THE UNIVERSITY OF CALGARY

# The Effects of Task Information Training and Frame-of-Reference Training with Situational Constraints on Rating Accuracy

by

Janine Keown

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF PSYCHOLOGY

CALGARY, ALBERTA

SEPTEMBER, 1997

0-612-24673-6

# ABSTRACT

Undergraduate students (N=96) were trained to complete performance evaluations using either 1) traditional frame-of-reference training on the performance dimensions (FORD); 2) FORD and FOR training on the situational constraints encountered by the ratee (FORD + FORS); 3) FORD and FORS plus training on the weighting strategy used to combine dimension and situational cues in determining deserved ratings (FORTI); or 4) control procedures. Participants then read twenty profiles which described lecturers' performance and rated each lecturer on their Observed Performance, Situational Constraints, and Deserved Performance on three dimensions relating to lecturing. Results suggest that FORD and FORS training increase rating accuracy for Observed Performance and Situational Constraint ratings, respectively. Furthermore, participants provided with a weighting policy successfully adopted this policy in determining Deserved Performance ratings.

# ACKNOWLEDGEMENTS

I would like to thank Dr. Lorne Sulsky, my thesis advisor, for the trust, respect, and guidance he has provided me throughout the thesis process. He has been one of my biggest supporters. His confidence in me has been fundamental in the confidence I have established in myself in my academic career.

I would like to thank Sterling for his continued support. He has been a constant reminder that there is a whole other world outside of graduate school. I would like to recognize my parents, family, and friends for their continued belief that I can accomplish anything I set my mind to.

I would like to thank Theresa and Dan for their support in the past two years. Finally, I would like to thank my fellow graduate students in the I/O/E area group. The hours of mindless chatter, deep discussion, laughter, antics, and sometimes tears we have shared have maintained my sanity throughout this process.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# THE EFFECTS OF TASK INFORMATION TRAINING AND FRAME-OF-REFERENCE TRAINING WITH SITUATIONAL CONSTRAINTS ON RATING ACCURACY

Training raters to improve the accuracy of their performance evaluations has been the focus of numerous studies in the performance appraisal literature. Frame-of-reference (FOR) training is one such rater training program that has been demonstrated to improve the accuracy of performance ratings (Athey & McIntyre, 1987; Bernardin & Pence, 1980; Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr, 1994). Although it has been recognized that situational constraints encountered by employees can have direct effects on their performance (O'Connor, et al., 1984; Peters, O'Connor, & Rudolf, 1980; Steel & Mento, 1986; Steel, Mento & Hendrix, 1987), previous FOR training studies have largely been conducted in laboratory settings in which situational constraints have been controlled. The present study examines how FOR training may be expanded to aid in handling situational constraint information and whether ratees can be trained to correctly utilize situational constraint information in determining performance ratings.

## Situational Constraints

Situational constraints are defined by Peters and O'Connor (1980) as "aspects of the immediate work situation...that interfere with the translation of abilities and motivation into effective performance" (p.391). This definition suggests that situational constraints are aspects of the work situation and not the individual employee and, furthermore, that these situational variables do have an effect on work outcomes.

Peters et al. (1980) attempted to identify situational variables by developing a taxonomy of situational characteristics which impact work outcomes. They asked a number of full-time employees to identify instances in their job in which they performed poorly and then to identify situational conditions which they believe may have affected their performance. Peters et al. used a sorting methodology to categorize these situational conditions into a number of situational resource variables. These variables as listed in Table 1 represent a number of situational resources needed by the employee to successfully perform their job. It is suggested that these resources may differ in their quantity, quality, or accessibility amongst employees. In other words, a given employee may have lower performance because they did not have enough of a given resource, a needed resource was of poor quality, or a needed resource was inaccessible. Peters et al. did identify, however, that the situational resources identified were for a number of different jobs in general, and the particular configuration of situational variables relevant to any particular job would be job specific.

Table 1

Situational Resource Variables Relevant to Performance

1. *Job-related information.* Refers to the information (from supervisors, peers, subordinates, customers, company rules, policies, and procedures, and so forth) needed to do the job assigned.
2. *Tools and equipment.* Refers to those specific tools, equipment, and machinery needed to do the job assigned.
3. *Materials and supplies.* Refers to those materials and supplies needed to do the job assigned.
4. *Budgetary support.* Refers to the financial resources and budgetary support needed to do the job assigned- the monetary resources needed to accomplish aspects of the job to include such things as long distance calls, travel, job-related entertainment, hiring new and maintaining/retaining existing personnel, hiring emergency help, and so forth. This category does not refer to an incumbent's own salary, but rather, to the monetary support necessary to accomplish tasks which are part of the job they have been assigned.
5. *Required services and help from others.* Refers to the services and help from others needed to do the job assigned.
6. *Task preparation.* Refers to the previous personnel preparation, through previous education, formal company training, and relevant job experience, needed to do the job assigned.

7. *Time availability.* Refers to the availability of the time taking into consideration both the time limits imposed and the interruptions, unnecessary meetings, non-job related distractions, and so forth, needed to do the job assigned.
8. *Work environment.* Refers to the physical aspects of the immediate work environment which are needed to do the job assigned- characteristics which facilitate, rather than interfere with doing the job assigned. For example, a helpful work environment is one that is not too noisy, too cold, or too hot; that provides an appropriate work area; that is well lighted; that is safe; and so forth.

Note. From "The Behavioral and Affective Consequences of Performance-Relevant Situational Variables" by L.H. Peters, E.J. O'Connor, and C. J. Rudolf, 1980, Organizational Behavior and Human Performance, 25, p.82. Copyright 1980 by Academic Press, Inc. Reprinted with permission.

Models of work performance have hypothesized that situational constraints affect variance in performance (Landy & Farr, 1980; Schneider, 1978; Terborg, 1977). Peters and O'Connor (1980) suggest the influence of situational constraints on performance is twofold; 1) situational constraints directly influence performance by hindering the utilization of ability, and 2) situational constraints have indirect effects in that the restrictive conditions created by situational constraints result in frustration and dissatisfaction amongst employees which result in decreased motivation and decreased performance.

Research has demonstrated that situational constraints are related to subjective measures of performance such as supervisor ratings (O'Connor et al., 1984; Peters, Fisher, & O'Connor, 1982; Peters et al., 1980; Steel & Mento, 1986; Steel et al., 1987), as well as to objective measures of performance such as cash shortages (Steel et al., 1987). It has also been empirically demonstrated that situational constraints are related to increased levels of frustration and decreased levels of satisfaction amongst employees (O'Connor et al., 1984; Peters et al., 1980).

Performance appraisal serves as a system for the supervisor to evaluate performance, to aid in personnel decisions, and to give employees feedback on their work performance. The focus of this evaluation is assumed to be factors internal to the employee such as their ability and effort. It is assumed that ratings do not include variance in performance not under the individual employee's control, such as variation due to factors of the situation. Recently researchers have expressed a concern that raters tend to focus ratings on performance outcomes without making any attributions to the cause of that outcome (Carson, Cardy, & Dobbins, 1991; Deming, 1986). According to Kelley (1973), one of the primary functions of a supervisor is to determine whether the performance of an employee is in fact caused by the employee, the task itself, the environment, or some combination of these.

When carrying out performance evaluations, managers may have difficulty recognizing and separating situational variance from variance in performance attributable to the individual. Deming (1986) suggested that performance appraisal is actually damaging to an organization due to its tendency to incorrectly attribute variance in performance to the employees rather than to problems within the system. Deming suggested the majority of variation in performance is due to situational or system factors, however, supervisors utilizing performance appraisal assume variation in performance to be caused mainly by factors internal to the employee. This incorrect attribution may cause managers to focus interventions on the employees rather than to the system and can cause morale problems among employees.

Carson et al. (1991) illustrated that subordinates and supervisors differ in their perceptions of the causes of variation in performance. They asked both supervisors and subordinates to estimate the percentage of variance in performance produced by situational factors and the proportion of variance attributable to subordinate characteristics. They found supervisors attribute 73.58% of variability in performance to subordinate characteristics while subordinates believe that only 44.78% of variability is due to this. It was also found that supervisors were more likely than subordinates to rate employee factors, such as low motivation and low ability, as significant contributors to poor performance and less likely to rate situational factors, such as inadequate tools or equipment and poor scheduling. In the absence of information concerning the effects of situational constraints in the environment, it is difficult to determine whether the supervisors' or the subordinates' perceptions are correct. The results do suggest, however, that there is variance amongst individuals in their understanding of the effects of situational constraints on performance. The results are consistent with the Fundamental Attribution Error which suggests that there is a tendency amongst observers to underestimate the extent to which the behavior of others is affected by situational sources (Ross, 1977).

Carson et al. (1991) also performed a policy capturing study to examine how raters weight information about situational constraints, effort, ability, and productivity in their performance judgments. They discovered participants based their performance evaluations almost entirely upon the employee's productivity and tended to overlook all other information. This suggests that raters tend to base performance evaluations on the outcome of the employee and do not attempt to make any attributions regarding whether that outcome was under the individual's control. The design of the study, however, manipulated the cue values so that all cues were orthogonal to one another. For example, the level of situational constraints experienced by the employee did not covary with the employee's productivity. Attribution theory suggests individuals examine the relationship between variables to guide their attributional judgments (Kelley, 1973). In other words, subjects may have recognized that situational constraints had no correlation with productivity and used this information to conclude that situational constraints had no effect on productivity. Therefore, this was an indication that situational constraints need not be considered in the resulting performance evaluations. In the environment, if situational constraints do have a consistent influence on the employee's performance, one would expect to find a negative correlation between the level of situational constraints experienced and an employee's productivity. In this instance, situational information may be weighted and integrated into a final performance judgment.

In the present study, I examine whether raters use information about the covariance between situational constraints and observed performance in determining the appropriate manner in which to weight situational variables in the performance evaluation. I also examine whether raters can be trained to weight observed performance and situational constraints according to an established theory of performance.

## Overview of FOR Training

There are numerous rater training programs that have been developed to increase the accuracy of performance ratings (Woehr & Huffcutt, 1994). To date, these programs have failed to examine contextual factors that may influence performance. FOR training, in particular, typically focuses on calibrating raters so they agree on the dimensions on which performance is judged and what constitutes levels of performance for each of these dimensions. FOR training involves emphasizing the multidimensionality of work performance, defining the performance dimensions, defining and describing behavioral examples of performance levels for each dimension, and practice and feedback in using these standards to evaluate performance (Bernardin & Buckley, 1981). This enables the rater to correctly categorize behaviors to the appropriate performance dimension and level of performance (Hauenstein & Foti, 1989). Overall, research has consistently shown that FOR training is effective in increasing the accuracy of performance ratings (Athey & McIntyre, 1987; Bernardin & Pence, 1980; Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; McIntyre et al, 1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr, 1994).

Recent research has investigated the cognitive mechanisms underlying FOR training to help explain the effectiveness of FOR training in increasing accuracy. Athey and McIntyre (1987) found that FOR-trained raters tend to remember more of the training content than raters trained with control training. Levels-of-processing theory was used to explain this finding. Levels-of-processing theory suggests that information that undergoes greater cognitive elaboration or is processed at a greater "depth" will be better remembered (Craik & Lockhart, 1972). Athey and McIntyre suggested that because raters engage in greater cognitive elaboration or deeper processing when given FOR training versus control training, the training content is better remembered and this leads to greater accuracy. This finding, however, does not address what specifically about the FOR training content leads to increased accuracy.

Sulsky and Day (1992) suggest that the specific theory of performance upon which FOR training is based allows the rater to develop precise prototypes of effective and ineffective performance. These prototypes aid the rater in correctly categorizing the behavior observed into the appropriate performance level of the appropriate dimension. The rater subsequently bases their performance judgments on these categorizations rather than on memory for specific behavioral information. In fact, Sulsky and Day found that FOR-trained raters did exhibit superior rating accuracy but tended to forget specific behavioral instances. Furthermore, FOR-trained raters tended to falsely recognize behaviors which were consistent to the ratee classification but not present for a given ratee.

FOR training research is typically conducted in laboratory settings in which ratee performance is illustrated with the use of videotape vignettes. This allows the researcher to manipulate performance so variables outside of the employee have no effect upon the ratee's performance. Therefore, situational variables which may influence performance and performance ratings are held constant for the purposes of studying FOR training. In other words, the theory of performance upon which FOR training is based is generally one which does not include situational constraint variables.

The present study examines the effectiveness of FOR training in situations in which performance is influenced by both ratee ability and effort as well as by situational constraints. Additionally, the FOR training methodology was expanded in certain experimental conditions for the purposes of rating situational constraints. For example, when raters are assessing the situational constraints encountered by an employee, it is vital they are operating from a similar frame-of-reference. Raters would thus benefit from a common theory of performance which dictates what constitutes severe, moderate, or limited situational constraints. For instance, if the situational resource in question is the tools and equipment available to the employee, the situational theory would specify what quantity, quality, and availability of tools and equipment would constitute a "1" (i.e., limited situational constraints) versus a "7" (i.e., severe situational constraints) on a 7-point Likert-type scale.

## Policy Capturing

By utilizing a policy capturing methodology, the present study attempts to determine if raters can be trained to integrate information about situational constraints in their performance ratings. Policy capturing is a strategy used to explain the processes decision makers use to weight and combine different pieces of information. Policy capturing requires an individual to make a series of judgments on profiles containing cues relevant to the situation. The individual makes judgments on multiple profiles and the levels of cues vary across these profiles. Multiple regression is then utilized to determine the relative predictability of the cues in the participant's overall judgment strategy. In other words, policy capturing allows the researcher to manipulate several cues concurrently to determine an individual's cue weighting system or judgment policy. Policy capturing also allows the researcher to examine whether groups of individuals consistently use the same cues when making decisions.

By way of illustration, a recent study by Kline and Sulsky (1995) examined the decision-making policies of professors in regards to the probability that they would accept applicants into graduate school. The cues or information upon which the judges based their decision were GRE scores, GPA, grade in statistics, etc. Participants were given a number of profiles in which the values of these cues varied and were asked to make a judgment of the probability in which they would accept the applicant in question. Kline and Sulsky were then able to regress the judgments onto these cue values to determine the relative predictability of each cue in determining the participant's responses. This methodology also enabled the researchers to examine whether individual variables such as the professor's age or gender helped explain differences between subjects in their rating policies.

The policy capturing methodology is well suited to the study of performance appraisal decision processes because it provides a means to empirically examine the complex judgment processes involved in making performance ratings. In performance appraisal research, policy capturing has been utilized to examine how raters combine and weight dimensional information to develop an overall rating for the employee (Hobson, Mendel, & Gibson, 1981). It can also provide an algorithm which objectively describes the process in which a rater weights and integrates performance and contextual information in arriving at performance ratings (Seitz, 1988).

Previous performance appraisal research utilizing policy capturing has found that the additive component of the general linear model is adequate in describing rater policies. In other words, raters generally do not tend to use non-linear or non-compensatory approaches in performance rating (Hobson et al., 1981; Zedeck & Cascio, 1982). As well, research has shown that raters lack insight into their own rating policies. Specifically, raters' subjective distributions of weights to performance dimensions are often more equally distributed than is observed in the multiple regression equation (Hobson et al., 1981; Taylor & Wilstead, 1974). That is, raters believe that they tend to equally weight performance dimensions in arriving at an overall rating; however, in reality raters tend to rely more heavily on some dimensions and less so on others when arriving at their rating. Finally, policy capturing research has shown distinct subgroups can be identified which possess different rating orientations. This implies it is impossible to use a single rating strategy to describe all raters (Hobson et al., 1981; Stumpf & London, 1981).

As suggested earlier, often there is variability amongst raters in their rating strategies. Therefore, it may be helpful to examine the training or information exchange necessary to unite raters in a similar strategy. The ultimate desired strategy is often based on knowledge of how the cues predict the criterion in the environment. In the present study, the desired strategy is developed by utilizing a theory of performance which specifies how the cues should be weighted in arriving at a performance rating.

The process of providing raters with information about the relationship between the rater's judgments and the task is called Cognitive Feedback (CFB). Doherty and Balzer (1988) have identified three components of CFB: task information (TI), cognitive information (CI), and functional validity information

(FVI). TI is information regarding how the cues predict the criterion in the environment, or for the purposes of this study, information regarding the desired weightings of the cues as suggested by the theory of performance. CI is information about how the rater has weighted the cues in arriving at the ratings. Finally, FVI is information regarding how the rater's strategy is related to the desired strategy. It is suggested it is the TI component of CFB that leads to increased validity and accuracy of the judgment (Balzer, Doherty, & O'Connor, 1989; Balzer, Sulsky, Hammer, & Sumner, 1992; Doherty & Balzer, 1988). In other words, informing raters of the desired policy is sufficient information for them to adopt this policy.

## The Present Study

Undergraduate participants were given one of four different types of training (see below) and then evaluated target ratees. The target ratees were presented in written profiles describing ratees' performance on three different performance dimensions and the situational constraints encountered by ratees on the job. Appendix A includes one of these profiles and the rating scales. Each profile describes how the employee was observed to perform on each of the three dimensions and information about the situational constraints faced by the employee that may have affected observed performance. For each dimension, the participant was required to rate the observed performance of the ratee and the severity of the situational constraints faced by that ratee. The participant was then asked to determine what they believed the deserved rating of the ratee was by taking into account the previous two pieces of information. In other words, participants were required to determine values for the cues by rating observed performance and situational constraints and then use these cue values in order to make a judgment of the deserved performance rating of the employee.

Specifically, participants were assigned to one of four groups: 1) FOR training on the performance dimensions (FORD); 2) FORD + FOR training on situational constraints (FORS); 3) FORD + FORS + TI training (this group will subsequently be referred to as the FORTI group); or 4) Control training. FORD training is based on traditional FOR training and focused on training raters to correctly classify observed performance into the appropriate level of the appropriate dimension. FORS training is an extension of FOR training which focused on training raters to correctly classify the severity of the situational constraints encountered by the ratee. Finally, TI training is a training program which focused on training raters to properly weight and integrate the cue values to determine the deserved performance rating. These training programs were evaluated by examining their effects on the accuracy of the observed performance, situational constraint, and deserved performance ratings. A completely crossed design was not used because it was not expected that there would be any interactions between the different types of training. Furthermore, the specified groups allowed for the evaluation of all types of training using a reasonable amount of participants.

It is believed that if the training programs are effective, those groups given training on the applicable frame-of-reference for observed performance and situational constraints will be more accurate on their ratings of observed performance and situational constraints, respectively, than those groups that did not receive training.

H1: The FORD, FORD + FORS, and FORTI groups will be more accurate on their ratings of observed performance than the Control group.

H2: The FORD + FORS and FORTI groups will be more accurate on their ratings of the situation than the FORD and Control groups.

When raters are given instruction on how to weight the cues in determining the deserved performance ratings, it suggests that they will consistently use the algorithm provided to them. However, those not given instruction in how to weight the cues may develop their own strategy. Given that the comparison scores for the deserved ratings are based upon a set algorithm, adoption of the prescribed weighting algorithm will lead to greater accuracy of the deserved performance ratings.

H3: The FORTI group will be more accurate than the other groups on their deserved performance ratings.

In addition, to address concerns arising from Cardy et al.'s (1991) study I was interested in examining the effects of the correlation of observed performance and situational constraints on the participants' weighting strategy. It is expected when situational constraints are highly correlated to observed performance, participants will tend to weight the situation more highly in their deserved performance rating than when situational constraints and observed performance are less correlated. Furthermore, I will only examine this effect within the control, FORD, and FORD + FORS groups. Participants in the FORTI group will not be included in this analysis because they were given instruction in the desired weighting policy. Therefore, as suggested by their hypothesized increased deserved performance ratings, they do not have the freedom to adopt their own weighting strategy but instead utilize the provided weighting strategy.

H4: Participants in the Control, FORD, and FORD + FORS groups will tend to give more weight/importance to the situational constraint cue as the correlation between the cues increases.

## Method

### Participants

Participants were 96 volunteer undergraduate students from The University of Calgary. Thirty-five percent of the participants were male and 65% were female. Participants ranged in age from 18 to 27 years old with a mean age of 19.6. Approximately 12.5% of the participants had worked previously in jobs which required them to formally evaluate other employees. Participants were randomly assigned to one of the four training conditions (n=24 for each condition).

Development of Stimulus Materials

To decide on the situational constraints relevant to each performance dimension, a series of focus groups and questionnaires were utilized. The first focus group consisted of five professors and graduate students with experience in lecturing. The performance dimensions were adopted from research conducted on lecturing performance by McCauley et al. (1990). Each performance dimension was defined for the group. Members were then asked to recall and share critical incidents related to that dimension (i.e., times in which they performed poorly on the particular dimension). Members were then asked to identify specific situational conditions that they believe negatively affected their performance in the given incidences. The findings of the focus group were used to establish a specific situational constraint to be evaluated for each dimension. The situational constraints chosen were based on general agreement regarding the relevance of the situational constraint amongst members of the focus group as well as the fit of the particular situational constraints with Peters and O'Connor's (1980) taxonomy of situational constraints.

A second focus group was conducted to determine how to scale the situational constraint measure for each dimension. The second focus group consisted of three graduate students with expertise in lecturing. Each situational constraint was defined for the group and the group then discussed what factors would suggest situational constraints for each level of the 7-point Likert-type scale. Any discrepancies between group members were discussed until members could come to a consensus on how to scale the situational constraint factors.

Finally, a brief questionnaire was developed that outlined the situational constraints chosen and the given levels of each. This questionnaire was distributed to ten professors, five of which responded. The questionnaire briefly explained the rationale of the study and listed the situational constraints and the different levels of constraints as developed in the previous groups. The different levels of constraints were randomly listed and respondents were asked to rate each on a 7-point Likert-type scale with "1" indicating "no constraint" and "7" indicating "severe constraint". Respondents were also asked to list any further situational constraints that they believed might also impact on each performance dimension.

A final focus group was conducted to determine the theory of performance of the performance dimensions and the weighting policy to be used. This group consisted of a graduate student and a professor. As mentioned previously, the dimensions used were developed based on previous research by McCauley et al. (1990). Three of the dimensions developed by McCauley et al. were adopted and the group discussed the definition of the dimension and what constitutes performance for each level of the 7-point Likert-type scale for each of the three dimensions. Finally, they discussed the policy for which observed performance and situational constraints would be weighted in determining deserved performance for each dimension. A final policy was decided upon based on group consensus. The final policy compensated individuals who encountered situational constraints by adding to the established observed performance rating to determine the deserved performance rating. For example, for ratees who did not have situational constraints (those who scored a "1" on the situational constraint scale), the rater gave them the same rating for deserved performance as for observed performance. For those that scored a 2 or 3 on the situational constraint scale, raters calculated the deserved performance score by taking the observed performance score and adding one. When ratees scored a 4 or 5 on situational constraints, deserved performance was calculated by adding two to the observed performance rating. Finally, when ratees scored a 6 or 7 on the situational constraint scale, raters added 3 to the observed performance rating to determine the deserved performance rating. If the algorithm ever provided the rater with a score greater than 7 for deserved performance, the rater would assign the ratee a 7 on the deserved performance scale.

## Stimulus Materials

The stimulus set consisted of twenty written profiles of twenty ratees. Each profile describes a university lecturer with performance information on each of the three dimensions as well as the levels of situational constraints encountered. The three dimensions evaluated were "Speaking Ability", "Organization", and "Fielding Student Questions". The situational constraints assessed were "Previous Lecturing Experience" (for "Speaking Ability"), "Quality and Availability of Equipment" (for "Organization"), and "Class Size" for ("Fielding Student Questions"). For each subject, one dimension in each of the profiles had a large ($r = -0.6$) correlation between the situational constraint and observed performance cues, one dimension had a moderate ($r = -0.3$) correlation, and for the third dimension the cues were not correlated ($r = 0$). These cue intercorrelations were counterbalanced across dimensions so that for each participant, the dimension/cue correlation pairing was randomly determined.

## Rater Training

Training was conducted in small groups (a mean of 2 people per group). Participants were asked to assume the role of a student responsible for evaluating a number of university lecturers. Each session including the training and completion of the tasks lasted approximately 90 minutes in duration.

FORD training. Participants were informed that they are to evaluate ratee performance on three separate performance dimensions, three situational constraints encountered by the ratee, and the deserved performance of the ratee. Following procedures employed in previous FOR studies (e.g. Sulsky & Day, 1992, 1994), participants were presented with the performance rating scales for observed performance and the trainer read aloud and defined each dimension. The trainer presented and discussed examples of behavior representing different levels of performance for each of the dimensions. The participants were asked to read practice vignettes of three ratees and provide ratings on observed performance, situational constraints, and deserved performance. The ratings of observed performance for each of the three dimensions were discussed and the trainer provided feedback indicating the performance level on each dimension that was appropriate for each ratee. The ratings of situational constraints and deserved performance were not discussed and no feedback was provided for these ratings. The ratees were also given a short lecture on performance appraisal to equate the training time to that of the FORTI training group.

FORD + FORS training. Participants received FORD training as discussed earlier. Additionally, the trainer read the situational constraint scales aloud and gave examples of the situational constraints that would constitute each level of the scales. The participants were asked to read practice vignettes of three ratees and provide ratings on the observed performance, situational constraints, and deserved performance. The ratings for observed performance and situational constraints were discussed and the trainer provided feedback indicating the level of each scale that was appropriate for each ratee. The ratings for deserved performance were not discussed nor was feedback provided. The ratees also received a short lecture on performance appraisal to equate the training time to that of the FORTI training group.

FORTI training. Participants received FORD and FORS training as discussed earlier. Furthermore, the trainer gave a short lecture emphasizing the importance of considering both performance and situational factors when evaluating performance. For each of the three dimensions, the trainer discussed how the situational constraints and observed performance should be weighted to generate the deserved performance rating. Participants were then provided with the profiles of three practice ratees and asked to provide an observed performance rating, situational constraint rating, and deserved performance rating. These ratings were discussed and the trainer provided feedback on observed performance, situational constraint, and deserved ratings.

Control training. The trainer gave a lecture describing the purpose of performance appraisal in general terms. To control for Hawthorne effects, participants were given an interactive exercise about performance appraisal to complete and discuss. This exercise was designed so that all groups received interaction with the trainer but did not include any information about the theory of performance or situational constraint effects on performance. The trainer distributed the rating scales to the participants and read the scales aloud with them, however, no specific training occurred regarding the rating scales. Participants read three practice vignettes depicting performance and situational constraints and were asked to rate the employees on the observed performance, situational constraint, and deserved performance scales. The ratings were not discussed and no feedback on the accuracy of the ratings was given.

## Procedure

Participants were informed that the purpose of the study was to examine how people evaluate performance. Participants then received training according to their experimental condition. Immediately following training, the participants were given the target vignettes. Participants were informed that they were to evaluate the observed performance, situational constraints, and deserved performance for a number of individuals. Observed performance was defined to the raters as "the performance of the employee that you observe, not accounting for any of the causes for this performance". Situational constraints were defined as "situational conditions not under the control of the individual which may affect performance outcomes". Deserved performance was defined as "the rating which you believe the individual deserves". Participants were instructed to evaluate the ratees as accurately as possible. Participants were given twenty profiles to rate which contained information on observed performance and situational constraints for each of the three dimensions. Immediately following each ratee the participants were asked to rate the observed performance of the ratee on the appropriate dimension (OP), the level of situational constraints encountered by the ratee (SC), and the deserved rating of the ratee (DR). OP and DR were rated on a 7-point Likert-type scale with 1 indicating extremely poor performance and 7 indicating extremely good performance. SC was rated on a 7-point Likert-type scale with 1 indicating no situational constraints and 7 indicating extreme situational constraints.

### Comparison Scores and Accuracy Measures

A panel of experts that consisted of a professor and a graduate student with experience in performance appraisal and lecturing determined comparison scores. This panel of experts was given a verbal and written description explaining the theory of performance for OP, SC, and the weighting policy. The experts read each of the vignettes and provided a rating on the observed performance on each of the performance dimensions (OP) and the situational constraints experienced (SC). Multiple opportunities were given to read the profiles and adjust their ratings if necessary. Any discrepancies on scores on the situational constraint cues or the observed performance ratings were discussed amongst the experts until resolved (cf. Sulsky & Balzer, 1988). Comparison scores for DR were obtained by entering the comparison scores on OP and SC into the comparison score weighting policy.

## Dependent Variables

Several performance measures were calculated based on the participants' ratings:

Observed performance accuracy. Cronbach's (1955) accuracy components (i.e., elevation (E), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA)) calculated using OP ratings and OP comparison scores were employed to assess OP rating accuracy. See Sulsky and Balzer (1988) for more information regarding these accuracy components.

Situational constraint accuracy. Cronbach's (1955) accuracy components calculated using SC ratings and the SC comparison scores were employed to assess SC rating accuracy.

Deserved performance accuracy. Cronbach's (1955) accuracy components calculated using DR ratings and DR comparison scores were employed to assess DR rating accuracy.

Usefulness index (UI). Usefulness indices were computed for each cue for each of the dimensions to assess the incremental variance in DR attributable to a particular cue. This index is an indicator of the use and relative importance of a given cue in a participant's weighting strategy (Darlington, 1968). The UI value represents the unique proportion of variance accounted for by each cue.

Therefore, UI equals the change in $R^2$ observed when the cue of interest is dropped from the regression equation.

## Results

### Observed Performance Accuracy

To determine if the groups receiving FORD training were more accurate than the group that did not (Hypothesis 1), a multivariate analysis of variance was conducted for the training conditions using Cronbach's accuracy components of OP ratings as the dependent variables. Results suggest a significant effect of training condition, $F$ (12,235) = 5.167, $p$< .05, Wilks = .539. Table 2 displays the mean values for the Cronbach accuracy components for the OP ratings for each group. Lower values on these accuracy components denote greater rater accuracy. Table 3 reports the correlations amongst the accuracy components for OP.

Table 2

Means and Standard Deviations for Cronbach's (1955) Accuracy Component Scores on OP Ratings for Each of the Training Conditions

| Training Group | | E | DE | SA | DA |
|---|---|---|---|---|---|
| Control | M | .4438 | .4549 | .3040 | 1.1115 |
| | SD | .2905 | .1621 | .2103 | .5985 |
| FORD | M | .1711 | .2333 | .1678 | .5174 |
| | SD | .1912 | .1664 | .2092 | .4278 |
| FORD + FORS | M | .0968 | .2020 | .1309 | .4377 |
| | SD | .0900 | .1511 | .1362 | .2803 |
| FORTI | M | .1398 | .2232 | .1777 | .4043 |
| | SD | .1928 | .1357 | .1780 | .1994 |

Note. Low values denote greater rater accuracy.

Table 3

Correlations between Cronbach's (1955) Accuracy Component Scores for OP Ratings

| Variable | E | DE | SA | DA |
|---|---|---|---|---|
| E | - | | | |
| DE | .54** | - | | |
| SA | .53** | .48** | - | |
| DA | .51** | .71** | .30** | - |

*p<.05. **p<.01. (one-tailed)

A follow-up discriminant function analysis (DFA) revealed one significant eigenvalue, *p* < .01, with training type accounting for 98.05% of the variance in the accuracy composite. DFA results also indicate that DA, DE, and E contributed most significantly to the composite (structure coefficients were as follows: DA, .795; DE, .746; SA, .385 and E, .748). Consistent with Hypothesis 1, the group centroids for the first function suggest a discrimination between the control group (1.52) and the FORD, FORD + FORS, and FORTI groups (centroids = -.31, -.66, and -.55, respectively) such that the control group evidenced the lowest levels of rating accuracy on OP ratings.

Given that each accuracy component was potentially interesting, univariate planned comparisons were conducted for each of the accuracy components. For each accuracy component, the control group was compared to the FORD, FORD + FORS, and FORTI groups. To protect the familywise error rate at .05, a Bonferroni correction was made and each test was conducted at an alpha rate of .0125. Significant (p<.0125) effects were found for elevation, *t* (92) = 6.409, $\eta^2$ = .309,differential elevation, *t* (92) = 6.474, $\eta^2$ = .313, stereotype accuracy, *t* (92) = 3.313, $\eta^2$ = .107, and differential accuracy, *t* (92) = 6.879, $\eta^2$= .340. Similar to the multivariate results, the univariate results suggest that those groups receiving FORD training (the FORD, FORD + FORS, and FORTI groups) were more accurate for OP ratings than the control group lending support to Hypothesis 1.

### Situational Constraint Accuracy

To test Hypothesis 2, which states that the groups receiving FORS will have greater accuracy on the SC ratings than those that do not, a multivariate analysis of variance of the training conditions was conducted using Cronbach's (1955) accuracy components for SC ratings as the dependent variables. A significant effect of training condition was found, *F* (12,236) = 7.919, *p* <.05, Wilks = .408. Table 4 displays the value of each of Cronbach's accuracy indices for the SC ratings for each of the four groups. Table 5 displays the correlations between the accuracy components for the SC ratings.

Table 4

Means and Standard Deviations for Cronbach's (1955) Accuracy Component Scores on SC Ratings for Each of the Training Conditions

| Training Group | | E | DE | SA | DA |
|---|---|---|---|---|---|
| Control | *M* | .6480 | .4025 | .4962 | .9160 |
| | *SD* | 1.0481 | .1698 | .4103 | .5920 |
| FORD | *M* | .5563 | .4987 | .4314 | 1.0292 |
| | *SD* | .7137 | .3035 | .2945 | .6574 |
| FORD + FORS | *M* | .0461 | .1971 | .0599 | .4049 |
| | *SD* | .0857 | .1493 | .0738 | .2756 |
| FORTI | *M* | .0432 | .1575 | .1345 | .3725 |
| | *SD* | .0582 | .1135 | .2486 | .2944 |

*Note.* Low values denote greater rater accuracy.

Table 5

Correlations between Cronbach's (1955) Accuracy Component Scores for SC Ratings

| Variable | E | DE | SA | DA |
|---|---|---|---|---|
| E | - | | | |
| DE | .18* | - | | |
| SA | .06 | .34** | - | |
| DA | .43** | .79** | .24* | - |

*p<.05. **p<.01. (one-tailed)

A follow-up discriminant function analysis (DFA) revealed one significant eigenvalue, $p < .01$, with training type accounting for 95.37% of the variance in the accuracy composite. DFA results also indicate that DA, DE, and SA contributed most to the composite (structure coefficients were as follows: DA, .539; DE, .621; SA, .577 and E, .394). Consistent with Hypothesis 2, the group centroids for the first function suggest a discrimination between the control and FORD groups (centroids = 1.08 and 1.17 respectively) and the FORD + FORS, and FORTI groups (centroids = -1.14 and -1.10 respectively) such that those groups which received FORS exhibited greater SC rater accuracy.

Given that each accuracy component was potentially interesting, univariate planned comparisons were conducted for each of the accuracy components. For each accuracy component, the control and FORD groups were compared to the FORD + FORS, and FORTI groups. To protect the familywise error rate at .05, a Bonferroni correction was made and each test was conducted at an alpha rate of .0125. Significant ($p<.0125$) effects were found for elevation, $\underline{t}$ (92) = 4.332, $\underline{\eta}^2$ = .169, differential elevation, $\underline{t}$ (92) = 6.779, $\underline{\eta}^2$ = .333, stereotype accuracy, $\underline{t}$ (92) = 6.327, $\underline{\eta}^2$ = .303, and differential accuracy, $\underline{t}$ (92) = 5.885, $\underline{\eta}^2$ = .273. Similar to the multivariate results, the univariate results suggest that those groups receiving FORS training (the FORD + FORS, and FORTI groups) were more accurate for SC ratings than the control and FORD groups lending support to Hypothesis 2.

### Deserved Performance Accuracy

Hypothesis 3 suggested that the FORTI group would be significantly more accurate on DR ratings than the other three training conditions. A multivariate analysis of variance testing this hypothesis found a significant effect of training condition, $\underline{F}$ (12, 236) = 7.199, $\underline{p} < .05$, Wilks = .438. Table 6 displays the value of each of Cronbach's (1955) accuracy components for DR ratings for each of the four groups. Table 7 displays the correlations between the accuracy components.

Table 6

Means and Standard Deviations for Cronbach's (1955) Accuracy Component Scores on DR Ratings for Each of the Training Conditions

| Training Group | | E | DE | SA | DA |
|---|---|---|---|---|---|
| Control | M | .4867 | .4876 | .3474 | .9313 |
| | SD | .4759 | .2215 | .2583 | .3487 |
| FORD | M | 1.0321 | .3677 | .2407 | .7001 |
| | SD | .7019 | .2327 | .2640 | .2837 |
| FORD + FORS | M | 1.0069 | .3895 | .1647 | .7353 |
| | SD | .6204 | .2119 | .1585 | .3093 |
| FORTI | M | .0502 | .2653 | .1296 | .4411 |
| | SD | .0440 | .1901 | .2214 | .2079 |

Note. Low values denote greater rater accuracy.

Table 7

Correlations between Cronbach's (1955) Accuracy Component Scores for DR Ratings

| Variable | E | DE | SA | DA |
|---|---|---|---|---|
| E | - | | | |
| DE | .16 | - | | |
| SA | -.04 | .43** | - | |
| DA | .23* | .72** | .39** | - |

*p<.05. **p<.01. (one-tailed)

A follow-up discriminant function analysis (DFA) revealed two significant eigenvalues, p < .01. Training type accounted for 69.25% of the variance in the accuracy composite in the first function. DA and E contributed most significantly to the composite (structure coefficients were as follows: DA, .51; DE, .29; SA, .20 and E, .88). Consistent with Hypothesis 3, the group centroids for the first function suggest a discrimination between the control, FORD, FORD + FORS groups (centroids = .14, .67, and .59, respectively) and the FORTI group (-1.40) such that the group which received TI training exhibited greater DR accuracy.

For the second function, training type accounted for 28.42% of the variance in the accuracy composite. All of the accuracy components appeared to contribute to the composite (structure coefficients were as follows: DA, .78; DE,

.51; SA, .57; and E, -.46). The group centroids for the second function suggest a discrimination between the control group (.92) and the FORD, FORD + FORS, and FORTI groups (centroids = -.35, -.35, and -.22 respectively) such that the control group exhibited less DR accuracy.

To further explore Hypothesis 3, a multivariate analysis of variance was conducted which compared the FORD + FORS group to the FORTI group on the accuracy of DR ratings. A significant effect was found, $F$ (4, 43) = 15.488, $p<.05$ with the FORTI exhibiting greater accuracy than the FORD + FORS group.

Given that each accuracy component was potentially interesting, univariate planned comparisons were conducted for each of the accuracy components. For each accuracy component, the FORTI group was compared to the control, FORD, and FORD + FORS groups. To protect the familywise error rate at .05, a Bonferroni correction was made and each test was conducted at an alpha rate of .0125. Significant (p<.0125) effects were found for elevation, $t$ (92) = 6.387, $\eta^2$ = .307, differential elevation, $t$ (92) = 2.958, $\eta^2$ = .087, and differential accuracy, $t$ (92) = 5.054, $\eta^2$ = .217. The effect was not significant for stereotype accuracy, $t$ (92) = 2.242, $p$ = .027, $\eta^2$ = .052. Similar to the multivariate results, the univariate results generally lend support to Hypothesis 3 suggesting that the FORTI group is more accurate than the other groups for DR ratings.

## Cue Intercorrelation

Prior to carrying out the within subjects analysis, the UIs were standardized using Fisher's r to z transformation. The standardization of all values was needed because of the violation of the normality assumption when correlations are not transformed. Hypothesis 4 suggested that for those groups that did not receive TI training, UIs for the SC cue would increase as the correlation between the cues increased. Support for Hypothesis 4 was not found in that a repeated measure analysis of variance found the linear effect of cue correlation on the UI of the SC cue was not significant, $F$ (1,71) = 2.92, $p$= .092, $\eta^2$ = .040.

## Exploratory Analyses

Because of the relative paucity of research that examines how raters utilize situational constraint information in determining performance ratings, it was deemed interesting to examine the weighting policies of the participants. Participants in the FORTI group were supplied with an algorithm by which to determine DR ratings and the adoption of this algorithm is reflected in the accuracy of their DR ratings. However, participants in the control, FORD, and FORD + FORS groups were given no such algorithm. Therefore, examination of weighting policies was restricted to the control, FORD, and FORD + FORS groups which had the freedom to develop their own individual weighting policies.

A number of questions may be posed in regards to the participants' weighting strategies: 1) What relative importance are the OP and SC cues in determining DR for each rater?; 2) Do raters consistently weight cues in the same manner across profiles?; 3) Do raters utilize linear, non-linear, or interactive policies in determining DR?; and 4) Are weighting policies relatively homogeneous across raters or are they rater-specific?.

For participants in the control, FORD, and FORD + FORS groups, each participant's DR's for the 20 ratee profiles were regressed on OP and SC ratings to produce a multiple regression equation for each dimension for each participant. Estimates of linear policy consistency were obtained for each of the three dimensions by examining the squared multiple correlation ($R^2$) values between the DR ratings and the SC and OP cue values for each participant. Collapsed across the three dimensions, the squared multiple correlations ranged from 0.4735 to 1 with a mean of 0.8827. Figure 1 illustrates a frequency distribution of the squared multiple correlations.

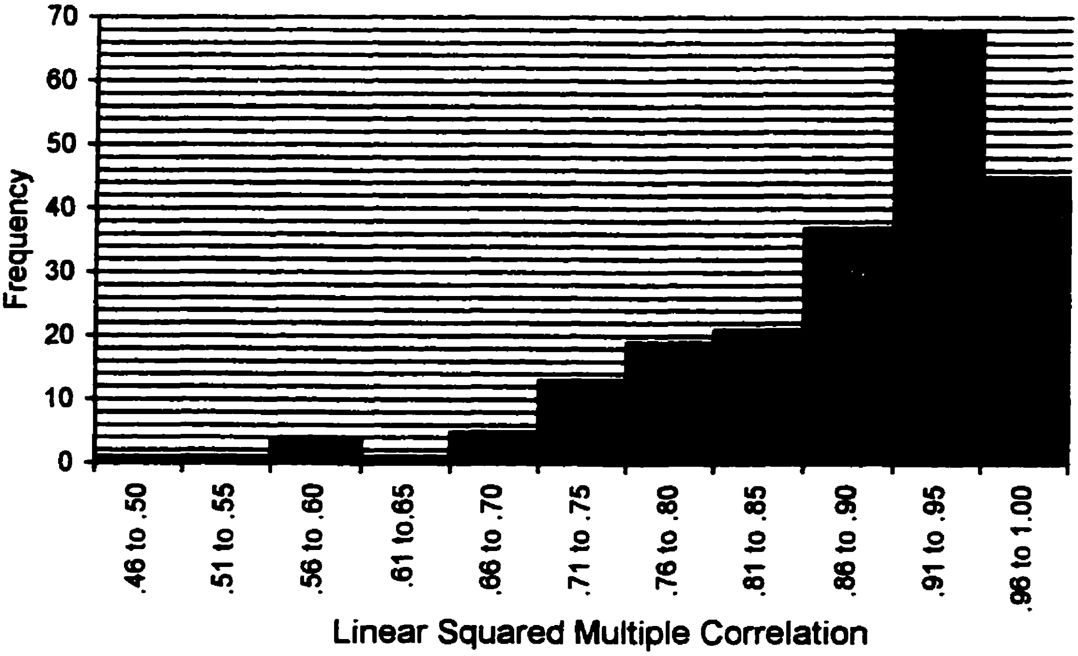

Figure 1. The Consistency of Raters' Linear Policies Across the Three Dimensions.

Configural analyses were conducted for each of the three dimensions for each participant to determine if cues were used in a nonlinear and/or interactive manner. First, the quadratic components of OP and SC were tested by calculating the change in $R^2$ resulting from the addition of these components to a regression equation containing the linear components of SC and OP. Secondly, the interaction of SC and OP was tested by calculating the increase in $R^2$ when the interaction term was added to the linear components of OP and SC. An increase in $R^2$ of 0.10 or greater that was statistically significant was used as an indication of the presence of a configural policy. Therefore, the incremental variance contributed by a nonlinear or interactive component had to be deemed both statistically and practically significant to be considered.

Of the 72 participants in the control, FORD, and FORD + FORS groups, four participants provided evidence of using an interactive policy for one or more of the dimensions. Four participants exhibited the use of a nonlinear policy for one or more of the dimensions. One participant had both a significant nonlinear and interactive component for their policy for one of the dimensions. These findings suggest that the majority of participants used a linear policy in deciding on DR ratings.

The statistical significance of each UI was computed, through an incremental $R^2$ significance test, to identify those cues that contributed significantly to each participant's rating strategy. For all participants, OP was a significant cue for all rating policies. However, SC only contributed significantly to 144 of the 216 rating policies. Therefore, approximately 33% of the rating policies did not use the information provided in the SC cue in determining the deserved performance rating.

One downfall of using tests of the UIs is that these numbers do not represent the manner in which cue information was used in determining DR ratings. In other words, it is impossible to determine the direction (up or down) DR ratings were adjusted according to cue information. Therefore, the positivity/negativity of the Beta weights were examined. A positive Beta weight would suggest that the participant provided increasingly higher DR ratings with increasing cue values. A negative Beta weight would suggest that the participant assigns decreasingly lower DR ratings with increasing cue values.

The Beta weights for OP for all 216 policies were in the positive direction. However, for the SC cue, 23 of the 216 Beta weights (11 %) were in the negative direction and the remaining 193 were positive.

## Discussion

The primary purpose of the present study was to examine the effects of FORD, FORS, and TI training on the accuracy of ratings of observed performance, situational constraints, and deserved performance. In addition, I examined the decision-making policies used to rate deserved performance by participants who were given no guidance in what weighting policy to use.

Examination of the rating policies used by participants in the control, FORD, and FORD + FORS groups suggest that participants do not use one universal rating policy in determining DR but policies tend to be participant-specific. In other words, there does not appear to be consensus amongst participants in how to combine information concerning observed performance and situational constraints in determining the rating the ratee is believed to deserve. The analysis of UIs suggests that while two-thirds of the participants incorporated information regarding situational constraints into their decision on DR ratings, one-third of the participants did not do so. Furthermore, examination of the beta weights for the SC cue show that while the majority of policies had a positive Beta weight associated with the SC cue, there were a few policies which had a negative Beta weight attached to the SC cue. This suggests that some participants compensated those ratees who experienced high situational constraints by increasing DR while others punished ratees experiencing situational constraints by decreasing DR.

These results suggest that it may be problematic to compare ratings completed by different raters because there is a general disagreement amongst raters regarding the importance of situational constraints in determining deserved performance ratings. Furthermore, even when there is agreement on the importance of situational constraints in the DR decision, there still may exist disagreement in regards to the best way to utilize situational constraint information. This implies that raters would benefit from any training or guidelines which instruct them on the preferred use of situational constraint information. Training may be one method of ensuring that a common theory of performance is used by all raters and ratings are, therefore, more easily comparable across raters.

Consistent with previous performance appraisal research, the present study found that those participants receiving traditional frame-of-reference training on the performance dimensions (the FORD, FORD + FORS, and FORTI groups) were significantly more accurate in rating observed performance than those that did not receive training (the control group) (Athey & McIntyre, 1987; Bernardin & Pence, 1980; Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; McIntyre et al, 1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr, 1994).

Support was also found for hypothesis 2 predicting that the accuracy of situational constraint ratings would be significantly higher for the FORD + FORS, and FORTI-trained participants compared to participants in the control and FORD groups. Similar to performance dimensions, raters may approach the ratings of situational constraints with differing theories concerning what constitutes a situational constraint and what the different levels of the rating scale represent. The results of this study suggest that raters may benefit from receiving a common frame-of-reference upon which to base ratings of situational constraints.

Support was found for hypothesis 3 predicting that the accuracy of deserved performance ratings would be significantly higher for the FORTI group compared to the control, FORD, and FORD + FORS-trained participants. The greater accuracy of DR for the FORTI group as compared to the other three groups suggests that the participants receiving the TI training adopted the weighting strategy as instructed.

To further determine whether the adoption of the prescribed weighting policy increased the accuracy of the DR ratings for the FORTI group, their accuracy was compared to the participants in the FORD + FORS group. These two groups differ only in the absence or presence of TI training; therefore, the statistically significant difference on DR accuracy found between these groups suggests that TI training was successful in aligning participants in one weighting strategy.

The discriminant function analyses on DRA highlighted two significant functions. The first function discriminated amongst the FORTI group and the other three groups. This function reflects differences in the groups due to adoption of the weighting policy as discussed previously. The second significant function, however, revealed a distinction between the control group and the other three groups. I believe DRA may have been influenced by two separate effects. The first would be the adoption of the instructed weighting policy. The second influence on DRA may be the accuracy of the cue ratings, namely OP and SC. If the ratings of OP and SC were extremely inaccurate, DR ratings as a result would also be inaccurate even when the appropriate weighting strategy was followed. The control condition tended to have low accuracy rates on both the OP and SC ratings. This may account for the distinction between the control condition and the other groups as highlighted by the second discriminate function.

Support was not found for hypothesis 4 stating that for those participants who did not receive training regarding the weighting policy the importance of the SC cue would increase as the correlation between the cues increased. No significant differences were found between the dimensions with high, moderate, or no correlation between the OP and SC cues in the UIs for the SC cue. This may suggest that participants do not use information about covariance in making attributional judgments. However, past research has shown support for attributional theory which suggests that decision makers assess covariance between variables and use this information to make cause/effect attributions (Cordray, & Shaw, 1978; Karaz & Perlman, 1975; Zuckerman, 1978).

Research does suggest, however, that participants often have difficulty assessing covariation in some situations and covariation is most accurately assessed when all information can be observed simultaneously, when the data are summarized for the participants, when instructions are clear, and when participants are repeatedly exposed to the data (Cordray & Shaw, 1978; Crocker, 1982). In the present study, information was observed serially, not simultaneously, the profiles were not summarized, and participants had only a brief single exposure to the profiles. Therefore, participants may have had difficulties assessing covariation between the cues and as a result did not use this information in the decision-making process.

Furthermore, there may have been difficulties in the design of the stimulus materials which also may explain the lack of findings. Participants were asked to make numerous decisions and to develop three separate weighting policies. Furthermore, for each dimension there was a different level of correlation between the cues. This may have been too much information for the participant to process simultaneously and participants may have had difficulty keeping dimensional information separated. If this is the case, covariation information may have not been readily available for participants to base their judgments.

## Implications and Future Research

Currently research in performance appraisal has focused on the cognitive processes of the rater, however, the attribution process has largely been neglected. The present study provides support for the need to consider contextual factors and the attribution process in performance appraisal research (cf. Feldman, 1981).

Future research will need to examine mechanisms other than training which may aid raters in integrating situational information in their decision process when providing performance ratings. For example, it may be suggested that having employees rated by a number of sources including self ratings instead of focusing merely on supervisor ratings may increase awareness of the influences of situational constraints. Carson et al. (1991) found supervisors and subordinates have different perceptions of the impact of situational factors on performance. Furthermore, different sources of ratings may also have differing opportunities to observe situational constraints. Perhaps organizations can take advantage of these different perceptions by utilizing multiple sources of ratings. Different perceptions of the influence of situational factors may also explain the discrepancies often found between supervisor and self ratings (Harris & Schaubroeck, 1988).

When developing the stimulus materials, it was noted that in the focus groups there was a fair amount of variability amongst individuals in beliefs regarding the impact of situational constraints on performance outcomes. It would be of interest to examine how individual difference or personality variables relate to beliefs regarding situational constraints. For instance, Locus of Control is one such variable that impacts how individuals perceive the world around them (Rotter, 1966). Those individuals with an internal locus of control believe that life events are controllable. Conversely, those individuals with an external locus of control believe that outcomes are generally beyond personal control. It may be expected that those with an external locus of control may be more likely to believe that situational constraints (or contextual factors beyond their control) impact on performance.

## Limitations of the Study

The present study may be limited in that the variables were manipulated in a laboratory setting. In the organization, supervisors have more knowledge about employees' behaviors, are able to interact personally with employees, and usually have more information about the tasks involved in the job. Therefore, an examination of the influence of situational factors on the accuracy of performance evaluations in an organizational setting would be valuable. It would also be beneficial to study the potential of the training program within an organizational setting.

The present study directly asked raters to assess situational constraints. Performance appraisals in an organization, however, often do not include the assessment of situational constraints. The process of evaluating the situational constraints in the present study may have made this variable much more salient to the rater than it otherwise would be. Therefore, raters may be more likely to include this information in their rating of deserved performance. An examination of raters' weighting strategies when they are not asked to directly evaluate situational constraints would be of interest.

The present study utilized "paper people" or written profiles to depict ratees' performance. Traditional frame-of-reference training tends to use videotaped vignettes. This difference may cause concern in the comparability of the results to previous frame-of-reference research. However, an empirical study by Woehr and Lance (1991) suggests that it is the amount of noise present in the two different media which account for differences in rater accuracy rather than differences in the type of cognitive processing required by each medium. In other words, ratings may be more accurate in written profiles because paper people have fewer irrelevant stimuli that may distract raters as compared to videotaped profiles but the cognitive processes used by the rater are similar in the two cases. This would suggest that results found via written profiles would be generalizable to cases where performance is directly observed.

Although the present study may be a simplified model examining how raters utilize situational constraint information in performance appraisal, it is a good starting point to a long neglected subject in performance appraisal research. It suggests that there is a lack of consistency amongst individual raters regarding the use of situational constraint information in final performance appraisal ratings. Furthermore, frame-of-reference training can be used to rate both performance and situational constraint dimensions. Finally, training may be an effective means to unite raters in one universal weighting policy and results in the greater accuracy of deserved performance ratings.

# References

Athey, T.R., & McIntyre, R.M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. Journal of Applied Psychology, 72, 567-572.

Balzer, W.K., Doherty, M.E., & O'Connor, R., Jr. (1989). The effects of cognitive feedback on performance. Psychological Bulletin, 106, 410-433.

Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance? Organizational Behavior and Human Decision Processes, 53(1), 35-54.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

Bernardin, H.J., & Pence, E.C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Cardy, R.L., & Keefe, T.J. (1994). Observational purpose and evaluative articulation in frame-of-reference training: The effects of alternative processing modes on rater accuracy. Organizational Behavior and Human Decision Processes, 57, 338-357.

Carson, K. P., Cardy, R. L., & Dobbins, G. H. (1991). Performance appraisal as effective management or deadly management disease: Two initial empirical investigations. Group and Organization Studies, 16(2), 143-159.

Cordray, D.S., & Shaw, J.I. (1978). An empirical test of the covariation analysis in causal attribution. Journal of Experimental Social Psychology, 14, 280-290.

Craik, F.I.M., & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 11, 671-684.

Crocker, J. (1982). Biased questions in judgment of covariation studies. Personality and Social Psychology Bulletin, 8, 214-220.

Cronbach, L.J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 52, 177-193.

Darlington, R.B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69(3), 161-182.

Day, D.V., & Sulsky, L.M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. Journal of Applied Psychology, 80(1), 158-167.

Deming, W.E. (1986). Out of the crisis. Cambridge: MIT Institute for Advanced Engineering Study.

Doherty, M.E., & Balzer, W.K. (1988). Cognitive feedback. In B. Brehmer & C.R.B. Joyce (Eds.), Human judgment: The SJT approach. Amsterdam: North-Holland.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Harris, M.M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-62.

Hauenstein, N. M., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. Personnel Psychology, 42(2), 359-378.

Hobson, C. J., Mendel, R. M., & Gibson, F. W. (1981). Clarifying performance appraisal criteria. Organizational Behavior and Human Performance, 28(2), 164-188.

Karaz, V., & Perlman, D. (1975). Attribution at the wire: Consistency and outcome finish strong. Journal of Experimental Social Psychology, 11, 470-477.

Kelley, H.H. (1973). The process of causal attribution. American Psychologist, 28, 107-128.

Kline, T.J.B., & Sulsky, L.M. (1995). A policy-capturing approach to individual decision-making: A demonstration using professors' judgements of the acceptability of psychology graduate school applicants. Canadian Journal of Behavioural Science, 27(4), 393-404.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87(1), 72-107.

McCauley, D.P., Jago, I.A., Gore, B., Lance, C.E., Quarles, F.K., Sledge, L., Pate, J.L., Logan, A.L., & Guest, F.D. (1990, April). The development of more ecologically valid videotapes for use in performance appraisal research. Paper presented at the annual meeting of the Southeastern Psychological Association, Atlanta, GA.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69(1), 147-156.

O'Connor, E.J., Peters, L.H., Pooyan, A., Weekley, J., Frank, B., & Erenkrantz, B. (1984). Situational constraint effects on performance, affective reactions, and turnover: A field replication and extension. Journal of Applied Psychology, 69, 663-672.

Peters, L.H., Fisher, C.D., & O'Connor, E.J. (1982). The moderating effect of situational control of performance variance on the relationship between individual differences and performance. Personnel Psychology, 35, 609-621.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influences of a frequently overlooked construct. Academy of Management Review , 5, 391-397.

Peters, L. H., O'Connor, E. J., & Rudolf, C. J. (1980). The behavioral and affective consequences of performance-relevant situational variables. Organizational Behavior and Human Performance, 25(1), 79-96.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69(4), 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. Organizational Behavior and Human Decision Processes, 38(1), 76-91.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (ed.), Advances in experimental social psychology, 10. New York: Academic Press.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80(1), 11-12.

Schneider, B. (1978). Person-situation selection: A review of some ability-situation interaction research. Personnel Psychology, 31(2), 281-297.

Seitz, C. (1988). Contextual factors in performance ratings: A policy capturing approach. Unpublished doctoral dissertation, Bowling Green State University, Bowling Green.

Stamoulis, D.T , & Hauenstein, N.M.A. (1993). Rater training and rater accuracy: Training for dimensional accuracy versus training for ratee differentiation. Journal of Applied Psychology, 78(6), 994-1003.

Steel, R. P., & Mento, A. J. (1986). Impact of situational constraints on subjective and objective criteria of managerial job performance. Organizational Behavior and Human Decision Processes, 37(2), 254-265.

Steel, R. P., Mento, A. J., & Hendrix, W. H. (1987). Constraining forces and the work performance of finance company cashiers. Journal of Management, 13(3), 473-482.

Stumpf, S.A., & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. Academy of Management Journal, 24, 752-766.

Sulsky, L.M., & Balzer, W.K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 497-506.

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. Journal of Applied Psychology, 77(4), 501-510.

Sulsky, L.M., & Day, D.V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. Journal of Applied Psychology, 79(4), 535-543.

Taylor, R.L., & Wilstead, W. D. (1974). Capturing judgment policies: A field study of performance appraisal. Academy of Management Journal, 17, 440-449.

Terborg, J. R. (1977). Validation and extension of an individual differences model of work performance. Organizational Behavior and Human Performance, 18(1), 188-216.

Woehr, D.J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. Journal of Applied Psychology, 79(4), 525-534.

Woehr, D.J., & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. Journal of Occupational and Organizational Psychology, 67(3), 189-205.

Woehr, D.J., & Lance, C.E. (1991). Paper people versus direct observation: An empirical examination of laboratory methodologies. Journal of Organizational Behavior, 12(5), 387-397.

Zedeck, S., & Cascio, W. F. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67(6), 752-758.

Zuckerman, M. (1978). Actions and occurrences in Kelley's cube. Journal of Personality and Social Psychology, 36, 647-656.

Appendix A. Example Profile

## Dr. Armstrong

Dr. Armstrong always covers the material outlined in the allotted time period and keeps the material at a pace which neither bores nor overwhelms the students. Dr. Armstrong has taught the same class one or two times previously. Dr. Armstrong appears to be very prepared and the lecture follows a somewhat logical sequence. Dr. Armstrong ordered an overhead projector but media services failed to deliver it, however, a blackboard was available for use. Dr. Armstrong often solicits questions from the class but appears to be impatient and rushed in answering questions. There are approximately 20 students in class.

## Dr. Armstrong

### Speaking Ability

Observed Performance

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (extremely poor performance) | | | | | | (extremely good performance) |

Prior Lecturing Experience

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (no constraints) | | | | | | (severe constraints) |

Deserved Performance

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (extremely poor performance) | | | | | | (extremely good performance) |

### Organization

Observed Performance

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (extremely poor performance) | | | | | | (extremely good performance) |

Availability and Quality of Equipment

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (no constraints) | | | | | | (severe constraints) |

Deserved Performance

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (extremely poor performance) | | | | | | (extremely good performance) |

### Fielding Student Questions

Observed Performance

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (extremely poor performance) | | | | | | (extremely good performance) |

Class size

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (no constraints) | | | | | | (severe constraints) |

Deserved Performance

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| (extremely poor performance) | | | | | | (extremely good performance) |

IMAGE EVALUATION
TEST TARGET (QA-3)