THE UNIVERSITY OF CALGARY

Impact of Frame-of-Reference and Behavioral Observation Training on Rating and Behavioral Accuracy in Performance Appraisals

by

Lisa Noonan

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

JULY, 1996

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled “Impact of Frame-of-Reference and Behavioral Observation Training on Rating and Behavioral Accuracy in Performance Appraisals” submitted by Lisa Noonan in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor, Dr. L. M. Sulsky, Psychology

Dr. T. J. B. Kline, Psychology

Dr. D. Skarlicki, Psychology

Barbara L. Marcolin

Dr. B. Marcolin, Management

July 23, 1996

Date

ABSTRACT

The effects of frame-of-reference (FOR) and behavioral observation training (BOT) on behavioral accuracy and various indices of distance accuracy were investigated. Infantry officers and senior non-commissioned members (N = 97) were trained using either FOR, BOT + FOR, or control procedures. Participants then observed three infantry soldiers on videotape, and rated them on four performance dimensions from the military Performance Evaluation Report (PER). Self-reported reactions to the FOR and BOT + FOR programs were collected as well. Four months following training (Time 2), knowledge of the FOR training material and self-reported use of the FOR and BOT training information were assessed. Results indicate that compared to control participants, FOR and BOT + FOR-trained participants produced more accurate performance ratings. Additionally, BOT + FOR participants were significantly higher on behavioral accuracy compared to other participants. The reaction to training questionnaire revealed that FOR participants reacted more favourably to the training than BOT + FOR participants. At time 2, the BOT + FOR and FOR-trained participants exhibited greater knowledge of the FOR training content than controls. Lastly, FOR and FOR + BOT participants indicated that they used the FOR information at Time 2, although reported use was higher for certain performance dimensions.

ACKNOWLEDGEMENTS

I would like to express my gratitude and sincere appreciation to my thesis advisor, Dr. Lorne Sulsky. He served as a motivator, technical director, wealth of knowledge and source of inspiration throughout all phases of this project.

I would also like to thank my parents for their emotional support throughout graduate school, and Mark, for helping me keep my sense of humour and enduring my panic attacks during the final phases of this project.

Finally, I would like to thank Lynn, Janine, Joy and all my other colleagues and professors in the I/O/E Area Group for helping to create a fun, interesting, and challenging experience during the past two years.

TABLE OF CONTENTS

LIST OF TABLES

IMPACT OF FRAME-OF-REFERENCE AND BEHAVIORAL OBSERVATION TRAINING ON RATING AND BEHAVIORAL ACCURACY IN PERFORMANCE APPRAISALS

Performance appraisal is a task facing many supervisors in work organizations. Most appraisal systems rely on subjective evaluations of subordinate performance provided by their immediate supervisors (Bernardin & Beatty, 1984). Over the years, research has repeatedly demonstrated that supervisor ratings are frequently contaminated by a wide variety of rater errors, such as halo, leniency and central tendency which may render them questionable in terms of reliability, validity and accuracy (Bernardin & Pence, 1980; Borman, 1977; Landy & Farr, 1980).

Two strategies have typically been advanced to address the potential problems associated with subjective performance judgments: rating scale development and rater training (Woehr & Huffcutt, 1994). Unfortunately, research on rating scale comparisons indicates that rating format modifications alone do not result in much improvement in performance ratings (eg., Landy & Farr, 1980). As a result, the focus of research has shifted over the past two decades to rater training which has been shown to have greater potential than rating scale formats to improve the effectiveness of performance ratings (Woehr & Huffcutt, 1994). The benefits of rater training are twofold: (a) it enhances raters' knowledge and skills for carrying out subjective evaluations, and (b) it "motivates" raters to use the skills and knowledge they have acquired in the training program (McIntyre, Smith & Hassett, 1984).

The potential value of rater training has been recognized for some time. For example, training provided to American army officers on the performance dimensions of the military evaluation scale improved officers' ratings of their soldier's performance (Bitner, 1948).

In an early review of the training literature, Spool (1978) concluded that rater training in general seemed to be effective; but no conclusions were drawn with respect to the degree of success for different training programs. In a subsequent review, Smith (1986) reported how both the method of presentation (lecture, group discussion or practice/feedback) and the content of training (rater error training, performance dimension training, or performance standards training) influenced the effectiveness of training as measured by leniency and halo errors and rating accuracy.

With regard to content, Smith (1986) found that rater error training was successful at reducing halo and leniency errors but had limited or no effect on accuracy measures while performance dimension training by itself was unsuccessful in reducing leniency error or rating accuracy. According to Smith, the largest improvement in accuracy was a combination of the performance dimension and performance standards approaches which are the elements comprising what is now referred to as frame-of-reference (FOR) training. He also noted that combining rater error training with the other types of training failed to produce any significant increment in rating accuracy.

In a more recent recent review of rater training research, Woehr and Huffcut (1994) further expanded on Smith's (1986) framework by identifying four models of rater training based on the content of training; specifically, rater error, performance dimension, FOR, and behavioral observation training (BOT). They performed a meta-analysis of effect sizes from these various programs, concluding that the greatest improvements in rating accuracy were to be found in the FOR and BOT approaches. They noted that in spite of data indicating that BOT may be an effective approach for increasing rating and observational accuracy, there were few studies focusing on the effectiveness of BOT or observational accuracy dependent measures in general. Furthermore, there was little research investigating the impact of combined rater training strategies.

The purpose of this study is to contribute to the growing literature on rater training by comparing a FOR training program to a combined BOT and FOR training program to determine if rating accuracy could be increased by amalgamating the two types of training. Previous studies merged rater error and FOR training into one program, but found no significant increment in rating accuracy when combining the two approaches (McIntyre, et al, 1984; Pulakos, 1984). However, the FOR and BOT approaches have not been provided together to determine their combined effectiveness.

To date, no studies have examined the effectiveness of FOR training in a field setting. The emphasis on laboratory based training studies may be contributing to the gap between performance appraisal research and practice as described by Banks and Murphy (1985). To address this concern, the current study was conducted using a military population. Finally, a number of training effectiveness criteria were assessed, some after a four month delay (see below). No performance appraisal studies have been conducted examining training effectiveness with such a lengthy temporal delay. The BOT and FOR training approaches are discussed in detail in the following sections. A summary of the research associated with the variables pertinent to this study is also provided. This is followed by a presentation of the study hypotheses.

Frame-of-Reference Training

In the early 1980's, there was a shift from training raters to avoid halo, leniency and other common rating "errors" to more proactive rater accuracy approaches (Athey & McIntyre, 1987). It was found that traditional rater error training facilitated the learning of a new rating response set which usually resulted in reducing leniency and halo errors, but also inadvertently lowered levels of rating accuracy in some instances (Bernardin & Pence, 1980; Landy & Farr, 1980). For example, if rater training is intended to eliminate rating errors such as halo, but the "haloed" ratings are actually based on real attributes or behaviors of ratees, then reducing such errors removes true variance as well as error variance. Thus, the ratings become less accurate after rater training.

Bernardin and Buckley (1981) concluded that there was a need to develop new rater-training programs that increase rating accuracy, and they proposed FOR training as an alternative strategy. FOR training "tunes raters" to a common frame of reference so that worker behaviors can be similarly assessed by different raters (McIntyre et al, 1984). The goal is to enable raters to share and use common conceptualizations of performance so that they can make more accurate evaluations (Athey & McIntyre, 1987). Specifically, it involves matching ratee behaviors to their appropriate performance dimensions and correctly judging the effectiveness levels of specific ratee behaviors (Sulsky & Day, 1992). In sum, theories of performance for individual performance dimensions are imparted to raters to assist them in the accurate evaluation of ratee performance.

The rationale Bernardin and Buckley (1981) used to develop FOR training was inspired in part by the development of behavioral anchored rating scales

(BARS), (Smith & Kendall, 1963), in which referent anchors facilitate agreement in evaluating recorded or recalled behaviors. However, FOR training takes the process one step further than the BARS because raters are not required to attain a particular frame of reference on their own.

Originally, Bernardin and Buckley (1981) proposed FOR training for idiosyncratic raters, defined as individuals who do not provide accurate ratings when compared to "true" scores. The goal is to eliminate the idiosyncratic standards of these raters through FOR training, thus bringing them into closer congruence with the rest of the organization. Hauenstein and Foti (1989) suggested identifying idiosyncratic raters before implementing FOR training. However, Sulsky and Day (1992) found that even after receiving FOR training, 8 - 15% of the sample was still idiosyncratic, suggesting that that there may be additional ability or motivational factors contributing to their idiosyncratic status.

The majority of studies have not focused on idiosyncratic raters but have simply employed a random sample of raters. All of these studies have demonstrated the efficacy of FOR training for improving various measures of accuracy (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). Woehr and Huffcut (1994) in a meta-analysis of rater training programs, found an average effect size of .83 for FOR training compared to control or no training groups. However, all of the FOR studies were conducted in a laboratory setting, raising possible concerns about the generalizability of this form of training to the workplace.

Across FOR training studies, there has been variability in the rating scales and protocols used in training; these are presented in Table 1. As Table 1 illustrates, there have been a wide variety of rating scales employed, although training procedures have tended to become fairly standardized over time.

Having demonstrated the efficacy of FOR training for improving rating accuracy, the research emphasis has shifted in recent years to understanding why FOR training leads to more accurate ratings. In particular, several researchers have focused on the cognitive mechanisms underlying the success of FOR training because raters receive vast amounts of information through observation and social interaction, and this data must somehow be organized in memory (Day & Sulsky, 1995; Sulsky & Day, 1992, 1994; Woehr, 1994). Athey and McIntyre (1987) found that FOR-trained raters remembered more training content than did raters trained with other procedures. This finding was explained through levels-of-processing theory which describes retention of information as a function of the depth at which information is processed; information requiring more cognitive elaboration is better remembered than information requiring less elaboration. In their study, the authors maintained that FOR training information was processed at a deeper level than control training information which explains why the training content in the FOR condition was better remembered (and therefore, more accurate ratings were produced).

Interestingly, memory for ratee behaviors and rating accuracy may not be highly correlated. Sulsky and Day (1992) found that FOR-trained raters demonstrated superior overall rating accuracy, but forgot many specific individual behaviors and made a series of recognition-related errors for specific ratee behaviors. The theory they advanced to explain their findings is based on

Table 1

Comparison of FOR Rating Scales and Protocols

Study	Scale	Protocol
McIntyre et al, 1984	12-item scale developed by Costin (1974) for instructors; measured the dimensions of organization, clarity of communication, elocutionary skills, and intellectual stimulation; items were constructed as positively worded statements in a 7-point agree-disagree format	12 items on the scale were discussed; one videotape was presented to participants as a practice exercise, followed by presentation of true scores and explanations of how they were derived; participants then rated three videotaped lectures
Pulakos, E., 1984	Developed by Borman (1977) to evaluate managers dealing with a problem subordinate; contained the dimensions of controlling the interview, establishing and maintaining	12-15 participants were given a lecture discussing the multidimensionality of jobs; dimensions were each presented with an

	rapport, resolving conflicts, motivating the subordinate, and developing the subordinate; each dimension contained seven scale anchors, and for the purposes of training, two levels of effectiveness were described for each dimension	explanation of the anchors and examples of behaviors; trainees practiced rating using videotapes of managerial performance; they were given feedback on their ratings in relation to true scores; participants were then tested on six videotaped ratees
Pulakos, 1986	Same scale used in the Pulakos (1984) study	Identical procedure used in the 1984 study for evaluative scales; those using observational scales were taught effectiveness levels based on the number of times a

		manager exhibited critical behaviors for each dimension; participants rated six videotaped ratees
Athey and McIntyre, 1987	12-item rating scale used by McIntyre et al. (1984)	Same procedure as employed in the McIntyre et al. study (1984)
Sulsky and Day, 1992	7-point BARS developed by Borman (1978) was used to assess ratee performance; critical incidents of managerial performance were those developed by Roberson and Banks (1986); the videotapes consisted of eight fictitious managers interviewing problem subordinates as described above in the Pulakos (1984) study	FOR training procedures followed those employed in the Pulakos (1984) study
Stamoulis and Hauenstein (1993)	Videotapes featuring interviewer behaviors from Hauenstein (1987) were used; 18 vignettes in all were	Same procedures as those used by Pulakos (1984), but

	scripted representing six dimensions for each of good, average and poor job-performance categories; a 7-item rating scale, also from Hauenstein (1987) was used	true scores were not derived by experts; instead, the mean ratings of each rating dimension were used
Cardy and Keefe, 1994	Five written vignettes of class-room teacher behavior representing 11-point scales for incidents on five dimensions as developed by Sauser et al. (1979); each vignette consisted of 10 behavioral incidents	Training was conducted using individual computer work stations; program contained instructions, allowed for rating practice ratees, provided feedback, presented target ratees and collected ratings/ reaction times; training was self-paced; accuracy feedback used true scores from Sauser

		et al. (1979)
Woehr (1994)	Instructor performance rating form and videotapes developed by McCauley et al. (1990) which featured six dimensions; each dimension was anchored on a 7-point Likert-type scale; four performance vignettes were developed using all six dimensions	Same procedure employed by Pulakos (1984) but training sessions lasted 45 minutes
Sulsky and Day, 1994; Day and Sulsky, 1995	Three 7-point BARS developed by Borman (1978); videotapes developed for these studies consisted of managers talking to problem subordinates; dimensions were the same as those used by Borman (1977)	Same procedure employed by Pulakos (1984)

the popular notion of cognitive categorization (Feldman, 1981) which suggests that raters categorize ratees on the basis of prototypes obtained through FOR training. These categorizations may serve as the basis for judgments to a greater extent than memory for specific behavioral information as they influence subsequent cognitive processing (Markus & Wurf, 1987).

Sulsky and Day (1992) hypothesized that FOR training is successful at enhancing rating accuracy because it allows ratees to be correctly categorized in terms of their performance. These categorizations, in turn, are based upon performance prototypes which they use to categorize ratee performance on each performance dimension. The prototypes are based upon a theory of performance which is developed before training to define the behaviors constituting varying performance levels on each dimension. Raters are assumed to categorize performance and engage in "on-line" processing rather than memory-based processing whereby raters rely upon memory for discrete ratee behavior (cf. Hastie & Park, 1986). In sum, the FOR program is not designed to facilitate memory for specific ratee behaviors.

Because the goals of many appraisal systems, (including the military system considered in this study), are to formulate personnel decisions (e.g., promotions) and to provide developmental feedback, it may be useful to consider an alternative form of training that serves to help raters attend to and remember specific ratee behaviors. Such a program is considered next.

Behavioral Observation Training

Another approach to rater training focuses on rater observation of behavior as opposed to rater evaluations of behavior. Subjective performance appraisal systems rely on observing behavior to provide a judgment. Unfortunately, the reliability and accuracy of observational methods have been questioned given that they are more vulnerable to the falibilities of human perceivers than most other methods (Weick, 1968).

The observer accuracy problem has been conceptualized as a function of three factors: (a) recording procedure characteristics, (b) conditions of observation and (c) observer characteristics (Weick, 1968). Recording procedure characteristics include such items as the complexity of categories in the coding system and the type of recording device used, while conditions of observation refer to such issues as (a) observee characteristics, (b) the number of subjects being observed, (c) the frequency and rate with which behaviors occur, and (d) the temporal sequencing of behavior (Cronbach, Gleser, Nanda & Rajaratnum, 1972). Together, these classes of factors constitute one approach to observer inaccuracy, what Weick (1968) referred to as "methodological solutions to bias". Another approach has focused on the third of these factors; observer characteristics which refer to the age, sex, expectancies, and intelligence of the observer, and includes any prior observational experience (Cronbach et al., 1972).

Of the observer characteristics, it is recognized that the ability level of the observer may be the key to minimizing observer error (Spool,1978). Spool noted that in order for the observer to be more accurate in observing, he or she must, among other things, be able to recognize the behaviors to be observed, be able to use the observer system with ease, and be able to make observations in accordance with some standard of criteria. Raters must not only be able to accurately observe behavior but must also develop the ability to recall the behavior of a number of ratees over a considerable span of time because performance appraisals are typically completed on a semiannual or annual basis (Murphy, Martin, & Garcia, 1982).

A logical means of developing these skills is to train the observer; however, little research has been conducted to determine what kinds of training programs increase observer accuracy and even fewer studies have attempted to reduce rater bias in the context of performance ratings (Woehr & Huffcutt, 1994).

One of the earliest BOT studies (Latham, Wexley & Pursell, 1975), compared a workshop, a group discussion and a control group. Managers in each of the groups were provided with information on common observational errors occurring in performance appraisal and selection interviews, including first impressions, similar-to-me, contrast and halo effects. Participants in the workshop group viewed videotaped vignettes of hypothetical job candidates being appraised by a manager who committed a different type of observational error in each vignette. Trainees then gave a rating of how they thought the manager in each vignette would evaluate the candidate and how they would evaluate the candidate themselves. Group discussion followed as to the reason for each trainee's ratings, and ways that they could avoid committing observational errors. Thus, trainees viewed a videotaped model, had the opportunity to practice, and received feedback.

In the group discussion, participants were given an example of each type of observational error in performance appraisal, selection and off-the-job situations. The trainees then generated and shared personal examples of observation-related problems and devised solutions to each of the rating problems that had been used in the workshop condition.

Finally, managers in the control condition were provided with a lecture on the errors mentioned above. Individuals were tested six months after the training program for halo, first impression, contrast and similarity errors. Results indicated that the control group exhibited significantly more errors compared to trainees in the other groups and the workshop group was marginally more effective than those involved in the group discussion in eliminating rating errors. A reaction measure administered six months after the training indicated that participants preferred the workshop format because it was more highly structured than the discussion group, provided more feedback from the trainers, and employed videotapes which allowed them to actually practice their newly learned skills.

Bernardin and Walter (1977) examined the use of a diary to increase observational accuracy. Student raters were required to keep an observational diary of their instructor in which they identified critical incidents relating to seven dimensions of performance featured on a behavior expectation scale (BES). They were advised that the diaries would be collected and verified for accuracy at the end of the semester. Dependent measures were leniency, interrater reliability, halo and discrimination across ratees. The group receiving psychometric training and exposure to the evaluation scale prior to and during observation showed significantly less leniency error and halo than all other groups. Although the members who were required to maintain a diary found this to be an extremely useful technique, the results of this study are questionable because subjects were not given the opportunity to practice with the rating scale, and no feedback was provided initially, or when the diaries were returned. Moreover, in the absence of true scores, it is unclear whether the results actually indicate improved rating quality.

The studies described above focused on improving the quality of ratings as opposed to directly enhancing observation processes. Thornton and Zorich (1980) maintained that previous research featuring observation training had not made a clear distinction between the processes of observation and that of judgment for performance ratings. They argued that judgment processes include the categorization, integration and evaluation of information, while observation processes are more basic, involving the detection, perception, and recall or recognition of specific behavioral events.

Thornton and Zorich (1980) were concerned with improving observational processes, maintaining that rating errors such as leniency and halo are primarily due to a lack of information stemming from problems in observation. For example, halo in ratings may occur when there is a lack of information about the ratee which results in overreliance on one type of information about the ratee. Consequently, they developed a rater training approach which focused on improving behavioral observations to increase performance rating accuracy.

Their training procedures consisted of three lectures, each representing an experimental condition. Participants in all three conditions were then shown a 45 minute videotape portraying three male and three female managers having a group discussion. In the behavioral instruction condition, raters were instructed to observe carefully, watch for specific behaviors in the videotapes and take notes. In the second error instruction condition, the previously mentioned precautions were provided in addition to a presentation of eight systematic errors of observation which people introduce into a communication system such as simplification, middle message loss, contrast effects, and various stereotypes. In the control condition, participants were provided with minimal information in a lecture format. Observational accuracy was operationalized as the number of correct answers on a questionnaire consisting of true-false, multiple choice and matching formats; items were selected from the videotaped discussion.

It was found that behavioral and error instruction groups evidenced significantly greater levels of observational accuracy than the control group. This study, however, employed undergraduate students; consequently, the authors were unable to state whether the results would generalize to a work population. In addition, performance appraisal research has indicated that a lecture training format, such as the one employed in the Thornton and Zorich (1980) study, may not have a lasting impact on trainees. Workshops result in more enduring behavioral changes, largely because of the feedback component and the opportunity to practice rating others (Latham, Wexley & Pursell, 1975).

Pulakos (1986) provided participants with a lecture on the importance of attending to relevant ratee behaviors (as opposed to traits) and on the difference between merely "looking for" certain behaviors and "forming judgments" of ratee effectiveness. Trainees were told that raters often make immediate judgments of ratees that are based on far too little information and are thus often incorrect. It was explained that focusing on observing and counting relevant behaviors should help them avoid premature judgments and hence rate more accurately.

Raters were then asked to memorize a list of behaviors corresponding to several performance dimensions and without referring back to the list, they were instructed to write down the dimension titles and the behaviors that fell within each dimension. This task was repeated and subsequent to each trial, subjects corrected their responses by consulting the list of behaviors. The goal was to sharpen raters' observational skills by teaching them what particular behaviors they should recognize when observing the ratees. Trainees were also told that the use of a mental checklist should help them to keep track of how often relevant behaviors occurred and while viewing videotaped ratees, the raters had to indicate whether the behaviors exhibited on the tapes was one of the target ones memorized.

Pulakos adopted Cronbach's (1955) accuracy measures (see below) and found that subjects receiving observational training evidenced significantly higher rating accuracy when using a behavioral observation scale format (Borman, 1979) compared to subjects receiving evaluative (identical to FOR) or control training.

Lastly, Hedge and Kavanagh (1988) compared BOT to control and decision-training groups (the latter was similar to FOR training). Training to improve observational skills involved instructions to observe carefully, watch for specific behaviors, and take notes whenever possible. In addition, several systematic errors of observation such as contamination from prior information, and overreliance on a single source of information were discussed in terms of the raters' ability to recognize and avoid them. Two workshop-style exercises were also conducted using videotapes to discuss observation errors and emphasize appropriate observation behaviors on the job. Significant improvements in rating accuracy for the decision-making and BOT groups were found while rating accuracy for the rater error group actually decreased.

The assumption in the studies outlined above is that better observation of behavioral information will result in improved rating accuracy. Given that one of the primary purposes of performance appraisals is to provide specific feedback for training and development, observational accuracy is likely to be at least as important a criteria as evaluative rating accuracy. Indeed, observational accuracy may be an important predictor of behavioral accuracy defined by Lord (1985) as a "rater's veridical encoding and recall of specific behaviors" (p. 67).

Unfortunately, only two studies have directly assessed observational accuracy (Murphy, Garcia, Kerkar, Martin & Balzer, 1982; Thornton & Zorich, 1980) and in both instances, observational accuracy was assessed by requiring raters to respond to a rating scale after viewing videotaped ratees. The ability of raters to recall specific information on the videos, however, may have influenced observational accuracy scores, and raises the question of whether observational accuracy was actually being measured in these studies (perhaps participants' short-term memory was also a determinant of their accuracy scores).

It is evident from the preceding discussion of BOT and FOR training studies that variability exists in the ways in which accuracy (e.g., evaluative vs. observational) has been conceptualized and measured. What follows next is a discussion of the various conceptual and operational definitions of accuracy across performance appraisal training studies.

Conceptualization of Accuracy in Performance Appraisal Training Research

Although the integrity of the appraisal process partly determines the accuracy of the appraisal, ultimately, the perceived correctness of rater judgments about the performance of others rests on the criterion against which rater accuracy is measured (Zalesny & Highhouse, 1992). Measures of rating error (e.g., halo, leniency and range restriction) have often been used to evaluate the quality of performance appraisals, particularly in the rater-error training (RET) and BOT studies of the 1970's (see for example Smith, 1986). This has proven to be problematic because there are numerous conflicting conceptual and operational definitions of these measures (Murphy & Balzer, 1981). It has also been noted that the absence of rating errors does not necessarily imply that ratings will be more accurate (Bernardin & Pence, 1980). Borman (1979) concluded that this situation was unfortunate because accuracy should be the "critical" criterion for judging the quality of performance ratings.

As a result of the inadequacies of using error indices to measure the quality of ratings, the use of accuracy scores has been prominent since the mid-1980's, and this is clearly evident in FOR training studies. Ratings are considered to be more accurate to the extent that raters evaluate performance in line with standards of performance (i.e., "true" scores) which are typically provided by "job experts". The external validity of studies involving the calculation of accuracy based on true scores depends, at least in part, on the relevance of the experts' true score estimates, with relevance defined as the degree to which the operationally defined true scores approximate the correct true scores (Smither, Barry & Reilly, 1989). Sulsky and Balzer (1988) reviewed a number of problems relating to true scores, arguing that the term "true score" is really a misnomer - they should be called target or comparison scores instead.

In an investigation of the validity of expert true score estimates, Smither et al. (1989) found that accuracy indices computed by using objective true scores were highly correlated with the same accuracy indices computed by using mean expert ratings, suggesting that expert true score estimates may serve as suitable substitutes when objective true scores are unavailable. However, they acknowledged that a high correlation between expert and objective true scores does not necessarily reflect a small difference between intended and estimated true scores.

For this reason, Smither et al. (1989) set out to test the validity of using "experts" to develop true scores. They compared true scores provided by student "experts" (who had enhanced opportunities to view videotaped job performances) with "nonexpert" true scores (provided by students who had only viewed the videotape once before rating the performances) using Cronbach's (1955) accuracy measures. The researchers found that expert raters were more accurate than nonexperts. The expert and nonexpert scores were then compared with the objective true scores using an overall distance measure (McIntyre et al., 1984) which provided an index of how close participant ratings were to the objective true scores. Borman's (1977) Differential Accuracy (DA) was also calculated by correlating the mean expert and nonexpert ratings for each dimension with corresponding objective true scores across ratees (see below). Borman's DA was above .90 for both experts and nonexperts (although expert scores were more highly correlated with objective true scores than those provided by nonexperts). However, with regard to distance accuracy, nonexperts' distance scores were larger, on average, than when expert true scores were used. In sum, these results provide some empirical evidence for the validity of expert ratings used as "true" scores in appraisal research.

Unfortunately, there is little agreement across studies concerning the specific accuracy measures to be employed in comparing raters' scores to these "true" scores (Stamoulis & Hauenstein, 1993). For example, many early studies examining FOR training used a simple distance index of accuracy (Athey & McIntyre, 1987; Bernardin & Pence, 1980; McIntyre et al, 1984; Sulsky & Day, 1992). More recently, however, studies have employed all four of Cronbach's (1955) accuracy component scores in examining the effects of FOR training (Day & Sulsky, 1995; Pulakos, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1994).

According to Cronbach (1955), a rater's overall accuracy is the sum of four separate components. Elevation (E) reflects the accuracy of the mean rating given by a rater over all ratees and dimensions. The rater whose overall average is close to the overall average true score will tend to be more accurate than one whose average rating is far from the true score average. Differential elevation (DE) is the accuracy of the average rating given to each ratee collapsed across performance dimensions (i.e., reflects a rater's accuracy in discriminating among ratees in terms of their overall performance). Stereotype accuracy (SA) is the component of accuracy associated with the average rating for each performance dimension collapsed across ratees (i.e., reflects a rater's accuracy in assessing the ratee group's strengths and weaknesses on various dimensions). Finally, differential accuracy (DA) is the interaction term, reflecting the accuracy with which ratees are rank ordered on each performance dimension.

Borman (1977) argued that differential accuracy (DA) is the most appropriate index for assessing the accuracy of performance judgments because correctly rank ordering target persons on each performance dimension seems to be the most important index of rating quality. Interestingly, Borman developed a distinct measure of DA which simply examines the correlation between ratings and true scores. According to Sulsky and Balzer (1988), Borman's DA measure is more useful as an index of rating validity.. Both of these measures of DA have been computed in some FOR training studies (e.g., Day & Sulsky, 1995; Jones, Sulsky & Day, 1995; Sulsky & Day, 1994; Sulsky, Day & Lawrence, 1994).

In recent research, there has been some debate as to which of Cronbach's (1955) components is most improved as a result of FOR training inasmuch as the four components represent different conceptualizations of the accuracy construct (e.g., Stamoulis & Hauenstein, 1993; Sulsky & Day, 1994; Woehr, 1994). FOR training has been shown to be superior at improving DA (Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994) which according to Day and Sulsky (1995) should be the component most affected by FOR training. Specifically, FOR training results in accurate ratee impressions on each performance dimension, and this is what DA is measuring - the ratee x performance dimension interaction.

Stamoulis and Hauenstein (1993) noted however, that FOR training should improve both DA and SA. They contended that FOR training is not designed to promote between-ratee differentiation as assessed by E (how well the overall mean rating matches the true score mean rating across all ratees and dimensions) and DE (how well each ratee was rated collapsed across all performance dimensions).

Alternately, Woehr (1994) argued that FOR training should have its strongest effects on DE and DA, and Sulsky and Day (1994) suggested that all four components may be equally affected in a given study. In sum, although this debate has yet to be resolved, it appears clear that DA is the one component of accuracy where there appears to be some consensus.

Thus far, the discussion has focused upon rating accuracy; however, another conceptualization of accuracy concerns observational accuracy. Thornton and Zorich (1980) admitted that they could not unequivocally say that observational training would improve properties of ratings because the "dependent measure in their study was a measure of recall and recognition of specific facts in the tape" (p. 353).

Murphy et al. (1982) were interested in the relationship between observational and evaluation accuracy and therefore, did not provide training to participants. Instead, they had participants view videotapes of lecturers and then asked them to evaluate the videotaped performances and rate the frequency of critical behaviors. Ratings were obtained immediately after viewing the short videotape which according to the researchers, minimized demands upon recall and therefore reflected differences in the accuracy with which raters observe ratee behavior. Nevertheless, as mentioned previously, the ability of the rater to recall details is a potential confound when measuring observational accuracy.

A final set of accuracy definitions that has some relevance to both FOR and BOT was developed by Lord (1985) who distinguished between classification and behavioral accuracy as a means of examining raters' cognitive processes. Classification accuracy (CA) refers to a rater's ability to correctly categorize ratees according to performance levels. It is operationalized on the basis of signal detection theory in that CA depends on the recognition of actually occurring behaviors (hits) or impression-consistent foils (false alarms) that would be expected on the basis of the ratee's performance level (Lord, 1985; Padgett & Ilgen, 1989).

Alternately, behavioral accuracy (BA) is based upon rater memory of specific behaviors and, unlike CA, does not depend upon the formation of ratee impressions (Sulsky & Day, 1992). In brief, BA involves examining the difference between the hit and false alarm rates on a recognition memory test and is conceptually close to sensitivity measures derived from signal detection theory (Lord, 1985; Padgett & Ilgen, 1989; Sulsky & Day, 1992).

Although rating accuracy and other conceptualizations of accuracy provide alternative means for evaluating training programs, our approach to program evaluation can be broadened even further. What follows next is another framework for thinking about how to evaluate rater training programs.

Training Program Evaluation

Kirkpatrick (1959) suggested that evaluation procedures for training programs should consider four levels of criteria - reaction, learning, behavior and results. What trainees think of a particular program constitutes the reaction component of his typology. In short, how did the trainees like the training program? Of course, this is multidimensional and may include such diverse issues as reactions toward the training content, training format and the specific trainer(s). Although there is not a consistent relationship between trainee reactions and performance (Alliger & Janek, 1989), it makes intuitive sense to design training programs that garner favorable reactions because training programs that make it miserable for trainees to learn will likely fail (Goldstein, 1993).

"Learning" refers to whether and the extent to which trainees learned the training content. An important issue here is retention. Assuming learning is successful, how long-lasting is the training material retained in memory? Learning can be assessed in a variety of ways. In performance appraisal research, it could be determined by raters' scores on a test of the training content and, less directly, by examining rating accuracy.

Sulsky and Day (1994) recently examined FOR trainees' learning after a two-day time delay, and found that even though there was significant forgetting of training material over time on a test of training content, it was insufficient to affect rating quality. In a real-world setting, time delays would be much longer because performance appraisals are normally only written every six months or annually, so a longer time delay may result in more extensive forgetting of the training material and thus, poorer rating accuracy.

"Behavior" refers to whether and the extent to which raters apply what they learned during training. This can be assessed indirectly through self-report data potentially provided by the rater, ratee or both. The real issue is the extent to which transfer occurred from training to the job (Landy, 1989).

Finally, "results" are measures of long-term payoff in organizational terms (i.e., the goal of training may have been to improve managerial skills, reduce the number of accidents or increase profits). In the case of performance appraisal training in the military, the results may be quite intangible. The "payoff" is more accurate performance evaluations which should result in improved morale among soldiers leading to higher levels of work performance. In addition, supervisors would have a greater understanding of the performance levels required for the various performance dimensions, and this should result in supervisors completing performance appraisals with greater ease, confidence and accuracy.

In sum, Kirkpatrick's (1959) typology is useful in so far as it provides a multidimensional framework for training program evaluation. Along with rating and behavioral accuracy, additional criteria (e.g., learning and subsequent use of the training material) allow for a more complete evaluation of training programs.

The Present Study

The study was conducted in two phases. At time 1, military personnel in a field setting were randomly assigned to either: (a) FOR training, (b) both BOT and FOR training, or (c) minimal control training conditions.

Following training, rating and behavioral accuracy were assessed. Specifically, participants rated videotaped soldiers who served the role of ratees which served as the rating accuracy measure, while the behavioral accuracy measure consisted of a test assessing recognition of ratee behavior. They were also given a measure which served as a manipulation check to determine whether or not FOR and BOT + FOR-trained participants paid attention to the FOR material taught during training. Moreover, following Kirkpatrick's (1959) typology, a measure was administered to participants in the FOR and BOT + FOR training groups to assess reactions to training.

Phase 2 of the study was conducted one week after the unit's annual performance evaluations had been prepared, (i.e., time 2), which was approximately four months following completion of training. Here, two criteria proposed by Kirkpatrick - learning and behavioral outcomes were assessed using a subset of the trainees. Participants from the three training conditions were administered a learning measure to determine whether they knew the training material initially presented during FOR training. To assess behavioral changes, BOT and FOR trained participants completed a self-report measure assessing whether and the extent to which they actually applied the training material over the time interval. Long-term results were not assessed in this study given the intangible nature of the "payoff" from improved rating accuracy in a military context, and the length of time required to assess such outcomes.

Because FOR training has been shown to improve rating accuracy across numerous studies, it was expected that participants exposed only to FOR training would evidence signficantly higher levels of rating accuracy at time 1 than participants in the control condition.

However, combining both BOT and FOR training was expected to lead to the highest levels of DA because BOT should help promote the formation of correct impressions (by helping to focus upon relevant behavior) which is the goal of FOR training. Additionally, BOT should have a direct effect on evaluative accuracy. Pulakos (1986), for example, found that observation training improved DA (compared to control training) and suggested that the training may have facilitated behavioral recall. Pulakos, in fact, recommended that future research examine a combined FOR and BOT protocol.

In summary, it was hypothesized that,

H1) the highest levels of rating accuracy would be obtained in the condition where both BOT and FOR training are provided, and the lowest levels of accuracy would be obtained for the control condition.

In particular, it was expected that the accuracy component most affected by FOR training would be DA (cf. Day & Sulsky, 1995). Nonetheless, all of the Cronbach (1955) accuracy components were examined in testing hypothesis 1.

Previous research suggests that FOR training leads to improvements in rating accuracy, however, the effects of FOR training on behavioral accuracy are unclear (cf. Sulsky & Day, 1992; Woehr, 1994). In contrast, the very nature of BOT is to improve observational skills: this should directly translate into improved memory for behavior in so far as more information is stored in memory in the first place. Thus, it was hypothesized that:

> H2) participants receiving BOT + FOR training would yield higher levels of behavioral accuracy than participants in the control or FOR only conditions.

Turning now to Kirkpatrick's (1959) criteria for training program evaluation, it was expected that because FOR training requires trainee assimilation of substantial amounts of material (i.e., the dimensional theories of performance), there would be some memory decrements over time for the material taught during training (Sulsky & Day, 1994). Memory loss, however, would perhaps be at least partly dependent upon initial reactions to training and the extent to which trainees actually applied what they have learned over the four month interval. Thus, it was predicted that for FOR and FOR + BOT trained participants at time 2:

> H3) both reactions to training at time 1 and self-report use of the FOR training content will predict scores on the time 2 learning measure of the FOR training content.

Lastly, we were interested in exploring whether significant differences exist between training groups on the learning measure completed four months following training. Specifically, we explored the possibility that

H4) FOR and BOT + FOR condition participants will yield significantly higher scores on the learning measure compared to controls.

METHOD

Participants

Participants for this study were 107 unpaid volunteers from the 1st Battalion of Princess Patricia's Canadian Light Infantry (1 PPCLI) stationed at Canadian Forces Base Calgary. 98% of the participants were males who ranged in rank from Master-Corporal to Captain, had served an average of 11 years in the military, possessed on average, six years of supervisory experience, and had written Performance Evaluation Reports (PERs) for an average of five years. Participants were randomly assigned to one of three experimental conditions; FOR training (n = 35), combined BOT and FOR training (n = 30), or the control condition (n = 32). An additional ten military supervisors from the same unit attended an initial pilot session so that the trainer could ensure that the FOR and FOR + BOT training sessions would be equivalent in length. To assess the time required to complete the various tests and questionnaires, supervisors in the pilot group also completed all of the dependent measures.

Stimulus Materials

The stimulus set consisted of 20 vignettes of infantry Corporals and Master-Corporals performing tasks typical of their occupation and rank level.

These vignettes were derived from critical incidents which were obtained from a focus group consisting of five senior Non-commissioned Members (NCMs) and two officers. Given the limited amount of time the unit could afford for such training, the researcher had the focus group generate incidents for only four of the performance dimensions contained in the 1996 version of the PER for NCMs. It was generally agreed by the group that these dimensions were ambiguous in meaning and, thus, would more easily result in different frames-of-reference when interpreted by supervisors. The dimensions were: "adaptability", "works on own", "military conduct" and "developing subordinates". As an example, for the dimension "developing subordinates", several types of situations where critical incidents occurred were identified by the focus group. These included: "when a supervisor is teaching new skills to subordinates in either a formal (i.e., classroom), or informal (i.e. field exercise) setting", " when a supervisor corrects a soldier because he/she is performing a task incorrectly", "when the supervisor is reviewing a soldier's work performance", and "when a supervisor is providing counselling for work-related or personal matters to a soldier". The theory of performance and critical incidents for each of the performance dimensions listed above is provided at Appendix A.

For each of the four performance dimensions, incidents exemplifying "satisfactory" performance at each level were differentiated from incidents representing "high" levels. Examples of behaviors at the "low" (A and B) levels were not generated because members of the focus group indicated that selection procedures and training courses typically eliminated individuals performing at the "low" level, and thus, these scores were rarely given to soldiers. In sum, theories of performance for the four dimensions were generated for satisfactory and high performance levels, and these theories formed the basis of FOR training.

Next, 20 vignettes were created to depict various performance effectiveness levels based on the critical incidents generated by the focus group. To enhance realism, vignettes related to winter warfare exercises conducted in Canada and peacekeeping tasks similar to those occurring in Bosnia, in addition to some routine tasks such as maintaining and repairing vehicles and weapons, instructing soldiers, and counselling subordinates. An infantry officer external to the unit and a senior NCM who was the Chief Instructor at the unit reviewed the vignettes and provided suggestions on how to further improve their credibility.

The vignettes were then videotaped using several infantry soldiers. In accordance with previous FOR research (cf. Sulsky & Day, 1992), two ratees were used for training practice. An additional three target ratees were used to assess rating accuracy. Each of the practice and target ratees demonstrated one behavior for each of the four dimensions.

Rating Scale and Comparison Scores

The military evaluation scale was used for the rating task. The scale contains nine performance dimensions (see Appendix B) and is a 7-point graphic-type rating scale containing "low" (A and B), "satisfactory" (C - E), and "high" (F and G) levels of performance (see Appendix C). This scale was used for the performance appraisal training programs and the rating task to facilitate transfer of training to the job.

Using a procedure recommended by Sulsky and Balzer (1988) and employed in other FOR training studies, (e.g., Sulsky & Day, 1992, 1994; Woehr, 1994), true or comparison scores were derived for the videotaped performances. The performance theories forming the basis of training were first discussed with the "experts" who were two infantry officers and an infantry senior NCM. Next, the experts provided ratings on all performance dimensions after examining each of the videotaped incidents. Following the suggestions of Smither et al. (1989), experts were provided with transcripts of the videotaped scenarios and allowed to take notes while the videotapes were being viewed. They were also provided with a copy of the performance theories while rating the videotaped vignettes. After all of the ratings had been assigned, the experts discussed rating differences, with the goal of arriving at a set of mutually agreeable comparison scores that were based on the performance theories. Initial agreement was based on ratings provided prior to the consensus meeting. The intraclass agreement index based upon the three experts equals .96. The comparison scores by performance dimension were as follows: Practice Ratees: E, C, F and C for Ratee 1; C, D, C and E for Ratee 2; Target ratees: C, F, D, and E for Ratee 3; F, E, C, and F for Ratee 4; and E, C, E, and C for Ratee 5.

Rater Training

FOR Training. Training sessions were conducted with groups of 5-9 participants. The same trainer was used throughout and sessions were just over 3 hours in duration. Initially, a one-hour lecture was provided on the purposes and types of performance appraisals, sources of performance appraisal error, and how these issues were linked to problems with the military performance appraisal system. Some actions that were being taken to rectify these problems were outlined.

The procedure for FOR training followed those developed by Pulakos (1984, 1986). Participants were told that they would evaluate the performance of NCM's on four separate dimensions of the PER. They were given the military rating scales and instructed to read them as the trainer read aloud the criteria defining each dimension and descriptions of various levels of performance effectiveness for the four dimensions. As the trainer explained the dimensions, participants were able to clarify meanings and ask questions. During this discussion, the trainer also gave numerous examples of ratee behaviors for the various performance levels on each dimension, (i.e. examples of behaviors representative of "satisfactory" performance were distinguished from behaviors representing "high" performance on that dimension). A common performance theory (i.e., frame of reference) was established among raters in the group for each dimension so that they could agree on which behaviors were relevant to which dimensions and the effectiveness levels of alternative behaviors.

Participants were then shown a videotaped vignette of a "practice ratee" and asked to evaluate the soldier using the written performance theories. Ratings were written on a whiteboard and discussed by the group. The trainer focused on a discussion of behaviors which were used in deciding on the assigned ratings and clarified any noted discrepancies among ratings. During this process, the trainer revealed the comparison scores and explained why the ratee should receive a certain rating on that dimension. This procedure was repeated with an additional "practice" ratee. Upon completion of training, participants viewed the videotapes of the three target ratees and rated them independently on the performance dimensions with the same scales used in training. They were also asked to complete the behavioral recognition measure, manipulation check and the reaction questionnaire. Participants were allowed to keep the written performance theory upon completion of training.

Combined BOT and FOR training. Sessions for the combined training program were also conducted with groups of 5-9 participants and were three hours in duration. The lecture given in the FOR training sessions was reduced to a short 10 minute introduction on the importance and uses of the PER in the military context, and the objective of the training that was to follow.

Initially, it was decided that a combination of techniques from the extant research literature would be included in devising the BOT portion of the training program. To this end, participants were advised of the importance of observational processes in the performance appraisal process, and several systematic errors of observation were described as outlined by Thornton and Zorich (1980) and Latham, Wexley and Pursell (1975). These included first impression, contrast, stereotype, similar-to-me and halo effects. Participants were also advised of the utility of keeping diaries on an ongoing basis, a technique used by Bernardin and Walter (1977). These would describe behaviors they observed throughout the year for each subordinate that corresponded to the dimensions introduced during training. The trainer then explained the criteria defining the four performance dimensions and descriptions of various performance effectiveness levels on the military rating scale, using the written performance theory, and provided examples for each dimension.

Participants were given the opportunity to practice with the diaries and observational techniques by rating the same practice vignettes used in FOR training, but they were advised to watch closely, take notes, and refer to the written performance theories while assessing the ratees, based on techniques taught during the lecture. Feedback on the accuracy of the raters' scores in relation to the comparison scores, and additional tips on note-taking were provided after each vignette. Trainees then rated the same three target ratees used to measure rating accuracy in FOR training. Finally, they were asked to complete the manipulation, behavioral accuracy and reaction measures. As in the FOR training condition, participants in the combined BOT + FOR training were allowed to keep the written performance theories upon completion of training.

Control Training. This session was also three hours in duration. Compared to the FOR and BOT only conditions, participants in this condition received a more lengthy introductory lecture on performance evaluations and appraisal research. Participants broke down into groups and were asked to brainstorm on uses of performance appraisals in the military context, and the problems they perceived were associated with military appraisals. A discussion of these issues followed, and the trainer provided an overview of a number of changes being introduced into the CF Performance Appraisal System (CFPAS) in the next couple of years which have been designed to rectify these problems. Participants were then given copies of the new PER, including brief written definitions of the performance dimensions and rating levels for the NCM PER provided by National Defence Headquarters, (NDHQ), (see Appendices B and C). These definitions would have been used by supervisors throughout the Canadian Forces in preparing the 1996 annual PERs. Participants reviewed both documents with the trainer, and provided feedback on the various performance dimensions and a variety of formatting issues; comments were later forwarded to a research sponsor at NDHQ. The control group participants then rated the two practice ratees used in the other training programs, but no feedback was provided regarding their accuracy. A short break was taken to prevent video fatigue. Then participants rated the same three target ratees used in FOR and BOT + FOR training. Finally, they completed the manipulation and behavioral accuracy measures.

Dependent Measures

Rating Accuracy. Rating accuracy was determined by comparing participants' ratings on the military performance evaluation scale to the comparison scores provided by the expert raters. Cronbach's (1955) four component indices were used to measure accuracy which included: a) Elevation (E), b) Differential Elevation (DE), (c) Stereotype Accuracy (SA), and d) Differential Accuracy (DA); (for the complete formulas of these accuracy indices, see Sulsky and Balzer, 1988).

Behavioral Accuracy. BA was computed (see below) from scores on a recognition measure. The 12-item measure (see Appendix D) developed for this study was modelled on the measure used by Sulsky and Day (1992). For this study, the measure was comprised of four behaviors actually occurring in the videotapes. Also included were four behaviors that did not occur but were each consistent with the performance of one of the ratees on one of the dimensions. Finally, there were four questions which did not occur during training and were not consistent with the performance of any of the ratees for any dimensions. Participants indicated (yes/no) whether or not they recognized each of these 12 incidents for specific ratees. Participants were not asked to indicate which ratee was involved in the incident.

BA was computed using the same procedure as Sulsky and Day (1992). In short, each participant's false alarm rate (saying a behavior occurred when it did not) was subtracted from the participant's hit rate (correctly identifying a behavior as occurring) to provide a BA score for each participant.

Trainee Reactions. Reactions to the various training programs were assessed by taking the composite score for each ratee on a 13-item reaction questionnaire which was developed for this study and modelled on one developed by Wexley and Latham (1991). Items tapped various aspects of reactions to training including responses to the training procedure, content, length of training and the trainer, (see Appendix E). The first seven items were based on a 5-point Likert-type scale while questions 8-12 asked participants to rate the various components and the length of the training program on a 3-point scale. Finally, question 13 asked the participants to rate the program from 1 (poor) to 5 (excellent). Coefficient alpha for the questionnaire in this sample (excluding questions 8-12 which were scaled differently) equals .85.

Learning Measure. The measure used to assess learning at time 2 (Appendix F) was an 8-item multiple choice test adapted from Sulsky and Day (1994). Four of the items required subjects to match specific behaviors to performance dimensions; the other four items involved assigning an effectiveness level to specific behaviors. A reliability analysis indicates that coefficient alpha for the scale in this sample equals .56. This low alpha should not be surprising given the heterogeneous nature of the measure and the small number of items (8).

Self-Report Usage. A self-report measure was used to evaluate the extent to which the information conveyed during the performance appraisal training was actually used following training. Participants in the FOR training received Form A, (Appendix G), while BOT + FOR participants received Form B

(Appendix H). Both forms contained questions concerning the frequency with which participants had used the performance standards and handout containing the performance theories while preparing their annual PERs. Participants in the BOT + FOR training condition were also asked how frequently they had used behavioral observation methods (avoiding errors of observation, note-taking techniques) while evaluating subordinates. Finally, participants were asked whether they had to refer to the written theory of performance for some of the four dimensions more often than for others. For all of the frequency ratings, a 5-point scale was employed from 1 (*never*) to 5 (*always*).

Procedure

Participants were advised at the outset of the study that the purpose of this study was to examine how people evaluate others in a military work setting and to provide them with additional insight and experience in rating the performance of their subordinates. They were then asked to complete a demographic survey which contained questions relating to their age, rank, gender, years of experience in the military and as military supervisors. After receiving either minimal control training, FOR, or the combination BOT + FOR program, participants viewed three target ratees and rated performance immediately afterwards by using the rating scales introduced during training. Just prior to scoring the target ratees, they were reminded that the vignettes only featured satisfactory and high performance levels on the four dimensions covered in training.

Next, participants were asked to complete the 8-item, multiple-choice measure designed as a manipulation check to determine whether FOR and

BOT + FOR-trained participants paid attention to the FOR material taught during training. Participants in the three training conditions completed the behavioral recognition measure, and those in the FOR and BOT + FOR training conditions also completed the questionnaire assessing reactions to their respective training sessions.

After a four month delay, a subset of participants from the FOR (n = 10) and BOT + FOR (n = 11) conditions returned and completed the learning measure. They also responded to the self-report measure, indicating the extent to which they had used the information provided in their respective training sessions when evaluating their subordinates and preparing performance reports. A subset of participants from the control condition (n = 12) were only administered the learning measure at time 2.

For ethical reasons, after time 2 measures were administered, participants were debriefed on the purpose of the study and given the opportunity to receive a copy of the final results. After the study was complete, participants were given a package explaining the purpose of BOT and FOR training, a copy of the performance theories, explanations of potential observational errors, and techniques for providing performance feedback to their subordinates. In addition, the trainer and an officer from the unit developed theories of performance for the remaining five dimensions on the military PER which were distributed to all supervisors in the unit after completion of the study.

RESULTS

The means and standard deviations for rating and behavioral accuracy, in addition to the manipulation, learning, reaction and behavioral outcome measures are reported in Table 2. The correlations among the study variables are reported in Table 3.

Manipulation Check for FOR Training Content

The manipulation check consisted of eight items which tested the knowledge of participants on FOR information presented during discussion of the practice ratees. The first four items required participants to match behaviors to their appropriate performance dimensions while the last four items had participants select the level of performance for certain behaviors (Appendix J). A one-way analysis of variance (ANOVA) with training group as the independent variable and manipulation scores as the dependent variable revealed a significant overall difference, $\underline{F}(2, 94) = 13.39$, $\underline{p} < .01$, $\text{eta}^2 = .22$. Results of a Tukey honestly significant difference test indicates that the participants in the FOR and BOT + FOR training programs obtained significantly ($\underline{p} < .05$) more correct responses ($\underline{M} = 5.69$, SD = 1.13 and $\underline{M} = 5.97$, SD = 0.93 respectively) in comparison with participants in the control condition ($\underline{M} = 4.72$, SD = 0.92). However, the difference between the FOR and BOT + FOR-trained groups was not significant, $\underline{p} > .05$.

Table 2

Means and Standard Deviations for Participants in Frame-of-Reference (FOR), FOR + Behavioral Observation Training (BOT), and Control Conditions

Dependent Variable	FOR	FOR + BOT	CONTROL
Elevation[a]			
M	.11	.14	.19
SD	.14	.21	.43
Differential elevation[a]			
M	.08	.14	.18
SD	.07	.11	.22
Stereotype accuracy[a]			
M	.31	.34	.54
SD	.26	.27	.44
Differential Accuracy[a]			
M	.42	.36	.77
SD	.42	.32	.50
Manipulation[b]			
M	5.69	5.97	4.72
SD	1.13	2.99	.92
Reaction[c]			
M	39.20	36.87	
SD	3.00	2.99	
Behavioral Accuracy[d]			
M	-.06	.03	-.17
SD	.37	.40	.45
Learn2[e]			
M	6.60	7.18	4.67
SD	.84	2.64	1.15
Frequency1[f]			
M	3.20	3.44	
SD	1.32	.53	
Frequency2[g]			
M	3.10	3.00	
SD	1.60	1.00	
OB[h]			
M		3.30	
SD		.67	

Note. [a]Low scores denote greater rating accuracy. [b]Total number of items responded to correctly on the manipulation measure (maximum = 8). [c]High values denote a more favourable reaction to training. [d]High scores denote greater behavioral accuracy. [e]Scored as the total number of correct responses on the learning measure assessed at time 2. (maximum = 8). [f]Frequency of use of the FOR performance standards assessed at time 2 (5-point scale). [g]Frequency of use of the performance theories handout assessed at time 2 (5-point scale). [h]Frequency of use of the behavioral observation techniques assessed at time 2 (5-point scale).

Table 3

Average Correlations Between Accuracy, Learning, Reaction and BA Scores

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
E[a]	-												
DE[a]	.30**	-											
SA[a]	.10	.29**	-										
DA[a]	.04	.16	.48**	-									
Manip[b]	-.07	-.17	-.14	-.22*	-								
React[c]	.24	-.20	.08	.30*	-.08	-							
Yrstot	.15	-.06	.07	.23*	.01	-.08	-						
Yrssup	.11	-.05	-.05	.08	.06	.01	-.11	-					
Yrsper	.07	-.02	-.03	.19	.05	-.09	.75**	.92**	-				
BA[d]	.19	.05	-.04	-.21*	-.04	-.01	.09	.10	.04	-			
Learn2[e]	-.04	-.30	-.22	-.13	.36*	.15	.24	.37	.31	.09	-		
Frequency1[f]	.15	-.40	.20	.19	-.19	.32	-.18	-.09	-.16	-.04	-.03	-	
Frequency2[g]	-.15	-.52*	.18	.35	-.16	.44*	.32	.30	.28	-.23	-.04	.49*	-

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

Note. [a] Low values denote greater rating accuracy.

[b] Scored as the total behaviors reported correctly on the manipulation measure.

[c] High values denote a more favourable reaction.

Yrstot = Number of years the individual has been in the military.

Yrssup = Number of years the individual has been a military supervisor.

Yrsper = Number of years the individual has been writing PERs.

[d] High values denote greater behavioral accuracy.

[e] Scores on the learning measure at time 2.

[f] Frequency of use of the performance standards assessed at time 2.

[g] Frequency of use of the theories of performance handout assessed at time 2.

Test of Hypotheses

To determine whether BOT + FOR-trained raters would produce the most accurate and control raters would produce the least accurate ratings (Hypothesis 1), a multivariate analysis of variance (MANOVA) was conducted among the three training conditions with the four Cronbach component measures as dependent variables. Results indicate a significant effect for training condition, $\underline{F}(8, 184) = 3.11$, $\underline{p} < .05$, Pillai's = .24. Note that the more conservative Pillai's criterion was used instead of Wilk's Lambda to evaluate multivariate significance given the unequal sample sizes across conditions. This test is recommended when sample sizes are unequal and violations of homogeneity may exist (Tabachnick & Fidell, 1996). A follow-up discriminant function analysis (DFA) revealed one significant eigenvalue, $\underline{p} < .01$, with training type accounting for 20% of the variance in the accuracy composite. DFA results also indicate that DA and DE contributed most significantly to the composite (structure coefficients were as follows: DA, .74; DE, .40; SA, .19 and E, .11). Partially consistent with Hypothesis 1, the group centroids for the first function suggest a discrimination between the control group (.70) and the FOR and BOT + FOR groups (centroids = -.39 and -.29, respectively) such that the control group evidenced the lowest levels of rating accuracy. However, centroid results indicate that the BOT + FOR groups were only marginally better than the FOR training groups on rating accuracy (see group centroids above).

Given that each accuracy index was potentially interesting, separate analyses of variance (ANOVAs) were conducted for the individual accuracy indexes. Levene's test for homogeneity of variance was conducted given the unequal sample sizes across conditions. For each of the accuracy scores, the test was significant ($p < .05$), indicating a violation of the homogeneity assumption. However, Tabachnick and Fidell (1996) indicate that there is only reason to be concerned about violations of homogeneity when significance tests become too liberal, and this will occur if the smallest group is associated with the largest variance. Given that the smallest group (BOT + FOR) did not produce the largest variances across dependent variables, concerns about violating the assumption are mitigated. Significant ($p < .01$) training effects were obtained for DA, $F(2, 94) = 9.09$, eta^2= .16; for DE, $F(2, 94) = 4.76$, eta^2= .09; and for SA, $F(2,94) = 4.39$, eta^2 = .09. Consistent with the multivariate result and Hypothesis 1, results of Tukey-Kramer honestly significant difference tests indicated that those participants in the control condition were significantly ($p < .05$) less accurate on DE, SA and DA than the FOR and BOT + FOR-trained participants (see Table 2). Similar to the multivariate results, the univariate analysis failed to yield support for the prediction that the highest levels of rating accuracy would be obtained in the BOT + FOR condition.

To test Hypothesis 2 predicting that participants receiving FOR + BOT training would yield higher levels of BA than participants in the other groups, a planned comparison was conducted, comparing the BOT + FOR condition on BA to the average BA of the other conditions. The decision to compute a planned comparison was predicated upon the idea that there were no expected differences between the FOR and control conditions on BA. Results of the planned comparison indicate that the BOT + FOR condition were significantly higher on BA ($M = .03$) compared to the other conditions ($M = -.12$), $t\,(93) = 4.7$, $p < .01$.

Hypothesis 3 predicted that both reactions to training at time 1 and use of the FOR training reported by the participants at time 2 would predict scores on the time 2 learning measure of the FOR training content. However, results indicate that initial reactions to the training and subsequent use of the performance standards, (frequency 1 item), were not significantly correlated (r = .15 and r = -.03 respectively, p > .05) with scores on the learning measure. In a review of twelve studies using various combinations of Kirkpatrick's criteria, Alliger and Janek (1989) failed to find any established relationships among reaction measures and the other three criteria. They suggested that reaction measures may simply be indicators of how much people enjoyed the course, and enjoyment may not necessarily result in learning.

Finally, it was hypothesized that participants in the FOR and BOT + FOR groups would yield significantly higher scores on the learning measure compared to control participants (hypothesis 4). Consistent with hypothesis 4, results of a planned comparison reveal that the FOR and BOT + FOR-trained groups were significantly higher on the learning measure (M = 6.89) than the controls (M = 4.67), t (30) = 3.5, p < .01.

Exploratory Analyses

Although not formally hypothesized, a number of additional exploratory questions were considered relating to (a) initial trainee reactions and, (b) specific performance dimensions. The following sections summarize the results of the various exploratory analyses conducted.

Analyses of trainee reactions. A two-tailed t-test comparing reaction to training for the FOR and BOT + FOR groups indicated a significant difference t(64) = 3.1, p < .01 between the groups (M = 39.2, and 36.9, respectively).

Means and standard deviations for the reaction questionnaire items are contained at Table 4. The highest scores were obtained for questions 3, "I acquired some useful information from this training program", question 6, "I would recommend this training program to other supervisors in the unit", and question 13, "Overall, how would you rate this training program?" Participants found the opportunity to receive feedback on their ratings and engage in a discussion of the performance standards with the trainer and other trainees to be the most valuable aspects of the program, while the lecture component in all three conditions was considered the least worthwhile part of the training (see Table 4).

Qualitative comments provided by the participants indicated that trainees were extremely in favour of the videos, but would have preferred longer vignettes for each ratee on which they could base their ratings. Typically, each vignette was two - three minutes in duration. They also indicated that background information on the individual would be useful prior to viewing each vignette, (i.e., types of qualifications that the soldier possesses which can vary and previous exposure to these types of situations/taskings) because these are important contextual issues that are taken into consideration when forming ratings on the performance dimensions. A large number of participants also indicated that they would have preferred that the training be extended so that all nine dimensions could be covered and more discussion allowed after the practice ratees. A number of participants indicated that they would have also liked performance standards and behavioral examples of "low behavior" which was contrary to the recommendations of the focus group. Finally, some participants provided feedback on question 4, indicating that they were not confident that they "could now rate their subordinates more accurately" because of existing systemic problems such as controls that are placed on the number of

"F" and "G" scores that are allocated. These high scores are controlled in an effort to deter rating inflation.

Individual performance dimension analyses. Based on anecdotal evidence obtained from the training sessions, it was determined that the performance theory for three of the performance dimensions (i.e., "works on own", "military conduct" and "adaptability") were relatively difficult to learn from the standpoint of correctly classifying behaviors to dimensions and assigning effectiveness levels to individual behaviors. Many of the trainees expressed difficulty in making assessment decisions because of the large number of criteria that defined these three dimensions. Thus, we decided to examine rating accuracy separately on a dimension by dimension basis. Unfortunately, however, such analyses do not permit computation of DA because only one dimension is considered at a time. Instead, an overall distance accuracy measure used in some previous FOR research (e.g., McIntyre et al., 1984; Sulsky & Day, 1992) was computed for each dimension. Results indicate there were significant differences in accuracy which parallel the original analyses for three of the four dimensions. The only dimension on which there were no significant differences on accuracy scores across groups was the "works on own" dimension, $\underline{F}(2, 94) = .35$, $\underline{p} > .05$.

Table 4

Means and Standard Deviations for Reaction Questionnaire Items

Item	Question	Mean	Standard Deviation
1.	Clarity of program objectives.	3.43	1.38
2.	Expectations for program.	3.31	1.78
3.	Some useful information acquired.	3.66	1.29
4.	Confident that I could now rate subordinates more accurately.	3.10	1.54
5.	Intend to use this information when assessing subordinates in the future.	3.23	1.69
6.	I would recommend this training program to other supervisors.	4.32	1.65
7.	Trainer was helpful.	2.98	1.96
8.	Usefulness of the videos.[a]	2.47	2.03
9.	Usefulness of the lecture.[a]	.97	.70
10.	Usefulness of practice ratings.[a]	2.11	1.90
11.	Usefulness of feedback/discussion.[a]	1.29	.46
12.	Satisfied with length of training.[b]	2.71	.61
13.	Overall rating of the program.	3.66	.73

Note. [a]Rated on a three-point scale. Higher scores indicate this component was considered more worthwhile.

[b]Rated on a 3-point scale. Higher scores indicate the length of the training was considered *just right*.

All other items are on a 5-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

DISCUSSION

The primary purpose of the present study was to compare the effects of FOR and BOT + FOR training on rating and behavioral accuracy in a field setting. In addition, we examined learning of the FOR training content and self-reported usage of the FOR and BOT + FOR training material after a four-month delay.

Some support was found for hypothesis 1 predicting that rating accuracy would be significantly higher for the FOR and BOT + FOR-trained participants compared to control participants. Moreover, as expected, the structure coeffficients from the DFA indicate that DA was the accuracy component that contributed most to the discrimination among groups. The univariate results indicate that FOR and BOT + FOR participants were superior in terms of DA, DE and SA, thus replicating findings by Pulakos (1986), and Sulsky and Day (1994). The effect sizes for DA, SA, and DE were .74, .19 and .41 respectively, which are comparable to previous FOR-training studies (Woehr & Huffcutt, 1994).

Although it was expected that DA would improve the most in terms of rating accuracy based on previous research, the results for DE and SA are not surprising. Previous research (Woehr, 1994) indicates that in addition to DA, DE should be improved by FOR training because it reflects distinctions in the overall performance of individual ratees - a distinction fostered by FOR training. Other researchers (e.g., Stamoulis & Hauenstein, 1993) have indicated that SA is one of the components most influenced by FOR training because it enables participants to develop stable, internal dimension standards that result in improved dimensional accuracy.

Support was not obtained for the prediction that there would be an incremental increase in rating accuracy provided by the addition of behavioral observation training material. This finding is mitigated by the fact that participants were required to rate video clips that were only a couple of minutes duration, thus, there may not have been sufficient opportunity to employ BOT techniques such as note-taking when assessing the ratees. In a sense then, the nature of the rating task in this study may have produced a "ceiling" effect in that FOR training was sufficient to improve rating accuracy compared to control training, so the addition of BOT techniques did not further enhance rating accuracy.

Although DA was improved by FOR and BOT + FOR training, the correlational analyses reveal that, paradoxically, initial reactions were significantly predictive of decreases in DA (r = .30, p < .05). One possible interpretation of this finding is that social desirability (cf. Crowne, Marlowe, 1964; Nunnally & Bernstein, 1994) may have infilitrated both responses to the reaction measure and the actual performance ratings. That is, some participants may have deemed it to be desirable to indicate that they perceived training favourably. Moreover, these same participants may have produced ratings in accord with their preconceived ideas about what constitutes socially desirable ratings (e.g., spreading out the ratings for each ratee).

A second interesting and unexpected result regarding DA is that supervisors who had served more time in the military demonstrated significantly lower levels of DA than their junior counterparts (r = .23, p < .05). Some research has indicated that older people may suffer a decline in cognitive ability which could then impact on the accurate completion of performance evaluations

(Salthouse, 1991). However, it is unlikely that these factors played a role in this study given that senior supervisors who had served upwards of twenty years in the military were only in their late thirties.

Judging from comments made during the training sessions, many of the senior supervisors perceived that a number of systemic problems limited the usefulness of a rater training program in the military context. In particular, written comments provided on the reaction measure cited such systemic factors as the limits placed on the number of high scores that can potentially be allotted to ratees which frequently prevent supervisors from providing accurate scores. Score controls were introduced several years ago in an effort to reduce rating inflation. They also commented on a number of political factors operating in the military which reduce accuracy such as the ratee's position and seniority in the unit, and the type of ratings the soldiers received the previous year. Finally, some of the more experienced supervisors expressed dissatisfaction with the fact that the rating forms and performance dimensions had been changed on a yearly basis for the past four years causing them to wonder whether the training would be of any use to them in the future.

In contrast, supervisors who had not yet written performance appraisals, or had only written appraisals for a year or two, appeared more eager to learn and participate in the training process. Many of these junior supervisors provided both verbal and written comments on the usefulness of the training in general, and in particular, the written performance theory and group discussion of the videotaped examples with the trainer and more senior personnel.

As noted by Banks and Murphy (1985), the raters' willingness to provide quality ratings is just as important as their ability to provide accurate ratings.

Given that the senior supervisors were generally more cynical concerning the utility of rater training, this may have diminished their motivation to use the performance theory which in turn adversely affected their accuracy scores.

Lastly, based upon comments made during the training sessions, some senior supervisors had their own firmly entrenched ideas of the criteria and levels of performance that should define each performance dimension, which were occasionally at odds with the performance levels established by the focus group or the experts providing the comparison scores for the practice ratees. Jones, Sulsky and Day (1994) found that disagreement with even a portion of the performance theory neutralized the beneficial effects of FOR training, presumably by diminishing motivation to use the theory appropriately. Given that seniority was not significantly correlated with scores on the learning measure, ($r = .24$, see Table 3), it appears that lower DA scores associated with greater seniority is probably not the result of a failure to learn the material taught during training. Rather, some of the more senior personnel may have chosen consciously or otherwise to ignore the training material and employ their own pre-existing performance standards when evaluating ratee performance.

The results support hypothesis 2 predicting that BOT + FOR-trained participants would be significantly higher in terms of BA compared to FOR and control participants. Given that there were no differences in rating accuracy between the two FOR-trained groups, this finding lends some support for the notion that the success of FOR-only training for improving rating accuracy may stem from the correct categorization of ratee performance (Sulsky & Day, 1992).

However, for the FOR-only group, BA was significantly correlated with DA ($r = -.29$, $p < .05$), suggesting that memory for ratee behavior was predictive of increases in rating accuracy as well. This pattern of results suggests that, consistent with data reported by Sulsky and Day (1994), FOR-only participants may have used a combination of stored ratee impressions and ratee behavior when forming their ratings.

Hypothesis 3 predicted that initial reactions to the training and self-reported use of the training would predict scores on the learning measure at time 2. Although the data do not support this hypothesis, reactions to training were correlated with use: more favourable reactions to the training predicted significantly greater use of the handout detailing the performance theories (frequency 2; r = .44, p < .05). In turn, handout use was significantly correlated with greater use of the performance standards while completing the PERs (frequency 1; r = .49, p < .05).

Interestingly, the manipulation check was significantly correlated with learning, (r = .36, see Table 3). Thus, paying attention to the FOR training material predicts performance on the learning measure completed after a four-month delay. This is encouraging for proponents of detailed training programs. After all, FOR and BOT + FOR participants had access to the performance theories (recall they left with a handout detailing the dimensional performance theories) and, evidently, paid some attention to the handout based upon the descriptive use data for the Frequency 2 item (see Table 2). Overall then, any ad-hoc learning of the FOR information following training was likely not sufficient; paying attention to the material taught at the time of training was apparently of significant incremental utility when completing the learning measure.

Finally, support was found for the fourth exploratory hypothesis: FOR and FOR + BOT participants scored significantly higher than the control participants on the learning measure at time 2. The learning measure required that raters know how performance would be manifested in new behavioral situations given that new scenarios were presented which had not been previously discussed. The fact that participants in the FOR and BOT + FOR-trained groups were able to successfully translate performance expectations to new situations indicates that these participants developed rules for observing and evaluating job behaviors in a variety of situations, rather than just being familiar with the behavioral examples provided in training. In addition, as just indicated, scores on the manipulation measure were significantly correlated with scores on the learning measure, suggesting that participants who paid greater attention to the FOR training evidenced greater knowledge of the FOR material in the long-term.

Exploratory Analyses

Reaction measure analyses. Considering that FOR participants reacted to the training more favorably than BOT + FOR-trained participants, it is possible that the addition of BOT material increased the complexity of the training and required participants to absorb a great deal more information. Nonetheless, for the BOT + FOR participants, this did not appear to have any deleterious consequences for any of the outcome variables. Future training efforts might examine the costs and benefits associated with a longer training protocol so the information is presented at a slower pace. Alternatively, dividing the BOT and FOR components into two separate protocols delivered at different times might overcome this problem as well.

The finding that videotapes were considered the most worthwhile component of the training program is not surprising given the television oriented society we live in. Videotapes provide graphic, vivid behavioral examples which, if made effectively, help instill the performance theory and provide more realism in terms of the vignettes. Practice in assigning ratings and subsequent discussion of the ratings were also considered important aspects of the program, while the lecture was generally considered negligible by most participants. This result lends support to the findings of Latham, Wexley and Pursell (1975) discussed previously.

Individual performance dimension analyses. Concerning the lower accuracy associated with the performance dimension of "works on own", it was evident from anecdotal evidence obtained during the training sessions that some trainees may have been confusing this dimension with "adaptability". The problem is that participants perceived these two performance dimensions to be somewhat interdependent. For example, to demonstrate adaptability (i.e., reacting effectively to changing circumstances), the soldier must be able to work on his own by being motivated to adjust, and by using his initiative to think of creative solutions for problems that may arise when adapting to the new situation. Thus, when viewing the vignettes, some participants incorrectly identified the "works on own" scenarios as ones depicting "adaptability".

Given that trainees reported that some dimensions were inherently more complex, we examined the descriptive data from the self-report usage measure which indicates that supervisors referred to the handout containing the theories of performance more frequently for the dimension "adaptability", (M = 3.44, SD =

.53), than for the dimensions of "works on own", (M = 3.00, SD = 1.00), "military conduct", (M = 3.20, SD = 2.70), or "developing subordinates", (M = 2.70, SD = 1.06).

Implications and Future Research

Although there were no significant differences between the FOR and BOT + FOR-trained groups on rating accuracy, BOT + FOR condition participants reported that they employed the BOT training material to some extent following training (M = 3.3, see Table 2). However, it is still unclear which components of BOT result in the greatest increases in rating and behavioral accuracy. Perhaps future research should focus on delineating the aspects of BOT that are most relevant to the tasks of observation and evaluation of performance. In addition, more detailed vignettes or role episodes might present a richer array of behavioral information that would potentially make the BOT intervention more useful for enhancing rating accuracy. The decision to use the shorter vignettes was based upon the advantages this afforded for the assessment of BA. Clearly, however, participants indicated on the reaction questionnaire that they would have preferred longer scenarios in which the various levels of performance on the dimensions were demonstrated. This is due to the fact that supervisors do not typically observe work-related behaviors in isolation, but in conjunction with other actions that tap into various performance dimensions.

Although FOR training was an effective means of improving rating accuracy in this field study, many organizations may hesitate to implement such a program because it is costly and time-consuming. Different means of delivering the FOR content (e.g., case studies with discussion) should be investigated to determine whether there is an equally efficient but less costly method of achieving a common frame-of-reference.

Stamoulis and Hauenstein (1993) also hypothesized that there may be conceptual and technical problems associated with generating target scores for practice vignettes. If this poses a problem to organizations, alternative protocols for delivering the FOR material should be examined. A study could be designed comparing the procedure used by most researchers such as Pulakos (1986) and Sulsky and Day (1994), where trainees practice rating and then receive feedback on their ratings in relation to true scores, with a program similar to the dimensional training discussed by Woehr and Huffcut (1994). For this latter type of training, true scores are not employed and there is simply a discussion of job behaviors and rating dimensions (accompanied by performance examples), in addition to practice and discussion of ratings.

If generating true scores is not at issue, but saving time is important, perhaps the traditional FOR training could be compared to a program in which the trainer “models” the appropriate way to rate the practice ratees. Modelling the correct way of evaluating and rating subordinates (without the use of videotaped performance episodes) may prove to be just as effective at enhancing rating accuracy, and less time-consuming/costly to the organization.

Finally, in typical performance appraisal scenarios, there is a time delay between training and ratee evaluation. At issue then, is whether the training material is retained in memory. In the present study, it was found that scores on the learning measure were significantly better for the FOR and BOT + FOR-trained participants, even after a four-month delay. However, future studies should also consider the effect of such a lengthy temporal delay on rating accuracy at time 2. A set of videotapes could be scored for rating and behavioral accuracy at time 1 and a second set of videotapes would reassess both forms of accuracy at time 2. To ensure that there was no confound introduced due to potential differences in the videotapes, the order of presentation of the videotapes could be counterbalanced.

Limitations of the Study

A number of potential limitations with the study should be noted. Although the results prove very promising for the use of BOT and FOR training in a field setting, military supervisors may be unique in that they are initimately familiar with job requirements because supervisors will have performed most jobs at each rank level. Training such a population and establishing a common frame-of-reference may thus be less difficult given the extensive job knowledge each of the supervisors already possesses. To ensure that the results obtained in this study generalize to the entire population, more field studies need to be conducted using a variety of organizations.

Second, the use of a small number of ratees and performance dimensions is a further limitation. Typically, a sergeant would be responsible for observing and evaluating at least 10 subordinates on nine performance dimensions, and more senior officers could be responsible for assessing up to 50 soldiers. Thus, the restricted stimulus set used in the study was less cognitively demanding than the actual rating tasks of military supervisors. In the future, these programs should be expanded to include all performance dimensions to determine whether increasing the number of dimensions and therefore, the complexity of the performance theories, still renders the programs effective.

As discussed previously, the rating task used in this study may not have adequately assessed the contributions of the BOT material to rating accuracy. Thus, the potential value of BOT training may have been underestimated.

Although a learning measure was administered at time 2 which indicated that FOR and BOT + FOR participants were able to apply the FOR training content to new behavioral scenarios, a parallel form of the learning measure could have been administered at time 1 to determine if participants in the above groups possessed greater knowledge of the FOR material immediately after training compared to controls. If FOR and BOT + FOR participants scored higher than controls on both learning measures, then it might be reasonable to infer that there was long-term memory retention of the FOR training content. Finally, the long-term effectiveness of these performance appraisal training programs needs to be more thoroughly investigated in terms of rating and behavioral accuracy by having participants rate another set of target ratees and complete a parallel form of the recognition measure several months after completion of the training.

In closing, this study examined FOR and BOT + FOR training programs in a field setting. In addition, various program evaluation criteria including learning of the training material, and reaction to and self-reported use of the training were assessed four months after the training (ie., time 2). Results were very promising in that FOR and BOT + FOR participants scored significantly higher than the controls on three of the four components of Cronbach's (1955) accuracy measures. Given that FOR has not been attempted previously in a field setting, these findings demonstrate the potential utility of a performance appraisal training program to organizations, thereby helping to lessen the gap between performance appraisal research and practice (Banks & Murphy, 1985). Future research needs to confirm these results and investigate alternative training strategies that will allow performance appraisal training programs to be employed in a variety of organizations.

The use of three of Kirkpatrick's criteria for program evaluation provided some interesting insights into the relationship of reactions to training with rating accuracy and subsequent use of the training material. Future studies should continue to incorporate these criteria. For example, self-reports of trainee reactions and use may prove useful for fine-tuning training programs. In addition to a learning measure, future studies should feature rating and behavioral accuracy measures at time 2 to further demonstrate the long-term effectiveness of performance appraisal training programs.

REFERENCES

Alliger, G. M., & Janek, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42, 331-342.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Level-of-processing theory and social facilitation theory perspectives. Journal of Applied Psychology, 72, 239-244.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. Personnel Psychology, 38, 335-345.

Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston: Kent Publishing.

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.

Bernardin, H. J., & Walter, C. S. Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Bittner, R. H. (1948). Developing an industrial merit rating procedure. Personnel Psychology, 1, 403-432.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human

Performance, 20, 238-252.

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

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. Organizational Behavior and Human Decision Processes, 40, 307-322.

Campbell, D. T. Systematic error on the part of human links in communication systems. Information and Control, 1958, 1, 334-369.

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

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavior measurements: theory of generalizability for scores and profiles. New York: Wiley.

Crowne, D. P., & Marlowe, D. (1964). The Approval Motive: Studies in Evaluative Dependence. New York: John Wiley & Sons.

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, 001-009.

DeNisi, A. S., Robbins, T., & Cafferty, T. P. Organization of information used for performance appraisals: role of diary-keeping. Journal of Applied Psychology, 74, 124-129.

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

Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. Research in Personnel and Human Resources Management, 4, 45-99.

Goldstein, I. L. (1993). Training in Organizations. Pacific Grove, California: Brooks/Cole Publishing Company.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. Psychological Review, 93, 256-268.

Hauenstein, N. M. A., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. Personnel Psychology, 42, 359-379.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. Journal of Applied Psychology, 73, 68-73.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. Research in Organizational Behavior, 5, 141-197.

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 64, 502-508.

Jones, M. C., Sulsky, L. M., & Day, D. V. Rater Motivation to Use Frame-of-Reference Training: Effects of Theory Agreement on Training Efficacy. Paper presented at the Ninth Annual Conference of the Society for Industrial and Organizational Psychology, Inc., Nashville, Tennessee.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. Journal of the American Society of Training Directors, 13, 3-9, 21-26.

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

Latham, G. P., & Saari, L. M. (1980). BOS, BES, and Baloney: Raising Kane with Bernardin. Personnel Psychology, 33, 815-821.

Latham, G. P., Skarlicki, D., Irvine, D., & Siegel, J. P. (1993). The Increasing Importance of Performance Appraisals to Employee Effectiveness in Organizational Settings in North America. International Review of Industrial and Organizational Psychology, 8, 87-131.

Latham, G.P., Wexley, K. N., & Pursell, F. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Lee, C. (1985). Increasing Performance Appraisal Effectiveness: Matching Task Types, Appraisal Process and Rater Training. Academy of Management Review, 10(2), 322-331.

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

Markus, H., & Wurf, E. (1987). The dynamic self-concept: A social psycholgocial perspective. Annual Review of Psychology, 38, 299-337.

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

Murphy, K. R., & Cleveland, J. N. (1991). Performance Appraisal: An Organizational Perspective. Englewood Cliffs, NJ: Prentice-Hall.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67, 562-567.

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. Journal of Applied Psychology, 71, 39-44.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74, 619-624.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Murphy, K. R., Philbin, T. A., & Adams, S. R. (1989). Effect of purpose of observation on accuracy of immediate and delayed performance ratings.

Organizational Behavior and Human Decision Processes, 43, 336-354.

Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric Theory (3rd ed.). New York: McGraw-Hill.

Padgett, M. Y., & Ilgen, D. R. (1989). The impact of ratee performance characteristics on rater cognitive processes and alternative measures of rater accuracy. Organizational Behavior and Human Decision Processes, 44, 232-260.

Pulakos, E. D. (1984). A Comparison of Rater Training Programs: Error Training and Accuracy Training. Journal of Applied Psychology, 1984, 69, 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, 78-91.

Smith, D. E. (1986). Training programs for performance appraisal: A review. Academy of Management Review, 1986, 11, 22-40.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Smither, J. W., Barry, S. R., & Reilly, R. R. (1989). An investigation of the validity of expert true score estimates in appraisal research. Journal of Applied Psychology, 74, 143-151.

Spool, M. D. (1978). Training programs for observers of behavior: A review.

Personnel Psychology, 31, 853-888.

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

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, 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, 1994, 79, 535-543.

Sulsky, L. M., Day, D. V., & Lawrence D. (1994, April). An examination of schema-development issues and frame-of-reference training: A possible boundary condition. In D. V. Day (Chair), Performance Schemas: Issues of development, similarity, and change. Symposium conducted at the meeting of the Society for Industrial and Organizational psychology, Nashville, TN.

Tabachnick, B. G., & Fidell, L. S. (1996). Using Multivariate Statistics. New York, NY: HarperCollins Publishers Inc.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. Journal of Applied Psychology, 65, 351-354.

Warmke, D. L., & Billings, R. S. (1979). A comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 64, 124-131.

Weick, K. E. (1968). Systematic observational methods. In G. Lindzey and E. Aronson (Eds.), The handbook of social psychology (Vol. 2). Reading, Mass.: Addison-Wesley Publishing Co.

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

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

Zalesny, M. D., & Highhouse, S. (1992). Accuracy in performance evaluations. Organizational Behavior and Human Decision Processes, 51, 22-50.

Appendix A

THEORIES OF PERFORMANCE

Adaptability

Criteria defining this assessment item: degree to which the soldier learns from experience, seeks advice, accepts constructive criticism, is flexible, and how he/she performs under mental/physical stress. The way in which the soldier adjusts to courses (which present new situations/information) is also important.

HIGH PERFORMANCE

Rating — Performance Examples

G — Adapts extremely easily to unusual circumstances; performs at an outstanding level, even when placed under extreme mental/physical stress; extremely flexible; fully accepts constructive criticism and learns from it to further improve performance.

F — Adapts quickly when thrown "fastballs"; high level of performance, even when placed under stress; when tasked, seeks advice and learns from experience to improve performance.

SATISFACTORY PERFORMANCE

E — Adapts very well to most changes; accepts constructive criticism; learns from experience; performs very well under physical/mental stress.

D — Adapts satisfactorily to changes; usually accepts constructive criticism; learns from experience; performs moderately well under physical/mental stress.

C — Limited adaptability in changing circumstances; derives minimal benefit from constructive criticism; sometimes becomes flustered in the face of mental/physical stress.

Works on Own

Criteria defining this assessment factor: initiative, motivation and reliability.

HIGH PERFORMANCE

Rating Performance Examples

G Soldier demonstrates a consistently high level of initiative; projects enthusiasm when given a tasking; volunteers suggestions/recommendations to improve taskings; can be fully relied upon to carry out tasks to the best of his ability, even in unusual situations; is always proactive as opposed to being reactive to situations/taskings.

F Soldier portrays a very positive attitude toward taskings; demonstrates superior initiative; can be relied upon to carry out tasks to the best of his ability.

SATISFACTORY PERFORMANCE

E Soldier demonstrates an above average level of interest and enthusiasm towards tasks; plans own workload without difficulty when tasked.

D Soldier usually demonstrates a satisfactory level of effort for the task; plans own work adequately when tasked; initiates suitable action with minimum delay.

C Soldier has the capability for initiative but it is not always used appropriately; demonstrates a variable level of interest in tasks; requires occasional checking to ensure tasks have been completed.

Military Conduct

Criteria defining this factor: loyalty, integity, dedication to the job/task/unit, adherence to CF regulations, dress/deportment.

Rating Performance Examples

G Soldier demonstrates a consistently high level of dedication to the job/task; extremely loyal to superiors, peers and subordinates; displays an outstanding level of dress/deportment.

F Soldier demonstrates a superior level of dedication to job; highly loyal to superiors, peers and subordinates; displays a superior level of dress/deportment.

SATISFACTORY PERFORMANCE

E Soldier displays a noteworthy amount of integrity and loyalty;demonstrates above average dedication to the job/task; very good dress/deportment.

D Average amount of dedication to the task/job; good level of dress/deportment; loyal to superiors, subordinates and peers; adheres to CF regulations; generally respected by others.

C Variable level of interest for the job or the tasks that he is assigned; sometimes has to be reminded about small details; sometimes slow to react to regulations and orders; acceptable level of dress/deportment.

Development of Subordinates

Criteria defining this assessment item: counsels, disciplines, develops skills, i.e., trains, assesses and provides feedback to subordinates.

Rating Performance Examples

HIGH PERFORMANCE

G Supervisor demonstrates a firm but supportive attitude towards soldiers; clearly communicates the soldier's strengths and weaknesses to him (accompanied by specific examples from notes); provides advice on how to improve performance (perhaps drawing from his own experience); generates enthusiasm in the soldier to perform even better in the future; demonstrates a high level of instructional ability; insists on work that is of a very high standard from subordinates.

F Same as above although the supervisor may not be able to motivate the soldier to the same degree as a supervisor at the G level, and may not draw on personal experiences to improve performance.

SATISFACTORY PERFORMANCE

E Feedback is organized and to the point; supervisor generally indicates the soldier's strengths/weaknesses and he does provide guidance on specific ways the soldier can improve his performance; good instructor, although his lesson plans are not quite as imaginative/polished as those at the F/G levels.

D Feedback is not comprehensive, but the supervisor does provide some examples of strengths/weaknesses; tone is supportive and not condenscending; he does allow subordinates to clarify any performance problems; average instructor; acceptable level of control exerted over subordinates.

C A minimum of information is provided to the soldier on where he performed well or poorly; supervisor is somewhat negative and though he still gets the point across.

Appendix B

PERFORMANCE DIMENSIONS FOR THE MILITARY NCM PER

N.B. AI = Assessment Item (typically described as a performance dimension in the performance appraisal training research).

AI 1 Application of knowledge and skills. Evidence of the effective application of job knowledge and skills is often provided by the quality of results obtained on the job. Care must be taken in assessing this item since poor results may, on occasion, be the result of factors beyond the member's control (e.g., member not trained). Problem solving is an important aspect of this item.

AI 2 Works on Own. Assess the member's ability to perform assigned duties effectively without the need for high levels of supervision. Initiative is perceived as a high level of "working on own".

AI 3 Adaptability. This item addresses the learning process as observed in the context of the job(s). This may be demonstrated by improvement of performance or not repeating mistakes. The numerous means of improving performance include; accepting constructive criticism, learning from experience, seeking advice and the completion of job-relevant courses or training.

AI 4 Team Work. This item addresses the member's ability to work effectively as a member of a team. Helping, assisting and cooperating are all important elements of this item.

AI 5 Military Conduct. Assess behavior and attributes which are valued by the CF and are often required by regulations. The item focuses upon conduct, loyalty, dedication and commitment to the CF, dress and deportment.

AI 6 Communications. Communication skills include both written and oral ability. It measures both the form (i.e., grammar and speaking skills) and the accuracy of the content of communications. This item also assesses whether the member passes information to others (e.g. keeps superiors informed) in a timely fashion.

LEADERSHIP FACTORS N.B. AI 7 and AI 9 are normally only for members who are assigned formal leadership responsibilities (e.g. instructors, section heads).

AI 7 Develops Subordinates. Assess how supervisors develop and maintain their relationships with their subordinates. This centres upon the supervisor's interpersonal skills, hence, the member's ability to effectively counsel, discipline and develop the skills of their subordinates is of great importance.

AI 8 Plans and Organizes. This item applies to the planning and organization of group work when the member is responsible for others. The item applies equally to the member's ability to plan and organize their own work, when the member has not been assigned subordinates. This item assesses the process of analyzing problems and identifying the resources and strategy required to accomplish the mission.

AI 9 Supervision. This item focuses upon the process of ensuring that the work of subordinates is completed accurately and on time. Monitoring, checking work, assigning tasks and delegation are important considerations.

Appendix C

RATING LEVELS FOR THE NCM PER

CATEGORY	DESCRIPTION	RATING LEVEL
HIGH	PROFESSIONAL ATTRIBUTES, PERFORMANCE LEVEL, OR DEGREE OF POTENTIAL FOR THE NEXT RANK IS WELL ABOVE THAT OF MOST NCMS IN THE SAME RANK.	G F
NORM	PROFESSIONAL ATTRIBUTES, PERFORMANCE LEVEL, OR DEGREE OF POTENTIAL FOR THE NEXT RANK IS COMPARABLE TO THAT OF MOST NCMS IN THE SAME RANK.	E D C
LOW	PROFESSIONAL ATTRIBUTES, PERFORMANCE LEVEL, OR DEGREE OF POTENTIAL FOR THE NEXT RANK IS WELL BELOW THAT OF MOST NCMS IN THE SAME RANK.	B A

APPENDIX D

BEHAVIORAL ACCURACY MEASURE

For the following scenarios, decide if MCpl Cooper (the first MCpl) actually engaged in this particular behavior. Use the following scale in deciding how certain you are that MCpl Cooper performed one of these behaviors.

1	2	3	4	5
Very likely did not occur	Probably did not occur	Possibly did occur	Probably did occur	Very likely did occur

1. When MCpl Cooper's Sgt asks him if he's finished the lesson plan for the field formation lectures, the Mcpl indicates that he was able to pull a lesson plan off the computer from last year.

Occur? Yes No Certainty Rating _____

2. During the post-exercise assessment, Mcpl Cooper commends the Cpl for demonstrating good initiative and performing well under stress.

Occur? Yes No Certainty Rating _____

3. MCpl Cooper indicates to one of his Cpls that he was pleased with the way he helped out the newer members of the sect during the last exercise.

Occur? Yes No Certainty Rating _____

4. When giving a post-exercise assessment to one of the Cpls in his sect, MCpl Cooper advises him of a couple of strengths and weaknesses but doesn't indicate how he can improve his performance.

Occur? Yes No Certainty Rating _____

5. When tasked by his Sgt to accompany some retired soldiers to a dinner for peacekeepers being held by the mayor of Calgary, Mcpl Cooper agrees that his sect will be willing to help out the veterans.

Occur? Yes No Certainty Rating _____

6. Mcpl Cooper asks the Sgt for some advice on safety precautions when they are discussing the display for Armed Forces day.

Occur? Yes No Certainty Rating ______

7. After MCpl Cooper has finished debriefing the Cpl on his performance during the last exercise, the Mcpl encourages him to keep up the good work, and indicates that he will talk to the pl comd to see if the Cpl can receive some form of recognition for his efforts.

Occur? Yes No Certainty Rating ______

8. When asked by his Sgt Lilly whether he is ready for the lecture tomorrow, MCpl Cooper indicates that he worked on the lesson plan all last evening.

Occur? Yes No Certainty Rating ______

9. When tasked with setting up a weapons display at Armed Forces day, MCpl Cooper indicates that the number/type of weapons were the important considerations for weapon displays when he was posted to the 2nd Bn.

Occur? Yes No Certainty Rating ______

10. MCpl Cooper advises the Cpl that he cannot continue to be a buddy to his peers when filling the role of 2IC of the sect.

Occur? Yes No Certainty Rating ______

11. When tasked with setting up a weapons display, MCpl Cooper complains that his sect is always getting these "dog and pony" shows lately.

Occur? Yes No Certainty Rating ______

12. MCpl Cooper indicates that he mactacked a set of patrol orders for each of the sect members so that they could carry them to the field in their butpacks.

Occur? Yes No Certainty Rating ______

Appendix E

REACTION MEASURE

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
1. The objective of this program was clear.	1	2	3	4	5
2. My expectations for this program were met.	1	2	3	4	5
3. I acquired some useful information from this training program.	1	2	3	4	5
4. I feel confident that I could now rate my subordinates more accurately.	1	2	3	4	5
5. I intend to use this information when assessing my subordinates in the future.	1	2	3	4	5
6. I would recommend this training program to other supervisors in the unit.	1	2	3	4	5
7. The trainer was helpful and contributed to the learning experience.	1	2	3	4	5

Please rate the relative value (1 = very valuable; 2 = worthwhile; 3 = negligible) of the following components of the training program to you:

8. Lecture _______
9. Videos _______

10. Practice in rating others _______
11. Feedback/discussion _______

12. The length of the training was (circle one):
(1) Too long (2) Too short (3) Just right

13. Overall, how would you rate this training program (circle one):
(1) Poor (2) Fair (3) Good (4) Very good (5) Excellent

14. What from this program was *most* valuable for you?

15. Please provide any additional comments, criticisms, or suggestions you might have for improving the program.

Appendix F

LEARNING MEASURE

Part A.
Directions: For each of the following, circle the letter associated with the assessment item for which the incident is relevant.

1. A MCpl is responsible for conducting a sect attack. After the task is completed, he decides to give the soldiers feedback on how the attack went and areas in which they can improve.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

2. A Cpl has been told that he is going to be part of the rear party for a peacekeeping mission. A couple of weeks before the unit is scheduled to leave for the mission, one of the other soldiers is injured and the Cpl is now told that he will be deploying with the unit. He responds by quickly sorting out his kit and his family affairs so that he can fill the vacancy in his pl.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

3. A Cpl arrives for roll call looking very tired. The sect comd asks the Cpl if he was out partying the night before and hadn't bothered to get his uniform ready for work. The Cpl tells his sect comd that he always makes sure his uniform is prepared but the sect comd knows for a fact that the Cpl was at a coy smoker until the wee hours of the morning. Note: Consider the actions of the Cpl in deciding which assessment item this scenario is depicting.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

4. Two soldiers have just returned to their bivouac area after a long day on a Bn exercise. One of the soldiers decides to clean his weapon before going to ground, while the other decides to leave it because he only fired one magazine that day and he is in dire need of some sleep before they go patrolling later that night.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

Part B.
Directions: For each of the following, circle the letter associated with the level of performance effectiveness for each incident.

5. A sect comd receives his orders for a fighting patrol and plans his patrol accordingly. When he submits his plans to higher, he is advised that the second lag of his route has been compromised. He quickly sets about changing his route so that he can adhere to his timings and carry out his mission successfully.

a. C
b. D
c. E
d. F/G

6. Several soldiers are cleaning their weapons in their pl room at F-16. One of the Ptes is questioning the leadership ability of his sect comd. The pl WO overhears the Pte's comments and tears a strip off the Pte for demonstrating a lack of loyalty towards his Sgt. Consider the actions of the WO when deciding what level of performance he is demonstrating.

a. C
b. D
c. E
d. F/G

7. A pl signaller is installing a 77 set and constructing an antenna for an AVGP. When the MCpl sees the signaller at work, he calls over the other members of his sect so that they can learn how to install a communications system. Note: Consider the actions of the Mcpl when deciding what level of performance this scenario is depicting.

a. C
b. D
c. E
d. F/G

8. A MCpl is counselling a Pte in his sect on performance problems. During the initial counselling session, the MCpl indicates the soldier's problems although he is vague on how the soldier can improve his performance in the future.

a. C
b. D
c. E
d. F/G

Appendix G

BEHAVIORAL OUTCOME MEASURE - FORM A

1. Did you use the performance standards taught during the PER training to help you while completing your 1996 annual PERs?

   Never _____ Rarely _____ Somewhat _____ Often ______ Always ______

2. How often did you consult the word picture handout during the PER period?

   Never _____ Rarely _____ Somewhat _____ Often ______ Always ______

3. Did you consult the word pictures for some of the assessment items covered in training (adaptability, works on own, military conduct, developing subordinates) more often than for other assessment items?

Assessment Item	Never	Rarely	Somewhat	Often	Always
Adaptability	_____	_____	_____	_____	_____
Works on Own	_____	_____	_____	_____	_____
Military Conduct	_____	_____	_____	_____	_____
Developing Subordinates	_____	_____	_____	_____	_____

Appendix H

BEHAVIORAL OUTCOME MEASURE - FORM B

1. Did you use the performance standards taught during the PER training to help you while completing your 1996 annual PERs?

   Never _____ Rarely _____ Somewhat _____ Often _____ Always _____

2. Have you used any of the information on note-taking and behavioral observation techniques taught during the PER training?

   Never _____ Rarely _____ Somewhat _____ Often _____ Always _____

3. How often did you consult the word picture handout during the PER period?

   Never _____ Rarely _____ Somewhat _____ Often _____ Always _____

4. Did you consult the word pictures for some of the assessment items covered in training (adaptability, works on own, military conduct, developing subordinates) more often than for other assessment items?

Assessment Item	Never	Rarely	Somewhat	Often	Always
Adaptability	_____	_____	_____	_____	_____
Works on Own	_____	_____	_____	_____	_____
Military Conduct	_____	_____	_____	_____	_____
Developing Subordinates	_____	_____	_____	_____	_____

Appendix J

MANIPULATION CHECK

Part A.

<u>Directions</u>: For each of the following, circle the letter associated with the assessment item for which the incident is relevant.

1. A soldier from Transport Pl is in the field driving his MLVW. When his packet stops for a 30 min halt, he performs a couple of basic maintenance checks and then rests until it's time to move.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

2. Two soldiers are in charge of a vehicle checkpoint in Bosnia. A car runs partly through the checkpoint and upon investigation, one of the soldiers discovers that the driver has experienced a heart attack. He immediately pulls the driver out of the car and begins performing CPR while instructing the other soldier to call for medical assistance.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

3. During a winter exercise in the Sarcee Training Area, a soldier reacts adversely to the cold and physical stress of the training by disregarding his sect comd's instructions and wandering over the ridge line during a reorg.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

4. Two Cpls are digging a snow defence while on exercise in the Sarcee Training Area. One of the Cpls is commenting on the exercise conditions and the sect 2IC.

a. Military Conduct b. Works on Own
c. Adaptability d. Developing Subordinates

Part B.
<u>Directions</u>: For each of the following, circle the letter associated with the level of performance effectiveness for each incident.

5. A MCpl provides performance counselling to a Pte in his sect who has been having problems. The MCpl advises the Pte of his shortfalls and warns him that he could be placed on C and P if his performance doesn't improve.

a. C
b. D
c. E
d. F/G

6. A Cpl is the pl comd's signaller for an infantry pl. During a road move, the pl stops for their first short halt, and the signaller hops out of his AVGP and proceeds to talk with the other drivers in his platoon to ensure their communication systems are operating smoothly. When he gets to one of the AVGP's, he discovers they have a problem with their radio so he offers to locate a new part and install it himself.

a. C
b. D
c. E
d. F/G

7. Two Cpls have been tasked to help organize some shelves in the CQ's storeroom. Cpl #1 notices several compasses lying on one of the shelves. He wants to steal one but Cpl #2 dissuades him, and when the CQ returns, Cpl #2 advises him that the compasses should probably be locked up because they are an attractive item.

a. C
b. D
c. E
d. F/G

8. A new MCpl is giving a lecture to his section on the GPMG. He is visibly nervous, and he has some difficulty organizing the material and using the audio-visual aids.

a. C
b. D
c. E
d. F/G