but rather a function of information processing at the time of performance observation. Raters who give high ratings seem to pay more attention to, and are more accurate in their recognition of, positive behavioral instances. In addition, purpose of appraisal information may also serve to concentrate the rater's attention to either positive or negative behavioral incidents and convey a priori expectations of performance ratings.

Neither the motivational nor the cognitive models are being endorsed as the absolute model of leniency. Instead, it is important to realize that leniency is a function of both motivational and cognitive processes operating independently and dependently of each other (Hauenstein, 1992). Although Murphy & Cleveland (1995) argued that motivated distortion is the primary explanation for leniency in performance ratings, they acknowledged that motivation is not the sole cause of leniency and that cognitive processes may exercise some effect on rating accuracy. It is reasonable to assume that motivational and cognitive processes mutually influence each other and "that the contextual factors that motivate raters to render lenient ratings can also bias the cognitive processes involved when raters form impressions of ratees" (Hauenstein, 1992, p. 486).

Purpose of Rating

Taylor & Wherry (1951) concluded that there is "an important and highly significant difference between the mean ratings collected for experimental purposes and that of those which have administrative implications" (p. 42). The results of their study proposed that raters bias their ratings more or less for some purposes versus others. In their review of several performance appraisal models (i.e., DeCotiis & Petit, 1978; Landy & Farr, 1980), Murphy & Cleveland (1995) concluded that the intended use of performance appraisal information is thought to have a significant impact on the properties of ratings (e.g., leniency, accuracy). It is hypothesized (Jawahar & Williams, 1997; Murphy & Cleveland, 1995; 1991) that raters may intentionally bias ratings when the results of their evaluations have administrative implications to a) avoid possible negative consequences associated with accurate ratings that may be less then favorable (e.g., no pay increase, refused promotion), b) to obtain positive consequences (e.g., pay increases, employee commitment), or c) to motivate poor performers.

Research studying the effects of purpose of rating on rater evaluations has focused on three main purposes: administrative (i.e., promotion, discipline, salary increases and selection), feedback (i.e., training and employee development), and research (i.e., scale, selection and predictor validation). This literature suggests that raters are more lenient when they believe that their evaluations will be used for administrative decisions than when they believe that their evaluations will be used for feedback or research purposes (Aleamoni & Hexner, 1980; Driscoll & Goodwin, 1979; Farh & Werbel, 1986; Harris, Smith, & Champagne, 1995; Heron, 1956; Taylor & Wherry, 1951; Waldman & Thorton, 1988; Williams, DeNisi, Blencoe, & Cafferty, 1985; Zedeck & Casio, 1982). Other studies have not found a difference in leniency errors as a function of purpose of rating

(Berkshire & Highland, 1953; Bernardin, 1978; Bernardin & Cooke, 1992; Centra, 1976; Hollander, 1965; McIntyre, Smith, & Hassett, 1984; Meier & Feldhusen, 1979; Murphy et al., 1984; Sharon, 1970; Sharon & Barlett, 1969).

Murphy & Cleveland (1995) noted that greater leniency for administrative decisions was found for field studies (e.g., Taylor & Wherry, 1951; Heron, 1956) than for laboratory studies (e.g., McIntyre et al. 1984; Murphy et al., 1984). They hypothesized that the inconsistencies in the literature may be an artifact of the different methods used. This finding is not intended to discount the possibility of conducting laboratory research to test for purpose of appraisal effects on leniency, but rather to highlight a difference in effect size based on the population used. Farh & Werbel (1986), Williams, DeNisi, Blencoe, & Cafferty (1985), Aleamoni & Hexner (1980), and Driscoll & Goodwin (1979) conducted laboratory research and found that ratings obtained for administrative purposes were more lenient than those obtained for developmental/research purposes.

Jawahar and Williams (1997) conducted a meta-analysis on performance appraisal purpose of effect. The results indicated that ratings for administrative decisions were nearly one-third of a standard deviation higher than ratings for research or developmental decisions. These results support Taylor and Wherry's (1951) performance appraisal model, which predicts that purpose of appraisal influences rating accuracy. Results of the meta-analysis also found that purpose of appraisal effects were higher for (a) field studies vs. laboratory research, (b) ratings done by organizational raters versus student raters, (c) ratings done on live observation of ratees versus paper people, (d) downward appraisal versus upward appraisal, and (e) when graphic rating scales were used rather than forced choice scales or behaviorally anchored ratings scales (Jawahar & Williams, 1997). The existence of purpose of appraisal moderators supports Murphy & Cleveland's (1995) hypothesis concerning methodological artifacts.

Criticisms of Field Research

One major criticism of the assumptions that characterize the typical measurement of leniency in field research is that "the true distribution of performance is always unknown" (Murphy & Cleveland, 1995, p. 276). Leniency error cannot be measured using the definition that leniency is a shift in mean ratings away from a scale midpoint, because organizational practices operate to ensure that performance distributions are not normal (Murphy and Cleveland, 1995). Organizations engage in activities ranging from selection to training to produce negatively skewed performance distributions (Saal et al., 1980).

Known "true scores" of performance in laboratory research allow for more accurate measures of rater leniency to be obtained. This increase in measurement precision that is typically associated with laboratory research has yielded results that are inconsistent with field research. Driscoll & Goodwin (1979) observed that student ratings of instructors were more lenient when the perceived purpose of rating was for administrative decisions versus feedback and research. Similar results were obtained by Dobbins, Cardy, & Truxillo (1988). However, they observed differences in leniency errors to be more a function of gender stereotypes than the purpose of rating. Conversely, research by Murphy et al. (1984) and Meier & Feldhusen (1979) found no effect of purpose of rating on leniency errors.

Rater Characteristics

Performance ratings are done by a rater and each rater brings a unique set of personal characteristics to the rating procedure. Rater characteristics vary in their relevance to the rating process, however, each characteristic represents a potential for variance due to personal bias (Landy & Farr, 1983). Research has explored a myriad of potential factors that may have a direct and/or indirect effect on the accuracy of performance ratings. One of the most popular demographic variable of interest in rating accuracy research has been rater gender. A review of the research on the effects of rater sex on rating accuracy has concluded that there have been no consistent effects (Landy & Farr, 1980). Results from instructional settings (e.g., Elmore & LaPointe, 1974), simulated work settings (e.g., Dipboye, Arvey, & Terpstra, 1977), laboratory/research settings (e.g., Jacobson & Effertz, 1974; Mischel, 1974), and "real" work settings (e.g., Gupta, Beehr, & Jenkins, 1980) have not found the gender of the rater to have a consistent effect on rating accuracy.

Research has examined the effects of rater race on the accuracy of ratings. Kraiger & Ford (1985), Hamner, Kim, Baird, & Bigoness (1974), and Crooks (1972) found that supervisory raters tended to give higher ratings to same-race subordinates than to different-race subordinates. Hamner et al. (1974) found that this observed effect only accounted for 2% of the variance in ratings. The race effect on ratings was found to be stronger for black raters (e.g., Schmitt & Lappin, 1980; Wendelken & Inn, 1981). Schmitt & Lappin (1980) also noted that both white and black raters gave more variable ratings and were more confident in the accuracy of the ratings given to ratees of their own social group. To the contrary, some researchers have not found a race-of-rater effect. Schmidt & Johnson (1973), while exploring peer ratings in an industrial setting did not find a race of rater effect on rating accuracy. However, they noted that the subject pool used in their study was potentially biased (i.e., subjects had been exposed to human relations training) and perhaps accounted for the observed results.

The age of the rater has generally been observed as a variable with no or very little effect on rating accuracy. Mandell (1956) found that there was an effect for rater age where younger supervisors were less lenient in subordinate ratings than older supervisors. However, other research on rater age has not revealed an age-of-rater effect (e.g., Schwab & Heneman, 1979; Klores, 1966). In a field study on rater-age effects, Cleveland & Landy (1981) did not find a significant main effect for rater age and performance ratings. They did find that of the six performance dimensions they explored, there was a significant age-of-rater effect for the interpersonal skills dimension.

Age of the rater is one demographic characteristic that can potentially exist as a proxy variable for a more job-related variable such as job experience. Older employees tend to be more experienced in the job due to their length of tenure with the organization and within the job category. Studies have found that with increased experience comes increased leniency in performance ratings (e.g., Cascio and Valenzi, 1977; Mandell, 1956). Mandell (1956) found that raters with more than 4 years of supervisory experience were generally more lenient than less experienced raters. Cascio and Valenzi (1977) found a significantly positive job-experience effect, however, they noted that the effect was weak and resulted in only a small percent of the variance in performance ratings. Klores (1966) did not find an effect for rater job experience.

In looking beyond demographic variables, traditional predictors of leniency include leadership style and personality (Villanova, Bernardin, Dahmus, & Sims, 1993). Research on leadership style and rating leniency has focused on initiating structure, but has yielded inconsistent results. Initiating structure is "the degree to which a leader defines and structures his or her own role and the roles of subordinates toward attainment of the group's formal goals" (Yukl, 1998, p. 47). Klores (1966) found that raters who measured high on consideration were more lenient in their ratings than those who measured high on initiating structure. To the contrary, Drory & Ben-Porat (1980) found that initiating structure and leniency where positively related. Finally, Villanova et al. (1993) found that initiating structure did not have a significant effect on student ratings of academic performance.

Wexley & Youtz (1985) investigated the relationship between personality variables and rating errors. Their results indicated that of the personality variables measured, only variability and positive human nature had a significant relationship with rating leniency. Variability (the extent to which people are seen as differing from one another in basic nature as well as being able to change over time) was found to be negatively related to leniency while positive human nature was positively related to leniency.

One cognitive variable believed to be the key to improving the performance appraisal process is cognitive complexity (Murphy & Cleveland, 1995). Cognitive complexity, which Schneier (1977) defined as "the degree to which a person possesses the ability to perceive behavior in a multidimensional manner" (p. 541), has been considered by some researchers as an important variable in the prediction of performance appraisal effectiveness (e.g., Landy and Farr, 1980). This assumption is based on the findings of Schneier (1977) where cognitive complexity was found to have a positive relationship with rating accuracy. The results of Schneier's (1977) research suggested that performance evaluation would be most effective when the cognitive complexity of the rater matched the complexity of the rating format. Studies have attempted to replicate the results of Schneier (1977), however, these attempts proved to be ineffectual (e.g., Bernardin & Orban, 1990; Wexley & Youtz, 1985; Bernardin, Cardy, & Caryle, 1982; Lahey & Saal, 1981). The lack of follow up support to Schneier's original study have lead to questions concerning not only the construct validity of the measures used to assess cognitive complexity (e.g., Murphy & Cleveland, 1995), but also the validity of cognitive compatibility theory itself (e.g., Bernardin et al., 1980).

Cognitive ability or intelligence has been shown a valid predictor of job performance in a wide range of jobs (Hunter & Hunter, 1984). Smither & Reilly (1987) explored the relationship between rater intelligence and rating accuracy. They predicted that a positive relationship would exist between the two variables. However, the results of their study revealed a curvilinear relationship between rater intelligence and rating accuracy where raters of moderate intelligence were the most accurate. Hauenstein & Alexander (1991) found a similar nonlinear, positive relationship between intelligence and accuracy. However, Hauenstein & Alexander (1991) observed that rater type (normative vs. idiosyncratic) moderated the relationship between intelligence (operationalized as verbal reasoning) and rating accuracy.

There are few general conclusions that can be made from the research surrounding rater characteristics and leniency (Landy & Farr, 1980). It appears that rater

sex and rater age do not generally effect ratings. Raters typically give higher ratings to ratees of the same race, although the effect sizes are not very convincing of the overall strength of this relationship. While leadership style and cognitive complexity have yielded inconsistent results, this research has provided evidence as to the importance of considering cognitive variables. Rater intelligence has been found to have a curvilinear relationship to rating accuracy with raters of moderate intelligence rendering the most accurate ratings.

The results of research surrounding rater experience are potentially perplexing. Hauenstein (1992) believed that one concern with using student raters was that their lack of familiarity with the performance dimensions and organizational expectations would result in less accurate ratings. It could be assumed that by following this logic, increased experience leads to increased familiarity with organizational performance dimensions and expectations, and subsequent increases in ratings accuracy. However, research has demonstrated a positive experience-leniency relationship (e.g., Mandell, 1956). Perhaps the variable of interest is not experience but rater confidence.

Mandell (1956) found that raters with high self-confidence gave more accurate ratings than raters with low self-confidence. Hauenstein (1998) and Neck, Stewart, & Manz (1995) have discussed the notion of rating-confidence as a potential variable that can effect the accuracy of ratings. However, there has not been research that has looked at a person's confidence (self-efficacy) in conducting performance evaluations and rating accuracy.

Self-Efficacy

Bandura (1986) defined self-efficacy as a person's judgement of his/her capabilities to organize and execute courses of action required to attain designated types of performance. Self-efficacy is not concerned with the actual level of one's abilities but rather one's subjective perceptions of what he or she can do with whatever abilities one possess. The concept of self-efficacy simply refers to one's belief in one's capability to perform a specific task (Gist, 1987). Self-efficacy arises from a gradual acquisition of cognitive, social, linguistic, and physical skills through experience (Bandura, 1982).

Bandura (1977) specified three dimensions of self-efficacy: magnitude, strength, and generality. Magnitude refers to the level of task difficulty at which a person believes he or she can perform. Strength indicates the level of confidence one has in performing at the level that he or she endorsed. Finally, generality refers to the global notion of efficacious perceptions extending to other situations. The most informative analysis of self-efficacy requires the assessment of all three dimensions (Bandura, 1986).

Self-Efficacy and Performance

Bandura (1977) argued that the type of behavior and the amount of risk a person was willing to undertake was a result of his/her perceptions of his/her own achievement and effectiveness. "Self-efficacy [is] the belief in one's ability to perform a task or more specifically to execute a specific behavior successfully" (Woodruff & Cashman, 1993, p. 423). If better performers have higher perceptions of personal ability, then it follows that self-efficacy potentially might explain why greater leniency errors occur when evaluations are used for administrative decisions. More specifically, if a rater is not confident that his or her ratings will be accurate, one approach suggests that giving high ratings may be better than giving low ratings.

Brill & Prowker (1996) observed that negative and positive judgements correlated with supervisor rating confidence in opposite directions (i.e., supervisor confidence was negatively correlated with negative judgements and positively correlated with positive judgements). A significant positive relationship was also observed between confidence and ratings. If confidence in ratings is an indication of the strength of one's performance rating efficacy, then the opposite directionality of negative and positive judgements may imply reluctance or uncertainty on the part of supervisors when dealing with poor performers, and vice versa with good performers. This process, when combined with the implications of decisions made for administrative vs. developmental decisions, may stem from a basic conflict-avoidance motivation strategy on the part of the rater who may have limited confidence in his or her ability to accurately appraise performance. Leniency may be negatively correlated to the rater's task-specific self-efficacy.

Task specific self-efficacy refers to an individual's perceptions of his/her ability to successfully perform the behaviors required to produce the desired outcome in a given situation under certain circumstances (Bandura, 1982). Self-efficacy, when tailored to the specific tasks being assessed, has demonstrated utility in the prediction of performance (Bandura, 1982). In the context of performance appraisal, task-specific self-efficacy refers to a person's perceptions of his or her ability to be an accurate and efficient evaluator of another individual's performance. In the context of this argument, self-efficacy is considered a state, rather than a trait, variable that is based on perceptions that vary across situations and activities (Bandura, 1982).

Consistent relations have been found between self-efficacy and performance. Research has demonstrated self-efficacy to be predictive both directly and indirectly of academic performance (e.g., Wood & Locke, 1987), sales performance (e.g., Barling & Beattie, 1983), smoking reduction (e.g., Barrios & Niehaus, 1985; Chabliss & Murray, 1979a), competitive performance (e.g., Weinberg, Yukelson, & Jackson, 1980), weight loss (e.g., Chabliss & Murray, 1979b), goal setting, task performance, and goal commitment (e.g., Locke, Frederick, Lee, & Bobko, 1984). These results give very strong support to Bandura's (1982) claim that self-efficacy is a key causal variable in performance (Locke et al., 1984).

Alternative Perspective

The results of Jawahar and Williams' (1997) meta-analysis seemingly quiet the purpose of appraisal debate and conclude that ratings used for administrative decisions tend to be more inflated than those used for developmental decisions. A question then follows: What effect does the purpose of appraisal have on the rater to cause him/her to be more lenient in one condition over another? It is reasonable to consider that the current purpose-leniency paradigm is underspecified and that there exists any possible number of variables that could serve as a generative mechanism through which appraisal purpose is able to affect leniency. If rater confidence is positively related to rating accuracy, then it is possible that if a rater perceives greater negative consequences for giving low ratings when the purpose is administrative verses developmental, rating purpose could serve to reduce rater confidence in his/her ability to accurately rater performance and subsequently increase rating leniency. Therefore, a new research paradigm in which self-efficacy exists as a mediating variable between purpose of appraisal and leniency may be a more appropriate model.

Hauenstein (1992) believed that “leniency depends more on rater attitudes within a particular organizational context and less on the rater’s ability to judge people”(p. 485). Rating errors are not necessarily indicative of an inability to discriminate among rating decisions but rather an example that the rater simply did not discriminate among good and bad performers. Murphy and Cleveland (1995) suggested that raters were equally capable of discerning good and bad performers when the purpose of the rating was either for administrative decisions or for development decisions. However, in the former, they simply did not. Apparently, the purpose of rating did not affect the rater’s actual evaluations of the ratee (Murphy et al., 1984) but it may have had an effect on the rater’s reported evaluations (i.e., the raters thought one thing and reported another). Perhaps administrative uses of appraisal information activates a rater's motivation to report more positive information than the rater privately perceives to be appropriate.

Saal and Knight (1988) gave several explanations for rater leniency: a desire for the rater to be liked by the ratee, an unwillingness to give negative feedback, a fear that other raters inflate their ratings, and an unusually high set of standards. The above mentioned explanations illustrate cognitive processes that may affect leniency errors, but the question remains: why are ratings used for administrative decisions plagued with higher instances of leniency than ratings used for developmental decisions? How do cognitive processes differ as a function of rating purpose?

One explanation is that raters do not want to feel responsible for negative consequences of a rating if they are uncertain about their ability to accurately evaluate the performance. More specifically, when the raters perceive that their evaluations of a person may have negative consequences (i.e., refused promotion or salary increase) and they do not have confidence in their ability to accurately evaluate a person’s performance, they may be reluctant to cause undue calamity for the person they are evaluating. Negative consequences associated with developmental decisions (i.e., increased training and supervision) may not be perceived as being as aversive as those associated with administrative decisions. Murphy and Cleveland (1995) “suggest that leniency is likely to occur when the appraisal has important consequences for the ratee (i.e., administrative decisions)" (p. 245). More simply, a person with low task-specific (i.e., evaluating an individual’s performance in a certain environmental context) self-efficacy may refrain from giving negative feedback as a function of his or her own perceived inability to accurately evaluate a person. Consistent with this idea of self-efficacy and its relation to leniency errors, a literature review by Landy and Farr (1983) noted that better performers were more accurate in their ratings of others (e.g., Kirchner and Reisberg, 1962).

Overview of Study

The purpose of the present study was to examine the relationship between a rater's self-efficacy specific to performance evaluation and leniency and behavioral accuracy. Hauenstein (1992) suggested leniency errors occur in performance evaluations because of raters’ lack of extensive knowledge of the performance content domain. Subsequently, raters’ evaluations are frequently based on unclear standards. Perhaps associated with raters’ lack of extensive knowledge comes lower efficacious perceptions about their ability to accurately evaluate a ratee’s performance. Low self-efficacy specific to performance evaluations may result in greater leniency errors. This study also explored the affect of purpose of appraisal on leniency. Leniency was expected to be higher in the

administrative condition and lower in developmental condition. These results were expected to follow the pattern observed in Jawahar & Williams' (1997) meta-analysis.

In addition, self-efficacy was explored as a mediating variable between appraisal purpose and rating leniency. It was hypothesized that the purpose of appraisal would augment/abate a rater's confidence and subsequently affect his/her tendency to bias his/her appraisal and give more/less lenient ratings.

Finally, given that performance evaluations are typically based on a rater's ability to recall behavioral examples of performance, leniency was thought to be a function of the accuracy of performance recall (Hauenstein, 1992). Leniency and behavioral accuracy were examined separately in attempts to investigate whether or not leniency is a result of bias or a result of inaccurate recall. It was expected that leniency and behavioral recognition accuracy would be negatively related.

Hypotheses

H1: Subject self-efficacy will partially mediate the relationship between purpose of appraisal and rating leniency.

H2: Leniency of ratings will increase as behavioral accuracy decreases and leniency of ratings will decrease as behavioral accuracy increases.

H3: Behavioral recognition accuracy will increase as subject self-efficacy increases and behavioral recognition accuracy will decrease as subject self-efficacy decreases.

Method

Subjects

Participants for this study were 109 undergraduate students originating from a pool of volunteers from an introductory psychology course at a large southern university. Fifty-six subjects were placed in the administrative purpose condition and fifty-three subjects were placed in the developmental purpose condition. Thirty subjects (27.5%) were men, and seventy-nine (72.5%) were women. Extra credit was given for their participation.

Self-Efficacy Scale

Three self-efficacy scales were constructed to assess students' confidence in their ability to rate the performance of an instructor. A 12-item general performance appraisal self-efficacy scale (GSE; see Appendix A1) was designed to measure students' overall confidence in their ability to rate performance. Prior to actual study, data on the internal consistency (see Table 1) of the GSE scale were collected in two pilot groups of 13 and 40 undergraduate subjects. The first group (n=13) was used to assess the original 7-item scale. The coefficient alpha in this phase was .55 (see Table 1). Items with low total correlations were rewritten and three new items were added. Data for the second scale (10-item) were collected from the second pilot group (n=40). The coefficient alpha in this phase was .74 (see Table 2). Items with low total correlations were rewritten and two new items were added. The 12-item GSE scale was analyzed to assess its internal consistency reliability (see Table 3). Results indicated that the scale had good internal consistency reliability (Cronbach Alpha = .80).

Two six-item task-specific scales were designed to measure students' confidence in their ability to perform at specific levels of performance. These scales were designed to measure accuracy self-efficacy (ASE; see Appendix A2) and behavioral recognition self-efficacy (BRSE; see Appendix A3). Reliability analyses of both scales indicated a 100% logical ordering of response patterns (i.e., subsequent responses were at same or

higher level of confidence than the previous response). The creation of all three self-efficacy scales were based on Bandura's (1995) "Guide for Constructing Self-Efficacy Scales."

Item Construction. Scale items were phrased in terms of "can do" as suggested by Bandura (1995) because "can is a judgement of capability" (p. 2) that focuses on current abilities and not on potential ability or expected future capabilities. Responses were measured by using a five item Likert-scale ranging from strongly disagree to strongly agree, with a neutral response in the middle (see Appendix A1-A3).

Typically, self-efficacy magnitude and strength are operationally measured by asking individuals whether they can perform at specific levels on a specific task (magnitude) and by asking for their percent confidence (strength) (Maurer & Pierce, 1998; Lee & Bobko, 1994). Lee & Bobko (1994) excluded measures of self-efficacy that utilize a Likert format on the premise that such techniques do not correspond to Bandura's (1986) recommendation for assessing both strength and magnitude. The results of their research indicated that the best method of measuring self-efficacy was to use composites based on combining only the strength items where the magnitude response was "yes" (Lee and Bobko, 1994).

Bandura (1995) recently has endorsed the usage of scales that measure item responses using a 100-point Likert scale of 10-unit intervals. The suggested response scale ranges "from 0 ('Cannot do'); through intermediate degrees of assurance, 50 ('Moderately certain can do'); to complete assurance, 100 ('Certain can do')" (Bandura, 1995, p. 2; also see Bandura, 1997, chapter two).

More recently, Maurer & Pierce (1998) explored the possibility of using a Likert-type measurement format as an alternative to the traditional format (e.g., Lee & Bobko, 1994) used to measure self-efficacy. A Likert-scale measure is essentially a combination of both the magnitude (response on the agree side may be equivalent to a "yes" response while a response on the disagree side may be equivalent to a "no" response) and strength (assessed as the distance away from the neutral position) aspects of traditional measures (Maurer & Pierce, 1998). Maurer & Pierce (1998) demonstrated using classical reliability, predictive validity, and confirmatory factor analysis techniques that the Likert format was empirically equivalent to the traditional format.

Manipulation Check

The effectiveness of the purpose of appraisal manipulation was assessed using two items that required subjects to indicate their perceived purpose of the study (see Appendix E). One item was used to assess the subjects' perceptions of an administrative purpose of appraisal while the other item was used to assess the subjects' perceptions of a developmental purpose of appraisal. Items where measured using a Likert style format.

Stimulus Material

A 15-minute videotape of a graduate teaching assistant lecturing on a topic of consumer psychology was used as the rating stimuli. Embedded in the taped lecture were 16 behavioral incidents (see Appendix B) representative of four performance dimensions: depth of knowledge, organization, delivery, and relevance. The 16 behavioral incidents embedded in the videotape were based on the behaviorally anchored rating scale study by Nathan & Lord (1983). A mix of positive and negative behaviors related to the four performance dimensions (i.e., four good behavioral incidents corresponding to depth of knowledge and organization; two good and two poor behavioral incidents corresponding

to delivery and relevance) were used to convey an average performance by the instructor (Nathan & Lord, 1983; Hauenstein & Alexander, 1991; Hauenstein, 1992). Average performance was chosen to allow raters latitude to distort in both a positive and negative direction while avoiding ceiling and floor effects.

The videotaped lecture used in this study was based on the videotape originally developed by Hauenstein & Alexander (1991). It was more recently updated (Fredholm, 1998) to reflect more current clothing and hair styles in order to increase subjects' belief that the graduate teaching assistant in the video is currently at a local university. The videotaped lecture was evaluated prior to the study by subject matter experts (graduate students in Industrial/Organizational psychology at the university) to determine new target scores for the performance dimensions (see Table 4). Target scores reflected an average performance (values between 3 and 5) with depth of knowledge and organization predictably higher (four good behavioral incidents) than delivery and relevance (two good and two poor behavioral incidents). Fredholm (1998) assessed the quality of the target scores by examining the convergent validity of the scores. An intraclass index of .61 indicated that there existed sufficient agreement of raters across dimensions Kavanagh, MacKinney, & Wolins, 1971).

Procedure

Subjects were randomly assigned to one of two treatment conditions (either the administrative or the development purpose of appraisal condition). Subjects in each treatment condition were run in small groups of six or twelve.

Subjects in both treatment groups were given a testing packet prior to the study's start. On the cover of the packet was a priming paragraph in which the purpose of appraisal manipulation was given. The paragraph was identical in content for both treatments except for the expressed appraisal purpose. Subjects were then asked to fill out the three self-efficacy scales (see Appendix A1-A3) prior to the performance appraisal part of the study. They were told that the purpose of the scale was to assess their level of confidence with rating instructors. Upon completion of the self-efficacy scales, subjects were asked to view a short videotaped lecture and to appraise the instructor's performance. After viewing the videotape, subjects were asked to fill out their extra-credit opscans (Fredholm, 1998). This pause in procedure was used to control for short-term memory effects. Subjects were then asked to fill out a rating form (see Appendix C) and a recognition-memory questionnaire (see Appendix D). Following the completion of the rating form and the recognition-memory questionnaire, subjects were asked to fill out two addition manipulation check items (see Appendix E). Anonymity of their ratings was guaranteed and subjects were assured that they would not have to interact with the instructor they evaluated.

Purpose of rating

The perceived purpose for subject evaluations of the instructor was manipulated with one of two sets of instructions.

Administrative Decisions Manipulation. Subjects were told that their participation in the study was part of a joint effort with a local university to evaluate graduate-teaching assistants in the psychology department. Subjects were told that because of biases commonly present when students evaluate their instructors, it is very difficult to determine which graduate-teaching assistants should receive future teaching assignments and funding increases.

Developmental Manipulation. Subjects were told that their participation in the study was part of a joint effort with a local university to evaluate graduate-teaching assistants in their psychology department. Subjects were told that because of biases commonly present when students evaluate their instructors, it is very difficult to provide accurate feedback about the instructor's teaching strengths and weaknesses.

Dependent Variables

Leniency. A 4-item rating scale (see Appendix C) was used by the subjects to evaluate the instructor's performance. Each scale item corresponded to a different performance dimension and was measured using a 7-point likert scale (1=low, 7=high). Subject ratings of the instructor's performance were converted into leniency measures (see McIntyre et al., 1984). Using the ratings' "true scores" as determined by the subject matter experts, it was possible to quantify the raters' tendencies to assign higher ratings than the instructor's performance would necessitate. Leniency was computed using the following formula:

$$\text{Leniency}_K = \sum_{i=1}^{d} \frac{(T_i - R_{ik})}{d}$$

where d is the number of rating scale items (4); k refers to the kth rater; R is the obtained rating; and T is the "true score" (McIntyre et al., 1984). Negative mean scores indicate that the subject was more lenient in his/her ratings than the true score. Positive scores indicate severity in ratings from the true score.

Behavioral Accuracy. A 32-item behavioral recognition questionnaire (see Appendix D) was used to measure subjects' true recall of the instructor's behaviors. The questionnaire consisted of eight behaviors of each of the four performance dimensions with a total of 24 good behaviors and 8 poor behaviors (this is due to the true ratio of good to poor behaviors represented in the videotaped lecture.)

Behavioral accuracy was assessed by analyzing each subject's false-positive rate (i.e., the percentage of nonoccurring behaviors incorrectly identified and occurring behaviors unidentified) and true-hit rate (i.e., the percentage of occurring behaviors correctly identified and nonoccurring behaviors unidentified). Behavioral accuracy was determined by subtracting each subject's false-positive rate from his or her true hit rate (Sulsky & Day, 1992).

Results

Scale Analysis

The distribution of scores on both the general self-efficacy scale and the task-specific self-efficacy composite were well within score maximums and minimums and thus it can be concluded that neither scale suffered from either ceiling nor floor effects (see Table 5). In addition, the range of scores on both scales was sufficient to conclude that there existed adequate variability on self-efficacy in the sample. The three efficacy scales were correlated to assess the level of scale convergence (see Table 6). Sufficient convergence between the ASE and BRSE scales ($\underline{r}$ = .68) allowed for a task-specific composite score to be computed by aggregating each subject's total score on each scale (Nunnally, 1978). Although the ASE and the BRSE scale only explain approximately 46% of the variance in each other, this was considered sufficiently close to the 50% value recommended by Nunnally (1978) for adequate scale convergence. This composite score of the ASE and BRSE scales will be referred to as the task-specific composite scale. The

GSE scale was not aggregated with the two task-specific self-efficacy scales because (a) correlations between the GSE scale and the ASE scale ($\underline{r}$ = .61) and the BRSE scale ($\underline{r}$ = .44) (see Table 6) were deemed inadequate for convergence, and (b) the scales were written using two separate formats (i.e., the GSE scale was written in terms of efficacy strength while the ASE and BRSE scales were written in terms of efficacy strength and magnitude) and thus there was not a logical rational for their aggregation (see footnote 1). Further analyses were run using both the GSE scale and the task-specific self-efficacy composite (from this point on this scale will simply be referred to as "composite").

Manipulation Check

First, t-tests were used to determine mean differences in manipulation check items by purpose. A significant mean difference ($M_{admin.}$ = 3.98, $M_{develop.}$ = 3.45) was observed for the administrative purpose item between experimental conditions (t = 2.59, $p < .05$). In addition, a significant mean difference ($M_{admin.}$ = 3.88, $M_{develop.}$ = 4.40) was observed for the developmental purpose item between experimental conditions (t = -3.15, $p < .05$). Further analyses of mean differences revealed medium effect sizes (Cohen, 1969) for both manipulation items ($ES_{administrative}$ = .5, $ES_{developmental}$ = .59).

Although both purpose items had significant mean differences by experimental condition, in the appropriate direction, the fact that subjects in both experimental conditions positively endorsed both purpose items is troublesome. It was expected that for those subjects in the administrative purpose condition, the response trend to the manipulation check items would be positive for the administrative item and negative for the developmental item. Consequently, an opposite trend was expected for subjects in the developmental condition. Failure to observe such a trend may be a result of (a) subjects' true beliefs that they were expected to accomplish both goals (although one more so than the other), and/or (b) a bias to use high anchors, regardless of what question is being asked.

To further examine the effectiveness of the experimental manipulation, a composite manipulation variable was computed by combining a subject's administrative purpose item score with their developmental purpose item reverse scored, such that a manipulation composite score of 10 would indicate maximum administrative purpose manipulation effectiveness and a manipulation composite score of 2 would indicate maximum developmental purpose manipulation effectiveness. This composite was regressed on purpose and a significant mean difference ($M_{admin.}$ = 6.11, $M_{develop.}$ = 5.06) was observed for the composite purpose manipulation item ($F_{(1, 107)} = 20.85$, $p < .0001$).

It can be concluded that the purpose manipulation was effective in causing subjects in the administrative condition to believe that the experimental purpose was for administrative decisions more so than subjects in the developmental condition. In addition, subjects in the developmental condition believed that the experimental purpose was for developmental decisions more so than subjects in the administrative condition. Finally, mean differences in the manipulation composite scores also indicates that the manipulation had a large effect ($ES_{manipulation\ composite}$ = .88). Although both means were above the scale midpoint, it is clear the manipulation worked in a relative sense.

Test for Mediation

A series of four regression equations were used to test the linkages of the mediational model (Baron & Kenny, 1986). To establish mediation, four conditions must be met. First, purpose of appraisal must significantly affect leniency; second, purpose of

appraisal must significantly affect self-efficacy; third, self-efficacy must significantly affect leniency; and finally, when leniency is regressed on purpose of appraisal and self-efficacy, the affect of purpose of appraisal must be less than when regressed on leniency independently. If any of these conditions are not satisfied, the proposed model of mediation is not supported.

First, leniency was regressed on purpose of appraisal. Results of this analysis proved nonsignificant and it was concluded that purpose of appraisal did not predict leniency in performance ratings (see Table 7). Examination of the mean differences in leniency between experimental groups revealed that although the groups did not differ significantly, subjects in the administrative condition were more severe in their ratings than subjects in the developmental condition (see Table 8). Given that a significant relationship between leniency and appraisal purpose was not found, hypothesis 1, which predicted self-efficacy as a mediating variable between purpose of appraisal and leniency, was not supported.

Continued test for mediation was not necessary given that the first condition of mediation was not satisfied. However, steps two and three in testing for mediation were carried out in order to investigate possible main effects within the model.

Self-efficacy was regressed on purpose of appraisal to test for possible purpose effects. Both the general self-efficacy scale (GSE) and the task-specific composite (composite) were tested independently. Both analyses yielded nonsignificant results and it was concluded that appraisal purpose did not affect self-efficacy (see Table 7).

Finally, leniency was regressed on both the general self-efficacy scale (GSE) and the task-specific composite (composite). Results from these analyses were mixed. General self-efficacy did not significantly predict leniency ($F_{(1, 107)} = 2.29$, $p < .13$). However, the task specific composite of self-efficacy was shown to be a significant predictor of leniency ($F_{(1, 107)} = 6.36$, $p < .05$) explaining 5.6% of the variance in leniency (see footnote 2). As a whole, subject task-specific self-efficacy was positively related to leniency such that as self-efficacy increased, rating leniency increased (see Table 7). Further examination of the independent rating dimensions and the task-specific composite revealed near significant positive correlations between the task-specific composite and the depth of knowledge rating item ($r = .16$, $p < .10$), the delivery rating item ($r = .16$, $p < .10$), and the relevance rating item ($r = .19$, $p < .052$), and a significant correlation between the task specific composite and the organization rating item ($r = .27$, $p < .01$) (see footnote 3).

As an alternative analysis strategy, the composite manipulation check variable was used instead of group membership effect codes in the mediation analysis. Results of these analyses were consistent with those stated above: the composite manipulation variable did not significantly predict leniency, general self-efficacy, or task specific self-efficacy.

Behavioral Accuracy

First, the relationship between rater leniency and subject behavioral recognition accuracy was assessed using simple linear regression. Hypothesis 2, which predicted a significant relationship between rating leniency and a subject's behavioral recognition accuracy, was not supported (see Table 9). Therefore, it was concluded that there existed no relationship between rater leniency and subject behavioral recognition accuracy.

Next, the relationship between subject behavioral recognition accuracy and subject self-efficacy was assessed using simple linear regression. Behavioral accuracy was regressed on both the general self-efficacy scale and the task-specific composite independently. Results of these analyses were mixed. Self-efficacy trends observed in the prediction of leniency were not found in the prediction of behavioral recognition accuracy. Rater task-specific self-efficacy did not significantly predict his/her behavioral recognition accuracy, however, a rater's general self-efficacy was found to be significantly predictive of his/her behavioral recognition accuracy ($F_{(1, 107)} = 3.99$, $p < .05$), explaining 3.6% of the variance in behavioral accuracy (see Table 9). Hypothesis 3 was supported and given these results it can be concluded that as a subject's general performance appraisal self-efficacy increases, his/her accurate recognition of performance behaviors increases (see footnote 4).

Finally, both the true-hit and the false-positive components of behavioral accuracy were regressed on general self-efficacy. Both components were significantly related to subject general self-efficacy [(true-hit rates ($F_{(1, 107)} = 3.99$, $p < .05$), and false-positives ($F_{(1, 107)} = 3.99$, $p < .05$)]. Examination of the regression parameters indicated that as general self-efficacy increased, true-hit rates increased, and false-positive rates decreased (see Table 9).

Discussion

The results of the current study indicate that there exists a positive relationship between rater self-efficacy and rater leniency and behavioral recognition accuracy. In addition, a relationship between a rater's behavioral recognition accuracy and his/her tendency towards rating leniency was not found. The opposite directionality of self-efficacy effects (i.e., increased self-efficacy leads to better behavioral recognition accuracy but also more lenient ratings) coupled with a nonsignificant behavioral recognition accuracy-leniency relationship suggests that leniency may be more a function of rating bias and less a function of observation bias. Finally, the proposed mediational model in which self-efficacy mediates the relationship between purpose of appraisal and rater leniency was not supported.

The lack of convergence between the general self-efficacy scale and the task-specific composite was contradictory to the findings of Lee & Bobko (1994). They found that measures of efficacy that were computed as strength indices (i.e., such as the general self-efficacy scale used in this study which measured a subjects' level of confidence with a general performance statement) were highly convergent with efficacy composites that were computed as strength and magnitude indices (i.e., such as the task-specific composite used in this study which measured a subject's level of confidence with specific levels of performance). However, further examination of the measurement content in both the general self-efficacy scale and the task-specific composite revealed that the general scale addressed more contextual issues (e.g., appraisal based on limited subject knowledge and ratee exposure) than the task-specific composite. The lack of convergence between the general self-efficacy scale and the task-specific composite serves to further support the notion that the scope of the general self-efficacy scale, although inclusive of the scope of the composite, was more broad and measured a fuller range of performance appraisal contexts. The additional appraisal contexts found in the general self-efficacy may have potentially reduced the correlations between the two efficacy measures.

The opposite directionality of self-efficacy effects on leniency is a potentially interesting result. The directionality of these results implies that raters that are more confident have more accurate recall of actual performance behaviors to base their performance rating on. However, these raters tend to be more lenient in their evaluations of performance. This would imply that when the subjects in the current study purveyed their ratings for the graduate teaching assistant, their rating decisions were seemingly independent of their recognition of actual performance behaviors. These implications support a model in which leniency is a function of rating bias and not observation bias.

Hauenstein (1992) explored both rendering bias and encoding bias explanations for leniency in performance ratings. Rendering bias asserts that inaccuracies in performance ratings are a function of motivated bias at the time of rating, while encoding bias asserts that inaccuracies in performance ratings are a function of contextual effects on selective attention and encoding processes at the time of performance observation. Hauenstein (1992) further noted that encoding biases occur in accordance with rater accountability, and that as rater accountability increases, biases at the time of information processing increase.

The results of the present study follow Hauenstein's (1992) logic concerning rater accountability and rendering/encoding bias cause of leniency. If behavioral recognition accuracy is considered a measure related at least in part to encoding accuracy, then it is possible to assume that high behavioral accuracy may be an indication of low encoding bias, or more specifically, lower observation bias. Hauenstein (1992) concluded that in situations of low rater accountability, leniency in performance ratings would result more from distortions at the time of ratings and less from distortions at the time of performance observation. To apply this reasoning to the present study in which there existed no contextual accountability variable, one would not expect there to be a relationship between leniency and behavioral recognition accuracy. Simply, the research paradigm associated with the current study lends itself to a rating bias model of leniency and not an observation bias model of leniency. Although the results of the current study did not support hypothesis 2, they were in accordance with the model proposed by Hauenstein (1992).

In a related manner, the question can be asked, "Why was it that the task-specific composite produced the leniency effect, but it was the general self-efficacy measure that produced the behavioral recognition results?" In reviewing the subject content measured by both scales, it becomes apparent that the general efficacy scale measures both encoding variables (e.g., recall of behaviors with limited exposure and accurate performance evaluation given personal dislike for the ratee or the performance context) and rendering variables (e.g., accurate judgement of performance). However, the content of the task-specific composite was directed more at the rendering process of performance evaluation. The encoding and rendering measurement associated with the general efficacy scale seems to indicate that this scale measures the entire rating process (i.e., both observation and evaluation) and not just the evaluation process. It is this observation efficacy measured by the general efficacy scale that allows better prediction of behavioral recognition accuracy than the task-specific composite. In addition, the task-specific composite appears better suited to measure a rater's perceptions of his/her ability to rate performance, not necessarily his/her ability to observe performance.

Failure to support the proposed model of mediation is primarily focused on the failure to produce a more salient purpose of appraisal manipulation. Although Jawahar & Williams (1997) concluded that there does exist a purpose of appraisal effect, they did report that purpose of appraisal effects were higher in organizational settings than in lab settings. The results of the current study are not inconsistent with those of other studies that used student samples in laboratory settings (e.g., Murphy et al., 1984) and failed to yield a purpose of appraisal effect.

Attempting to produce a successful purpose of appraisal effect in a laboratory setting is potentially more difficult due to students' approach to the appraisal process itself. Hauenstein (1992) criticized the use of student samples in performance appraisal research on the notion that they often evaluate performance on unclear standards. Student subjects simply may not understand the process of performance appraisal or the effects their ratings may have on the ratee. Performance ratings obtained from student samples may be based less on operationalized criteria and more so each subject's general affect at the time of the ratings. This idea is not intended to imply that actual organizational raters are free from this bias, quite the opposite. Research has demonstrated that rater affect can influence performance ratings (e.g., Demuse, 1987; Isen, Shalker, Clark, & Karp, 1978) even to the degree that affective standards are more influential that organizational standards (Murphy & Cleveland, 1995). Affect is a variable that may effect both student and organizational raters alike, however, if student raters do not have a full appreciation for the performance standards from which they are supposed to base their ratings on, then ultimately their ratings may be nothing more than an affective "gut feeling".

The inability to demonstrate a purpose of appraisal effect in laboratory research may be a result of more than just unclear performance appraisal standards. The motivational model of leniency discussed earlier describes a process in which raters compare the potential negatives of giving low rating with the potential positives for giving high ratings. These processes function within organizations where the potential negatives associated with giving low ratings are very much real for rater. However, the potential negatives consequences for giving low ratings are likely perceived as being infinitely less real for student subject raters. One element that is often not present in a research paradigm that is present in an organizational paradigm is rater accountability. Rater accountability engages the motivational and political processes generally considered existing in leniency models. A lack of perceived accountability by student raters can perhaps explain why purposes of appraisal effects are less salient in laboratory research. Research on accountability in a performance appraisal context has revealed that raters who felt more accountable for their ratings gave higher ratings than those who felt less accountable (e.g., Klimoski & Inks, 1990; Sharon, 1970; Sharon & Bartlett, 1969). In addition, Mero & Motowidlo (1995) noted that subjects who were accountable for their ratings were more attentive to and engaged in the rating task.

Given that rater accountability in laboratory settings can potentially draw a research paradigm more inline with an actual organizational paradigm, it may be appropriate to conclude that making students more accountable for their ratings may serve to augment a purpose of appraisal effect. Realizing this, the proposed mediational model should not be completely discounted, but rather it should be recognized that limitations of the current methodology (i.e., lack of rater accountability) may have

resulted in the failure to support said model. Future research exploring possible mediating variables for the purpose of appraisal-leniency model should be explored with a research methodology that incorporates a rater accountability component.

Finally, subjects were randomly selected at the conclusion of the experiment and questioned as to the rationale for their endorsements of the two purpose of appraisal manipulation items. Subjects generally indicated either one of two explanations why they positively endorsed both purpose of appraisal items: a) the nature of the questions asked in the general self-efficacy questionnaire indicated that the purpose of their ratings was for both developmental reasons and administrative reasons, and b) regardless of the expressed purpose by the experimenter, performance appraisals are generally used for both purposes, even if for one more than the other. Given these responses, it is reasonable to assume that the manipulation was possibly diffused due to the items contained in the general self-efficacy scale.

Conclusion

One issue surrounding research is the generalizability of results from the laboratory to practice. First, it needs to be recognized that leniency as a measure of rater bias, is not necessarily synonymous with rater accuracy (Borman, 1979). Although a component of accuracy, leniency represents a compensatory measure of a rater's evaluating tendencies in which instances of leniency on performance dimensions can be offset by instances of severity on other performance dimensions. Near zero scores on a leniency measure may not reflect rating accuracy: a rater who scores low on leniency as a result of consistent severity-leniency compensation is no more accurate that a rater who is consistently lenient or severe. Leniency as it was measured in the current study is a more appropriate measure of accuracy than traditional measures of leniency (i.e., deviation from the scale mid-point) due to the usage of performance true scores. However, to fully explore accuracy in performance appraisal, future research should consider additional measures of rating accuracy such as Cronbach's (1995) differential and elevation accuracy or Hauenstein and Alexander's (1991) dimensional accuracy in conjunction with leniency.

Hauenstein (1998) suggested two type of rater training content approaches to improving rating accuracy: increase observational skills and increase rater confidence. The results of the present study have implications for the practice of performance appraisal within organizational contexts. The relationship between rater self-efficacy and behavioral recognition accuracy suggests that one way to increase rating accuracy is to increase rater confidence in his/her ability to rate. If more confident raters have better true recall of actual performance behaviors, steps for increasing rater self-efficacy should be part of behavioral observation training in which the goal of rater training is to improve observational skills (Hauenstein, 1998). Self-efficacy appears to have an affect on the encoding process at the time of behavioral observation, and by improving rater confidence, it may be possible to improve rater-encoding strategies. The connection between rater self-efficacy and rating accuracy was not properly tested in the current study. Due to the relationship between rater self-efficacy and behavioral recognition accuracy, one would expect self-efficacy to effect rating accuracy only in situations of perceived rater accountability when encoding bias strategies have a more causal relation to performance rating errors. Within this framework, it would be expected that

behavioral recognition accuracy would partially mediate the relationship between rater self-efficacy and rating accuracy.

Finally, it should be recognized that self-efficacy explained only a small portion of the variance in both rater leniency and behavioral recognition. This is not to downplay the importance of self-efficacy in understanding these variables, but rather it is important to realize that self-efficacy represents only one of potentially many variables related to rating accuracy and the processes that underlie this phenomena. Future research in the area of performance appraisal accuracy should continue to explore the effects of rater self-efficacy on rating accuracy. In addition, more research is necessary to explore the cognitive/motivational processes involved in rating judgements.

References

Aleamoni, L.M. & Hexner, P.Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instruction on student course and instructor evaluation. Instructional Science,9, 67-84.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84, 191-215.

Bandura, A. (1982). Self-efficacy mechanisms in human agency. American Psychologist, 37, 122-147.

Bandura, A. (1986). Social foundations of thought and actions: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1995). Manual for the construction of self-efficacy scales. Available from Albert Bandura, Department of Psychology, Stanford University, Stanford CA 94305-2130.

Bandura, A. (1997). Self-efficacy: The exercise of control. New York, NY: W. H. Freeman and Company.

Barling, J. & Beattie, R. (1983). Self-efficacy beliefs and sales performance. Journal of Organizational Behavior Management, 5, 41-51.

Baron, M.R. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.

Barrios, F. & Neihaus, J. (1985). The influence of smoker status, smoking history, sex, and situational variables on smoker self-efficacy. Addictive Behaviors, 10, 425-429.

Berkshire, J.R. & Highland, R.W. (1953). Forced-choice performance rating – A methodological study. Personnel Psychology, 6, 355-378.

Bernardin, H.J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 63, 301-308.

Bernardin, H.J., Cardy, R.L., & Carlyle, J.J. (1982). Cognitive complexity and appraisal effectiveness: Back to the drawing board? Journal of Applied Psychology, 67, 151-160.

Bernardin, H.J. & Cooke D.K. (1992). Effects of appraisal purpose on discriminability and accuracy of ratings. Psychological Reports, 70, 1211-1215.

Bernardin, H.J. & Orban, J. (1990). Leniency effect as a function of rating format, purpose of appraisal, and rater individual differences. Journal of Business and Psychology, 5, 197-211.

Bretz, R.D., Milkovich, G.T., & Reed, W. (1990). The current state of performance appraisal research and practice: Concerns, directions, and implications. Journal of Management, 18, 321-352.

Brill, R.T. & Prowker, A.N. (1996). Supervisor information processing: Rating complexity and confidence. Paper presented at the Eighth Annual Meeting of the American Psychological Society. San Francisco; CA.

Borman, W.C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 410-421.

Cascio, W.F. & Valenzi, E.R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 62, 278-282.

Centra, J.E. (1976). The influence of different directions on student ratings of instruction. Journal of Educational Measurement, 13, 277-282.

Chambliss, C.A. & Murray E.J. (1979a). Cognitive procedures for smoking reduction: Symptom attribution versus efficacy attribution. Cognitive Therapy and Research, 3, 91-95.

Chambliss, C.A. & Murray E.J. (1979b). Efficacy attribution, locus of control, and weight loss. Cognitive Therapy and Research, 3, 349-353.

Clevelend, J.N., & Landy, F.J. (1981). The influence of rater and ratee age on two performance judgements. Personnel Psychology, 34, 19-29.

Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". Psychological Bulletin, 53, 177-183.

Crooks, L.A. (1972). An investigation of the sources of bias in the prediction of job performance: A six-year study. Princeton, NJ: Educational Testing Service.

Cohen, J. (1969). Statistical power analysis for behavioral sciences. New York, NY: Academic Press.

DeCotiis, T. & Petit, A. (1978). The performance appraisal process: A model and some testable propositions. Academy of Management Review, 3, 635-646.

Demuse, K.P. (1987). A review of the effects of non-verbal cues on the performance appraisal process. Journal of Occupational Psychology, 60, 207-226.

Dipoye, R.L., Arvey, R.D., & Tempstra, D.E. (1977). Sex and physical attractiveness of raters and applicants as determinants of resume evaluations. Journal of Applied Psychology, 62, 288-294.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of individual differences in stereotypes of women and purpose of appraisal on sex differences in performance ratings: A laboratory and field study. Journal of applied Psychology, 73, 551-558.

Driscoll, L. A. & Goodwin, W. L. (1979). The effects about varying information about the use and disposition of results on university students' evaluations of faculty and courses. American Educational Research Journal, 16, 25-37.

Drory, A. & Ben-Porat, A. (1980). Leadership style and leniency bias in evaluations of employees' performance. Psychological Reports, 46, 735-739.

Elmore, P.B., & LaPointe, K.A. (1974). Effects of teacher sex and student sex on the evaluation of college instructors. Journal of Educational Psychology, 66, 386-389.

Farh, J.L. & Werbel, J.D. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. Journal of Applied Psychology, 71, 527-529.

Fredholm, R.L. (1998). Effects of dual accountability and purpose of appraisal on accuracy. Masters Thesis submitted for degree completion at Virginia Tech.

Feldman, K.A. (1986). Instrumentation and training for performance appraisal: A perceptual cognitive viewpoint. In K. Rowland & J. Ferris (Eds.), Research in personnel and human resource management (Vol. 4). Greenwich, CT: JAI Press.

Gist, M.E. (1987). Self-efficacy: Implications for organizational behavior and human resource management. Academy of Management Review, 12, 472-485.

Gupta, N., Beehr, T.A., & Jenkins, G.D. (1980). The relationship between employee gender and supervisor-subordinate cross-ratings. Proceedings of the Academy of Management, 40, 396-400.

Hamner, W.C., Kim, J.S., Baird, L., & Bigoness, N.J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work sampling task. Journal of Applied Psychology, 59, 705-711.

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. (1998). Training raters to increase the accuracy of appraisals and the usefulness of feedback. In J.W. Smither (ed.), Performance Appraisal: State of the Art in Practice (p. 404-444). San Francisco, CA: Josey-Bass Publishers.

Hauenstein, N. M. A. (1992). An information-processing approach to leniency in performance judgements. Journal of Applied Psychology, 77, 485-493.

Hauenstein, N. M. A. & Alexander, R. A. (1991). Rating ability in performance judgements: The joint influence of implicit theories and intelligence. Organizational Behavior and human Decision Processes, 50, 300-323.

Heron, A. (1956). The effects of real-life motivation on questionnaire response. Journal of Applied Psychology, 40, 65-68.

Hollander, E.P. (1965). Validity of peer nominations in predicting a distant performance criterion. Journal of Applied Psychology, 49, 443-438.

Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternate predictors of job performance. Psychological Bulletin, 96, 72-98.

Isen, A.M., Shalker, T.E., Clark, M.S., & Karp, L. (1978). Positive affect, accessibility of material in memory, and behavior. A cognitive loop? Journal of Personality and Social Psychology, 36, 1-12.

Jacobson, M.B., & Effetz, J. (1974). Sex roles and leadership: Perceptions of the leaders and the lead. Organizational Behavior and Human Performance, 12, 383-396.

Jawahar, I.M. & Williams, C.R. (1997). Where all the children are above average: The performance appraisal purpose effect. Personnel Psychology, 50, 905-925.

Kavanagh, M.J., MacKinney, A.C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

Kirchner, W.K. & Reisberg, D.J. (1962). Differences between better and less effective supervisors in appraisal of subordinate. Personnel Psychology, 15, 295-302.

Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. Organizational Behavior and Human Decision Processes, 45, 192-208.

Klores, M.S. (1966). Rater bias in forced-distribution ratings. Personnel Psychology, 19, 411-421.

Kraiger, K., & Ford, J.K. (1985). A meta-analysis of race effects in performance appraisal. Journal of Applied Psychology, 70, 56-65.

Lahey, M.A., & Saal, F.E. (1981). Evidence incompatible with a cognitive complexity theory of rating behavior. Journal of Applied Psychology, 66, 706-715.

Landy, F. J. & Farr, J. L. (1983). The measurement of work performance: Methods, theory and applications. New York, NY: Academic Press.

Landy, F.J., & Farr, J.L. (1980). Performance ratings. Psychological Bulletin, 87, 72-107.

Lee, C. (1985). Increasing performance appraisal effectiveness. Matching task types, appraisal processes, and rater training. Academy of Management Review, 10, 322-331.

Lee, C. & Bobko, P. (1994). Self-efficacy beliefs: Comparison of five measures. Journal of applied Psychology, 79, 364-369.

Locke, E.A., Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. Journal of Applied Psychology, 69, 241-251.

Mandel, M.M. (1956). Supervisory characteristics and ratings: A summary of recent research. Personnel, 32, 435-440.

Maurer, T.J. & Pierce, H.R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. Journal of Applied Psychology, 83, 324-329.

Mero, N.P., & Motowidlo, S.J. (1995). Effects of rater accountability on the accuracy and the favorably of performance ratings. Journal of Applied Psychology, 80, 517-524.

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.

Meier, R. A. & Feldhusen, J. F. (1979). Another look at Dr. Fox: Effect of stated purpose of evaluation, lecturer's expressiveness and density of lecture content on student ratings. Journal of educational Psychology, 71, 339-345.

Mischel, H.N. (1974). Sex bias in the evaluation of professional achievements. Journal of Educational Psychology, 66, 157-166.

Murphy, K. R., Balzer, W. K., Kellam, K. L., & Armstrong, J. (1984). Effects of purpose of ratings on accuracy in observing teacher behavior and evaluating teacher performance. Journal of Educational Psychology, 76, 45-54.

Murphy, K. R. & Cleveland, J. N. (1995). Understanding performance appraisal: Social, organizational and goal-based perspectives. Thousand Oaks, CA: Sage Publications.

Murphy, K.R. & Constans, J.I. (1987). Behavioral anchors as a source of bias in rating. Journal of Applied Psychology, 72, 523-579.

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.

Neck, C.P., Stewart, G.L., & Manz, C.C. (1995). Thought self-leadership as a framework for enhancing the performance of performance appraisers. Journal of Applied Behavioral Science, 31, 278-302.

Nunnally, J. (1978). Psychometric theory. New York, NY: McGraw-Hill.

Saal, F. E. & Landy, F. J. (1997). The mixed standards ratings scale: An evaluation. Organizational Behavior and Human Performance, 18, 18-35.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the quality of rating data. Psychological Bulletin, 88, 413-428.

Saal, F. E. & Knight, P. A. (1988). Industrial/organizational psychology: Science and practice. Pacific Grove, CA: Brooks/Cole Publishing Company.

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.

Schmidt, F.L., & Johnson, R.H. (1973). Effect of race on peer ratings in an industrial setting. Journal of Applied Psychology, 57, 237-241.

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean variance of performance ratings. Journal of Applied Psychology, 65, 428-435.

Schneier, C.E. (1977). Operational utility and psychometric characteristics of behavioral expectation scales. Journal of Applied Psychology, 62, 541-548.

Schwab, D.P. & Heneman (1978). Age stereotyping in performance appraisal. Journal of Applied Psychology, 63, 573-578.

Sharon, A. & Bartlett, C. (1969). Effect of instructional conditions in producing leniency on two types of ratings scales. Personnel Psychology, 22, 252-263.

Sharon, A. (1970). Eliminating bias from student rating of college instructors. Journal of Applied Psychology, 54, 278-281.

Smither, J.W. & Reilly, R.R. (1987). True intercorrelations among job components, time delays in ratings, and rater intelligence as determinants of accuracy in performance ratings. Organizational Behavior and Human Decision Processes, 40, 369-391.

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, 501-510.

Taylor, E. & Wherry, R. (1951). A study of leniency of two ratings systems. Personnel Psychology, 4, 39-47.

Villanova, P., Bernardin, H.J., Dahmus, S.A., & Sims, R.L. (1993). Rater leniency and performance appraisal discomfort. Educational and Psychological Measurement, 53, 789-799.

Waldman, D.A. & Thornton, G.C. III (1988). A field study of rating conditions and leniency in performance appraisal. Psychological Reports, 63, 835-840.

Weinberg, R.S., Yukelson, D., & Jackson, A. (1980). Effect of public and private efficacy expectations on competitive performance. Journal of Sports Psychology, 2, 340-349.

Wendelken, D.J., & Inn, A. (1981). Nonperformance influences on performance evaluations: A laboratory phenomenon? Journal of Applied Psychology, 66, 149-158.

Wexley, K.N. & Youtz, M.A. (1985). Rater beliefs about others: Their effects on rating errors and rater accuracy. Journal of Occupational Psychology, 58, 265-275.

Williams, K. J., DeNisi, A. S., Blencoe, A. G., & Cafferty, T. P. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. Organizational Behavior and Human Performance, 35, 314-339.

Wood, R.E. & Locke, E.A. (1987). The relation of self-efficacy and grade goals to academic performance. Educational and Psychological Measurement, 47, 1013-1024.

Woodruff, S. L. & Cashman, J. F. (1993). Task, domain, and general efficacy: A reexamination of the self-efficacy scale. Psychological Reports, 72, 423-432.

Yukl, G.A. (1998). Leadership in organizations. Upper Saddle River, NJ: Prentice Hall.

Zedeck, S. & Casio, W.F. (1982). Performance appraisal decision as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67, 752-758.

Footnotes

[1] To test the assumption of insufficient scale convergence, later analyses of the mediational model were run using the GSE scale, an ASE - BRSE composite, a GSE - ASE - BRSE composite, a GSE - ASE composite, and a GSE - BRSE composite. Analyses using these composites did not effect the interpretation of the overall results.

[2] Aggregated composites of both general self-efficacy - accuracy self-efficacy and general self-efficacy - behavioral recognition self-efficacy predicted leniency ($F_{(1, 107)} = 3.99$, $p < .05$; $F_{(1, 107)} = 4.83$, $p < .05$), however, they explained less variance than the task-specific composite.

[3] Leniency was significantly predicted by both accuracy self-efficacy ($F_{(1, 107)} = 5.89$, $p < .05$) and behavioral recognition self-efficacy ($F_{(1, 107)} = 4.81$, $p < .05$), however, the task-specific composite of self-efficacy explained more of the variance in leniency.

[4] Behavioral recognition accuracy was not significantly predicted by the aggregated composites of general self-efficacy - accuracy self-efficacy, general self-efficacy - behavioral recognition self-efficacy, accuracy self-efficacy, and behavioral recognition self-efficacy.

Table 1

Pilot Study 1: Coefficient Alpha and Item-Total Correlations.

Cronbach Coefficient Alpha for RAW variables: 0.548640

Deleted Variable	Correlation with Total	ΔAlpha when item is removed from scale
GSE	0.359641	0.488514
NGSE2**	0.227655	0.528988
GSE3	0.244617	0.528221
GSE4	0.414612	0.466431
GSE5	0.324543	0.491489
GSE6	0.264841	0.520152
GSE7	0.224690	0.549903

**note: question two was reversed scored because original item correlated negatively with scale total.

[1]Nomenclature refers to the General Self-Efficacy Scale and the corresponding survey item

Table 2

Pilot Study 2: Coefficient Alpha and Item-Total Correlations.

Cronbach Coefficient Alpha for RAW variables: 0.741448

Deleted Variable	Correlation with Total	ΔAlpha when item is removed from scale
GSE1	0.445440	0.722556
GSE3	0.201354	0.748163
GSE4	0.244758	0.741425
GSE5	0.420494	0.718859
GSE6	0.441589	0.715294
GSE7	0.504411	0.702302
GSE8	0.588062	0.683702
GSE9	0.501794	0.702492
GSE10	0.466358	0.715565

**note: item two was removed from scale due to low item-total correlation.

[1]Nomenclature refers to the General Self-Efficacy Scale and the corresponding survey item

Table 3

General Performance Appraisal Self-Efficacy Scale:
Coefficient Alpha and Item-Total Correlations.

Cronbach Coefficient Alpha for RAW variables: 0.801899

Deleted Variable	Correlation with Total	ΔAlpha when item is removed from scale
GSE1	0.486629	0.788117
GSE2	0.574578	0.778874
GSE3	0.526563	0.782139
GSE4	0.424166	0.789888
GSE5	0.375154	0.796972
GSE6	0.578509	0.773609
GSE7	0.352594	0.798204
GSE8	0.456654	0.787135
GSE9	0.407083	0.792239
GSE10	0.391795	0.792607
GSE11	0.544662	0.778443
GSE12	0.398800	0.793104

[1]Nomenclature refers to the General Self-Efficacy Scale and the corresponding survey item

Table 4

Mean Target Scores for Performance Dimensions.

Performance Dimension	Descriptive Statistics			
	N	Minimum	Maximum	Mean	SD
Knowledge	12	2.00	7.00	5.0000	1.2792
Delivery	12	2.00	7.00	3.4167	1.1645
Relevance	12	2.00	7.00	4.1667	1.2673
Organization	12	4.00	7.00	5.2500	1.0553

Table 5

Means and Standard Deviations for the Efficacy Scales and Dependent Variables.

Variable	N	Mean	Std Dev	Min	Max
GSE	109	45.807339	4.888530	27	55
ASE	109	23.770642	3.576241	13	30
BRSE	109	22.366972	4.104335	13	30
COMPOSITE[a]	109	46.137615	7.042816	-2.65	1.68
LENIENCY[b]	109	0.350552	1.263523	-2.54	2.96
BEHACC[c]	109	15.3211	5.313994	0	26

[a] The task-specific composite was aggregated using a standardized average of both the ASE and the BRSE.
[b] Overall measure of leniency for the entire sample.
[c] Overall measure of behavioral recognition accuracy for the entire sample.

Table 6

Correlations of Self-Efficacy Scales and Dependent Variables

	GSE	ASE	BRSE	Comp	LTOT	LADM	LDEV	BTOT	BADM	BDEV
GSE	1.00									
ASE	0.61**	1.00								
BRSE	0.44**	0.68**	1.00							
Comp	0.57**	0.92**	0.92**	1.00						
LTOT	-0.14	-0.22*	-.021*	-0.24*	1.00					
LADM	-0.21	-.025	-0.11	-0.19	1.00	1.00				
LDEV	-0.07	-0.21	-0.33*	-0.29*	1.00	…….	1.00			
BTOT	0.19*	-0.11	-0.18	-0.16	-0.02	-0.16	0.15	1.00		
BADM	0.19	-0.07	-0.16	-0.13	-0.16	-0.16	…….	1.00	1.00	
BDEV	0.19	-0.15	-0.21	-0.20	0.15	…….	0.15	1.00	…….	1.00

** p < .01
* p < .05

GSE = General Self-Efficacy Scale
ASE = Accuracy Self-Efficacy Scale
BRSE = Behavioral Recognition Self-Efficacy Scale
Comp = Task-Specific Self-Efficacy Composite
LTOT = Leniency, across sample
LADM = Leniency, within administrative purpose condition
LDEV = Leniency, within developmental purpose condition
BTOT = Behavioral Recognition Accuracy, across sample
BADM = Behavioral Recognition Accuracy, within administrative purpose condition
BDEV = Behavioral Recognition Accuracy, within developmental purpose condition

Table 7

Regression Table for Tests of Mediation

Model	F Value	Pr > F	R-Square	Parameter Estimate	Pr > |T|
Step 1:					
Leniency = Purpose	0.25	0.62	0.0023	-.123	0.62
Step 2:					
GSE = Purpose	0.43	0.51	0.004	-.6166	0.51
Composite = Purpose	0.02	0.89	0.0002	-.1944	0.89
Step 3:					
Leniency = GSE	2.29	0.13	0.0209	-.037	0.13
Leniency = Composite	6.36*	0.013	0.0561	-.04225*	0.013

* $p < .05$

Table 8

Means and Standard Deviations for Mediation Model Variables

Variable	Purpose of Appraisal	Mean	Standard Deviation
Leniency	Administrative	0.409	1.316
	Developmental	0.289	1.215
GSE	Administrative	46.11	5.259
	Developmental	45.49	4.492
Composite	Developmental	46.03	7.338

Table 9

Regression Table for Tests of Behavioral Accuracy Hypotheses

Model	F Value	Pr > F	R-Square	Parameter Estimate	Pr > |T|
Hypothesis 2:					
Leniency = BA	0.07	0.798	0.00062	-.0059	0.798
Hypothesis 3:					
BA = GSE	3.99*	0.048	0.0359	.206*	0.048
BA = Composite	2.89	0.094	0.026	-.122	0.094
T-Hit Rate = GSE	3.99*	0.048	0.0359	.103*	0.048
T-H Rate = Composite	2.89	0.094	0.026	-.061	0.094
False-Pos. = GSE	3.99*	0.048	0.0359	-.103*	0.048
False-Pos. = Composite	2.89	0.094	0.026	.061	0.094

* $p < .05$

Appendix A1
General Performance Appraisal Self-Efficacy Scale

Please rate your level of agreement with the following statements, as you perceive them as of now.

I am an accurate judge of the performance of others.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can review a person's performance and give an honest opinion.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can tell the difference between good performance and poor performance.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can remember a person's behaviors with only limited exposure to them.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can evaluate performance based on standards given to me by someone else.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can evaluate people without letting my personal feelings interfere with my conclusions.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can accurately evaluate the performance of an instructor based on a limited sample of his or her teaching behaviors.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can accurately evaluate the performance of someone for whom I personally dislike.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can judge an instructor's performance in a subject that I have limited knowledge in.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can render an objective and accurate judgement concerning a person's performance even if the judgement may result in a negative outcome for the person.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can accurately evaluate the performance of an instructor in a course that I am not personally interested in.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I can render an objective and accurate judgement concerning a person's performance even if the judgement may result in a negative outcome for me.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

Appendix A2
Accuracy Self-Efficacy Scale

I am confident that I can evaluate an instructor's performance with **at least** 100% accuracy.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I am confident that I can evaluate an instructor's performance with **at least** 90% accuracy.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I am confident that I can evaluate an instructor's performance with **at least** 80% accuracy.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I am confident that I can evaluate an instructor's performance with **at least** 70% accuracy.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I am confident that I can evaluate an instructor's performance with **at least** 60% accuracy.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

I am confident that I can evaluate an instructor's performance with **at least** 50% accuracy.

1----------2----------3----------4----------5
strongly disagree disagree undecided agree strongly agree

Appendix A3
Behavioral Recognition Self-Efficacy Scale

I am confident that I can accurately recall **at least** 100% of an instructor's behaviors.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

I am confident that I can accurately recall **at least** 90% of an instructor's behaviors.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

I am confident that I can accurately recall **at least** 80% of an instructor's behaviors.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

I am confident that I can accurately recall **at least** 70% of an instructor's behaviors.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

I am confident that I can accurately recall **at least** 60% of an instructor's behaviors.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

I am confident that I can accurately recall **at least** 50% of an instructor's behaviors.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

Appendix B1
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 topic 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 and Demand.
6. Lecturer presents multiple citations concerning the effect of price on purchasing behavior.
7. Lecturer presents results of personnel 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

Appendix B2

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 the lectern or at the blackboard; refrains from pacing.
14. Lecturer writes legibly when using the blackboard to report results from his single cue study.
15. Midway through the lecture, the lecturer begins pacing.
16. Lecturer does not label or explain the 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 of she uses the board to clarify and emphasize important point 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 D1

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 lecture 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 in a 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 to student (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 TV 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.

Appendix D2

Y N 17. Instructor illustrated how price affects perceptions of quality using various product (e.g., motor oil, shaving cream, razor blades) as examples.

Y N 18. Instructor had handouts available for certain lecture topics.

Y N 19. Instructor had handouts available outlining the material covered in lecture.

Y N 20. Instructor mentioned many different articles concerning the effect of behavior.

Y N 21. To show the effects of brand names on perception of quality, the instructor used designer jeans as an example.

Y N 22. Instructor used the chalkboard appropriately.

Y N 23. Instructor discussed each topic in the same order as was presented in the lecture outline.

Appendix D2

Y N 24. Instructor was involved in a research project with an important figure in the area of consumer psychology.

Y N 25. Instructor used an example of how many franc people in France were willing to spend on a diamond.

Y N 26. Instructor had a distracting habit of removing his glasses and pinching the bridge of his nose.

Y N 27. Prior to the lecture, the instructor had set-up all necessary audio-visual equipment for presenting the lecture.

Y N 28. Instructor was familiar with a body of research because of his intention to do future research in the area.

Y N 29. Instructor used unrealistic price levels when presenting examples of consumer purchasing decisions (i.e., buying a stereo for $50.00).

Y N 30. Instructor did not label or explain his graph of the demand curve.

Y N 31. Instructor put an outline of the day's lecture on the board.

Y N 32. Instructor presented multiple examples of research studies illustrating the Theory of Absolute Price Threshold.

*Correct responses are underlined

Appendix E

Manipulation Check Items

Please rate your level of agreement with the following statements.

The purpose of my ratings was to provide performance information that will be used to make decisions concerning future teaching assignments and funding increases.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

The purpose of my ratings was to provide performance information that will be used to provide developmental feedback.

1----------2----------3----------4----------5

strongly disagree disagree undecided agree strongly agree

VITA

Adam N. Prowker

602 McBryde Drive • Blacksburg, VA 24060 • (540) 953-0529 • aprowker@vt.edu

EDUCATION

M.S. Industrial/Organizational Psychology, May 1999
Virginia Polytechnic Institute and State University, Blacksburg, VA

B.A. Psychology, Magna cum Laude with Honors in Psychology, 1997
Moravian College, Bethlehem, PA

RESEARCH EXPERIENCE

Virginia Polytechnic Institute and State University, May 98 to May 99

Examined the effects of purpose of appraisal and rater self-efficacy on rating leniency and behavioral recognition accuracy. Designed and validated a general rating self-efficacy scale and two task-specific rating self-efficacy scales. Used alpha reliability and confirmatory/exploratory factor analytic methods to assess scale integrity. Used multiple regression to analyze data.

Virginia Polytechnic Institute and State University, Sept 97 to May 98

Worked on a research team exploring current organizational research philosophies and various methods of theory development and testing. Conducted extensive reviews of literature concerning logical positivism, the hypothetico-deductive model, and falsificationism. Presented new perspectives on and suggestions for theory development, research methodology, statistical analysis of theory.

Moravian College, May 96 to May 97

Designed research methodology to test perceptions of compensation equity in work and academic settings. Tested various research models based on equity theory, cognitive dissonance, and equity sensitivity theory. Used various parametric and nonparametric statistical methods to test hypotheses and draw conclusions.

Moravian College, May 96 to May 97

Assisted in the construction of a survey instrument used to measure supervisor information processing in association with performance appraisal in an organizational setting. Coded qualitative response data and performed various statistical analyses of survey data.

TEACHING EXPERIENCE

Virginia Polytechnic Institute and State University, Jan 99 to May 99

Teaching Assistant: Physiological Psychology & Advanced Social Psychology Advised and tutored students on course related topics. Assisted in test construction, administration and evaluation. Lectured on various occasions when necessary.

Virginia Polytechnic Institute and State University, Sept 98 to Dec 98

Teaching Assistant: Personality Psychology & Industrial/Organizational Psychology Advised and tutored students on course related topics. Assisted in test construction, administration and evaluation. Lectured on various occasions when necessary.

Virginia Polytechnic Institute and State University, Sept 97 to May 98

Teaching Assistant: Introductory Psychology Recitation. Full responsibility for teaching course.

PRESENTATIONS

Brill, R.T., & Prowker, A.N. (1996). Supervisor information processing: Rating complexity and confidence. Paper presented at the annual conference of the American Psychological Society, San Francisco, CA.

Prowker, A.N., Heverly, A., & Shelton, E. (1995). The effects of task significance on productivity and work related attitudes. Presented at the 11th Annual Lehigh Valley Undergraduate Psychology Conference.

PROFESSIONAL AFFILIATIONS

Society for Industrial and Organizational Psychology
American Psychological Association