NOTE TO USERS

Page(s) not included in the original manuscript are unavailable from the author or university. The manuscript was microfilmed as received.

24

This reproduction is the best copy available.

UMI

UNIVERSITY OF CALGARY

Good, Better, Best: Incorporating Frame-of-Reference (FOR) Training in a Centralized Selection Board's Process for Selecting Prospective Candidates

By

Ronald Douglas Porter

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF PSYCHOLOGY

CALGARY, ALBERTA

JUNE, 1999

© Ronald D. Porter 1999

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-48035-6

Canada

ABSTRACT

Frame-of-reference (FOR) training has emerged as an effective training method for improving the rating accuracy of performance raters. FOR training calibrates raters, such that they agree upon ratee levels of work performance. Although research has consistently demonstrated the efficacy of FOR training for improving rating accuracy in the context of performance appraisal, this study has extended FOR research by examining a FOR training program designed to calibrate selection based assessments. This study also represents a field application of FOR training. The Canadian Forces Regular Officer Training Program 1998 selection board was given FOR training before it reviewed the over 863 applicant files for that year's competition. FOR training increased both participant's assessment accuracy and agreement compared to rater error training (RET). In addition, there was partial support for the hypothesis that the level-of-success of selected applicants at Basic Officer Training would correspond with their selection board assessment.

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude to Dr. Lorne Sulsky, my thesis advisor, for his patience, and the continued encouragement and guidance he provided throughout this endeavor. To my committee members (Drs. Daniel Skarlicki, Lary Mosley and Daphne Taras) whose comprehensive and insightful review of this document added incrementally to its overall quality. As well, I would like to express my appreciation to Debbie Pankratz, who unselfishly donated her time as a research assistant. Finally, I am especially indebted to Major Jim Uchiyama, whose active support and participation made this study possible.

TABLE OF CONTENTS

Approval Page ii

Abstract iii

Acknowledgments iv

Table of Contents v

List of Tables vii

List of Figures viii

INTRODUCTION 1

Frame-of-Reference (FOR) Training 2

Limitations of FOR Training Research 7

FOR Applied to the Selection Context 9

Selection Boards 12

The ROTP selection board procedure. 15

THE PRESENT STUDY 17

Hypotheses 18

METHOD 20

Participants 20

Stimulus Material 21

Focus Group Phase 22

Development of Target Scores 27

Rating Scales 28

Phase I Training 28

Phase II (FOR) Training 30

Phase III 33
Criterion Measures 35
Reaction measure. 35
Training Course success. 35
Rating accuracy. 36
Data Analysis 36
RESULTS 37
Additional Analyzes 42
DISCUSSION 44
Practical Implications 53
Practical Limitations 54
Future Research 57
REFERENCES 58
Appendix A: Officer Training Performance Factors 62
Appendix B: Assessment Factor Matrix 63
Appendix C: Phase I Military Potential (MP) Scores 64
Appendix D: Pre-1998 ROTP file Review Form 65
Appendix E: 1998 ROTP File Review Form 66
Appendix F: Phase II Military Potential (MP) Scores 67
Appendix G: Participant Background Questionnaire 68
Appendix H: Utility and Benefit Survey 70

LIST OF TABLES

Table	Title	Page
1	Categories and Assessment Factors for the ROTP 1998 File Review Form	24
2	Summary of Events in Experimental Phases	34
3	Distance Accuracy Scores	38
4	Correlations Between Phase I MP Ratings and BOTC Results	40
5	Correlations Between Phase II MP Ratings and BOTC Results	41

LIST OF FIGURES

Figure	Title	Page
1	Sample ROTP Word Pictures	27

Good, Better, Best: Incorporating Frame-of-Reference Training (FOR) in a Centralized Selection Board's Process for Selecting Prospective Candidates

INTRODUCTION

Frame-of-reference (FOR) training has emerged as an effective training method for improving the rating accuracy of performance raters (Bernardin & Buckley, 1981; Cardy & Keefe, 1994; Hauenstein & Foti, 1989; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Woehr & Huffcutt, 1994). FOR training attempts to calibrate performance raters such that they agree upon the effectiveness levels of ratee work performance and thus provide similar assessments of ratee performance. Although research has consistently demonstrated the efficacy of FOR training for improving rating accuracy in the context of performance appraisal, published research has yet to examine FOR training in the context of personnel selection. Because personnel selection decisions are often based upon the subjective assessment of interviewers, assessment center/work sample assessors etc., FOR training may also be beneficial for calibrating selection-based assessments.

The purpose of this study was to extend previous FOR research by examining a FOR training program designed to calibrate selection-based assessments. Specifically, I examined a FOR training program designed to give the Canadian Forces Personnel Selections Officers (PSOs) a common FOR for evaluating candidates for Officer training. The contribution of this

study is two-fold. First, as already indicated, I expand FOR training into the domain of personnel selection. Second, this study represents a field application of FOR training; the extant FOR training research has largely been laboratory-based (see below). In the following pages, I (a) review previous FOR training research, (b) describe the potential utility of FOR training in the selection context, (c) briefly outline the Canadian Forces (CF) selection process for the Regular Officer Training Program (ROTP) candidates, (d) propose a number of predictions regarding the effects of FOR training on assessor evaluations and selection validity, and (e) present the study methods, research findings and provide a discussion of the results.

Frame-of-Reference (FOR) Training

It has long been recognized that those who rate the performance of others within an organization are influenced by subjective factors, which ultimately affect their ability to provide accurate ratings (Muchinsky, 1996; Stone & Kendall, 1964; Ungerson, 1975). It was not until the 1980's that there was a shift from conventional training, that had aided raters in avoiding common cognitive errors including halo, central tendency and leniency, to more proactive rater accuracy procedures (Athey & McIntyre, 1987; Muchinsky, 1996). Landy and Farr (1980) determined that while traditional rater error training facilitated the learning of a new rating response set (that normally lowered the incidence of common rater errors), it coincidently lowered levels of rating accuracy, in some cases. For instance, if the intention of the rater

training was to eliminate a common rater error such as central tendency, but the centralized ratings were actually based on real attributes or ratee behaviours, then the effect of removing such errors also eliminated both true and error variance.

Bernardin and Buckley (1981) concluded that to improve rater reliability new emphasis had to be placed on rater training programs that would increase rater accuracy. They proposed that through the implementation of one (or a combination) of the following programs: diary keeping, FOR training and mastery-based training, rater effectiveness could be increased. Bernardin and Buckley concluded that both diary keeping and developing common frames-of-reference for raters had similar components. First among these (as in personnel selection) was that critical work behaviours had to be identified and raters had to be instructed in how to rate each behaviour relative to its effectiveness (in the context of those tasks that comprised a particular occupation). FOR training focuses raters on common frames-of-reference so they can similarly assess a worker's performance (McIntyre, Smith, & Hassett, 1984). More specifically, FOR consists of matching ratee behaviours to the appropriate performance dimension and precisely assessing the effectiveness levels of alternative ratee behaviour (Sulsky & Day, 1992).

Bernardin and Buckley (1981) originally proposed the use of FOR training for individuals who did not provide accurate ratings when compared to target scores (i.e., idiosyncratic raters). The goal was to bring those raters into

congruence with other raters within the organization by eliminating the idiosyncracies through FOR training. It was suggested that idiosyncratic raters could be identified before the implementation of FOR training (Hauenstein & Foti, 1989); however, a more recent study found that even after receiving FOR training, between eight to fifteen percent of the sample's ratings remained idiosyncratic (Sulsky & Day, 1992). Consequently, Sulsky and Day suggested that there may have been additional ability or motivational factors contributing to the participant's idiosyncratic behaviour.

Notwithstanding the fact that a minority of raters may remain idiosyncratic following training, FOR training research has found uniform support for the proposition that FOR training substantially improves rating accuracy (Athey & McIntyre, 1987; Bernardin & Buckley, 1981; Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr & Huffcutt, 1994). In a recent meta-analysis of four rater training programs, FOR training (effect size = .83) was found to be the single most effective training strategy (versus control or no training groups), for increasing rating accuracy (Woehr & Huffcutt, 1994). Although the effectiveness of FOR training has been well documented, almost all of the studies were conducted in laboratory settings. Consequently, the generalizability of FOR training to field settings is still uncertain.

Bernardin and Buckley (1981) noted that the effects of even a very effective rater training program appeared to dissipate over a relatively short period of time. Presumably, this is due, in part, to trainees forgetting the material imparted during the training. These findings may have partially contributed to a recent shift in the focus of FOR training research. This recent change has seen a departure from researchers examining the effects of FOR training from a purely procedural intervention to examining the cognitive effects FOR training has on the rater participants during training. There have been several studies, which have examined the cognitive effects of FOR training, focusing on the cognitive mechanisms underlying FOR training success (Athey & McIntyre, 1987; Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; Sulsky & Day, 1992, 1994).

One study, for example, found that raters who had received FOR training remembered more training content. Specifically, Athey and McIntyre (1987) hypothesized that the FOR training protocol requires trainees to elaborately encode the training information. Therefore, simply giving the FOR information without the entire protocol was expected to attenuate retention of the training material. The authors employed levels-of-processing theory (that describes retention of information as a function of the depth at which the information is processed) to explain that information requiring greater cognitive elaboration would be more effectively remembered than information that required less elaborate processing. Athey and McIntyre (1987) found support

for their prediction; however, levels of processing theory was not tested directly. Rather, a levels of processing explanation was offered ad-hoc to explain the finding that participants receiving the full protocol retained significantly more training information compared to participants receiving only the FOR information.

In addition, some studies have examined the notion that FOR training assists raters in properly categorizing ratee performance, thus facilitating the formation of correct impressions concerning ratee performance (Sulsky & Day, 1992, 1994). That is, FOR training was expected to provide new categorizations or schemas for relevant performance dimensions, replacing the existing mal-derived preconceptions a supervisor might have about performance dimensions and related/unrelated work behaviour(s).

Sulsky and Day (1994) concluded that FOR training was a viable method for improving rating accuracy and enhancing categorization accuracy. They stated that while raters may forget specific behaviours and rely on long term categorizations based on the established theories performance learned in FOR training, these categorizations contribute to rating accuracy. Consequently, when a delay occurs between the observed performance and the actual rating of the employee, rating accuracy will not necessarily be attenuated even though raters become increasingly unable to recall specific ratee behavioural information.

Although these categorizations should result in a generally more accurate overall assessment of ratee performance, it must also be recognized that categorization has a potential negative side effect, causing rater impressions to be potentially affected by preconceived stereotypes (Cardy & Keefe, 1994). Raters may, in the formation of the general impressions, be overly quick to decide whether they recognize the existence of a performance dimension or behaviour in a particular ratee (Sulsky & Day, 1992).

There is a second concern with this and other similar findings regarding impression formation (Cardy & Keefe, 1994; Hauenstein & Foti, 1989). If specific behaviours for performance indicators are not remembered at the time of a performance appraisal interview/feedback session, it could be quite difficult to provide employees with effective performance feedback. As well, if an assessment resulted in a redress or litigation the defensibility of an impression (as apposed to the recording of specific behaviours) would potentially put an organization in a tenuous legal position (Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; Sulsky & Day, 1994).

Limitations of FOR Training Research

In spite of the preponderance of evidence suggesting that FOR training is a viable and useful training approach, some limitations associated with the FOR training research should be addressed. First, as noted earlier, almost all of the FOR training studies (Athey & McIntyre, 1987; Bernardin & Buckley, 1981; Cardy & Keefe, 1994; Day & Sulsky, 1995; McIntyre, Smith, & Hassett,

1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr & Huffcutt, 1994)(the notable exception is Hauenstein and Foti (1989)), have been conducted in a laboratory setting with undergraduate students who possessed limited work experience and little or no managerial experience, threatening the generalizability of the studies. In addition, the generalizability of the studies is threatened by the small number of ratees and the limited number of performance dimensions used in the experimental designs. In particular, Sulsky and Day (1992, 1994) stated that with more performance dimensions, the theory of performance imparted during training would become more complex, and might detract from the participants' ability to learn the training program materials, thereby reducing the program's effectiveness. It would seem as though the generalizability is reduced to near zero if the assessment/performance dimensions do not reflect a relatively similar number, as one might expect to find in an average work environment. Further, if an adequate training package cannot be developed to encompass a realistic number of performance dimensions, then no matter how overwhelming the findings of studies in support of FOR training, they will be of no, or very limited, value to an organization (Hauenstein & Foti, 1989). Therefore, FOR training must, by necessity, move out of the laboratory and into applied settings. It is not enough that a training program simply enhance a rater's accuracy; the program and the subsequent learned behaviours that result from the training must improve organizational effectiveness and be legally

defensible. At this point in time, based on the articles reviewed, there appears to be no clear evidence that this has been accomplished.

In field settings, it would appear that if a FOR training program is to have a lasting positive effect then the program must be thoroughly embraced by the organization including superiors, supervisors and subordinates. To be successful the FOR training program would have to include not only commitment by an organization for the initial and follow-up training, but in advance (if possible) the culture of the organization would have to be evaluated to determine the level of acceptance for the new program (Schein, 1990). This would facilitate determining whether the change initiated by the new training program would be readily accepted or whether the training program would have to modified before its implementation. In addition, raters must accept the FOR (i.e., theory of performance) which forms the basis of the training. Otherwise, raters may not be motivated to adopt the FOR when appraising performance (Schleicher and Day, 1998).

FOR Applied to the Selection Context

To date all FOR training research has centered on performance appraisal; however, the area of personnel selection/placement provides an interesting and possibly fruitful forum for extending FOR training. I propose that the utilization of FOR training in the area of personnel selection is a useful line of inquiry.

It is not difficult to understand why published FOR training has not as yet been extended into the province of employee selection. Although personnel appraisal is normally made after repeated observations of an individual's performance; personnel selection assessments are generally conducted with only brief exposure to a prospective candidate, which may involve only a single contact between the applicant and the human resource person. Moreover, not all types personnel selection procedures would benefit from FOR training. Whenever there exists an objective scoring system (e.g., cognitive ability testing, structured personality inventories), the development of a common FOR is unnecessary. However, whenever the selection test in question involves some measure of subjective assessment by an assessor (e.g., interviews, assessing work samples or assessment center exercises), the development of a commonly held FOR for all assessors arguably would provide the same benefits as it does in the context of performance appraisal.

In fact, the popularity of patterned behavior descriptive interviews (Janz, 1982) and situational interviews (Latham, Saari, Pursell, & Campion, 1980) is partially due to the fact that interviewers are provided with a set of scale anchors for evaluating interviewee responses. Thus, interviewees are given a common FOR in the form of behavioral anchors. The provision of these anchors is helpful from the standpoint of standardizing - or making more reliable - interviewee assessments (Cook, 1993). However, published research applying FOR training to these types of interviews does not exist.

Consequently, there is no research demonstrating the benefits of a formal FOR training program, which would not only review the scale anchors, but would also go beyond the anchors to consider other possible interviewee responses beyond those anticipated by the scale anchors. That is, one of the hallmark's of FOR training is that the training considers a variety of ratee behaviours/responses which may or may not "match" the contents of the scale anchors. By considering a variety of behaviours/responses that ratees may exhibit, the training attempts to calibrate raters do that they agree on the performance levels for these behaviours/responses. In addition, the training teaches raters how to integrate information which may be inconsistent (e.g., information indicative of both high and low performance levels) to arrive at a single rating for a given performance dimension.

Overall, the inherent similarities between subjective assessments during personnel selection and the assessment of ongoing work performance by incumbents (i.e., performance appraisal) is evident by recognizing that the desired outcome for both selection and appraisal is to correctly identify an individual's current performance (in many cases based on past events) with at least some emphasis being placed on the assessment of the individual's future potential. For example, in a performance appraisal system a rater is normally expected to rate a ratee's work performance that s/he may have observed directly or indirectly, which could include not only an assessment of performance, but also feedback, recommendations for training, promotion

and/or transfer within the organization (Muchinsky, 1996). These latter aspects of the performance assessment are somewhat similar to that of the selection context, in that a position offered could include training, promotion and/or transfer within the organization.

Selection Boards

Personnel selection is of paramount importance to all organizations. Being able to attract, select and place prospective employees in the right place at the right time is a process wrought with complexities and methods that could by no means be considered fool-proof (Muchinsky, 1996; Stone & Kendall, 1964; Ungerson, 1975). This is especially true when some or all of the selection process is dependent upon the fallibility of subjective assessment (Ungerson, 1975).

Many large organizations in Canada including the Canadian Forces, police and fire departments, and universities, employ a selection process that requires the use of a board or panel to review applicant information on prospective applicants and render a selection decision based on that information. Subjective assessments are central to these selection systems.

In some cases these Selection Boards will interview candidates, in addition to conducting aptitude and/or intelligence testing, physical fitness and/or medical examination and a thorough file review of relevant applicant information (biographical data, letters of reference, etc). However, this is not always the case; for example, in the Canadian Forces (Canadian Forces

Recruiter's Handbook, 1997), and graduate student selection at some universities (Kline & Sulsky, 1995), Selection Boards are conducted without the benefit of a selection interview. The Selection Board procedure normally involves several personnel from within an organization who review relevant applicant information and offer them a position (in the case of the Canadian Forces) or an offer of acceptance (in the case of university graduate school programs). In this study, the context for the FOR training program was the Regular Officer Training Plan (ROTP) board of the Canadian Forces. In short, the goal of the training is to improve the board's ability to select qualified candidates for officer training.

For the purposes of this research it is important to know both the purpose of officer training and the centralized Selection Board, as well as, understand its process. The following background information has been extracted and adapted from the Canadian Forces Recruiters' handbook, 1997.

Simply put, the purpose of Regular Officer Training Plan (ROTP) is to develop selected applicants for full-time service in the Canadian Forces as career officers. The purpose of the ROTP selection process is to evaluate each applicant's potential to successfully integrate into the Canadian Forces and, more specifically, successfully complete the Basic Officer Training Course (BOTC) and initial military occupation (MOC) training.

The formal selection process begins with an initial assessment by recruiters. Specifically, the selection process begins when a candidate enters a recruiting

centre. The recruiters are responsible for compiling a file on each applicant. These files consist of a variety of information (academic transcripts, reserve/cadet information, letters-of-reference, and all necessary Canadian Forces documentation) and allow the centre to monitor the candidates' progress through the selection process. One of the final components of the selection process is a selection interview conducted by a Military Career Counsellor (recruiter). As part of this interview the recruiter is responsible for completing a Canadian Forces Applicant Assessment report. The written report provides a summary of all relevant applicant information and is the primary document for recording information relating to assessment and selection. The information within the recruiter's assessment report is culminated in the military potential (MP) rating (a single score from 1 - 9), which reflects the recruiter's prediction of a candidate's likelihood of successfully integrating into the Canadian Forces with at least average performance. To accomplish this, each recruiter weighs all of the information received during the course of the interview and the other components (e.g., transcripts, letters of reference, etc.) and assigns the MP rating.

Once the report has been completed, it is forwarded to the centralized Selection Board for final review and assessment. Because the centralized Selection Board is not able to interview each applicant personally, it relies upon the detailed information provided by the recruiters in the recruiter report narrative as its primary source of assessment information. Similar to the

recruiters, the board members must determine a military potential (MP) score. It must be stressed, however, that there is no attempt by the board members to use the recruiters' MP ratings as an anchor or to attempt to validate the initial recruiter generated MP ratings. Rather, the board determines a new MP rating - which may or may not be numerically close to the recruiter's initial rating.

In recent years, the Selection Board has come under criticism for the number of candidates selected who were unsuccessful at the entry level training course for officers in the Canadian Forces (Basic Officer Training Course (BOTC)). This course must be completed by applicants who are selected in that years competition. Traditionally, the Selection Board has measured the validity of their selection decisions by this criterion (successful completion of BOTC).

The ROTP Selection Board procedure. The Selection Board process, consists of several distinct elements. First, a number of officers (6 to 8) within the Canadian Forces are selected by the recruitment Personnel Selection Officer at the Canadian Forces Recruiting, Education and Training System Headquarters (CFRETS). Then, the Selection Board conducts an applicant file calibration process. The board selects a number (no more than 5) of applicant files from the current competition and reviews them until all board members reach a consensus on the files MP ratings. An applicant's file contains: a recruiter's written applicant assessment report, academic transcripts (for a minimum of three years), two standardized letters of reference, the

employment application blank, and past course reports from the reserves or cadets. However, it is the written recruiter report which is the focus for evaluating each file. The other information (e.g., reference letters) is consulted in the event that information contained in the report requires clarification/confirmation or to "break ties".

After consensus has been reached on the MP scores, all files are re-inserted with the remainder of the files for the current competition. Next, the board members, working in teams of two, review a file and assign a military potential (MP) rating. The MP rating is subjectively assigned by the raters and reflects the raters assessment of the applicant's ability to succeed on the Basic Officer Training Course (BOTC). This rating is based on the board member's assessment of critical requirements (e.g., para-military experience, leadership ability, and motivation), which they chose during file calibration and believe will be predictors of success. Upon completion of the board's file assessment all applicant files are rank ordered based on the assessed MP rating.

Finally, it is also important to note how the rated applicants are offered positions within the Canadian Forces. In any given year, available training positions within an occupation are determined by training space allocations for each occupation (normally based on a five year forecast), which can vary significantly from year-to-year. As a result, applicants who were assigned lower MP ratings are, sometimes (by virtue of having selected an open position), given offers of enrollment over competing applicants who received

higher MP ratings during the Selection Boards assessment. This contributes to an increase in MP score ranges for the selected sample, and thus may help attenuate the inevitable restriction of range that would be expected in the MP scores for selected applicants. However, it should also be noted that some screening of the files occurs during the initial stages of the selection process, when applicants are assessed in the recruiting centres. As a result, the lowest scoring applicants are "screened out" prior to the Selection Board. This has the effect of restricting the range of applicant scores, and thus would be expected to contribute to a range restriction in MP scores assessed by the board.

THE PRESENT STUDY

It was my intent to develop a frame-of-reference (FOR) training package for the Regular Officer Training Plan (ROTP) centralized Selection Board of the Canadian Forces, with the goal of improving the board members' ability to assess prospective applicant files and select those most likely to be successful at the Basic Officer Training Course (BOTC). As previously mentioned, while FOR training has been successful in improving rater evaluation accuracy, the majority of the supporting studies have been conducted in laboratory settings with inexperienced raters (i.e., undergraduates). In addition to this lack of applied support for FOR training, this concept has not been incorporated into any personnel selection models. This study was conducted in three phases. The first two phases were conducted over three days. These two phases

consisted of a within-subjects design, in which participants reviewed and assessed 44 randomly selected Regular Officer Training Plan (ROTP) applicant files from the 1998 year's competition, before and after they received FOR training. Phase I provided the participants with pre-Selection Board training that was consistent with that given for the previous year's (1997) Selection Boards before the files were reviewed. In phase II, the participants received FOR training and then re-assessed the same files. Finally, in phase III the participants reviewed and assessed all 1998 applicant files for the current year's competition (conducted over seven days), immediately following the conclusion of phases I and II.

Hypotheses

As already noted, FOR training should lead to improvements in rating accuracy. Therefore, employing a random sample of files from which target scores are generated, it is hypothesized that:

H_1) Rating accuracy will be significantly higher for phase II assessments compared to the phase I assessments.

An important goal of FOR training is to improve inter-rater agreement. Because FOR provides participants with the same frames-of-reference for assessing candidates, interrater variability should be significantly less following FOR, thus positively contributing to inter-rater agreement.

Therefore, the second hypothesis is that:

H_2) There will be greater interrater agreement associated with the phase II assessments compared to the phase I assessments.

As a result of the uniform success FOR training has demonstrated in terms of improving rater accuracy, it was expected that following FOR training, participants would be better able to select candidates who would be successful at BOTC. Therefore, it was hypothesized that:

H_3) As a result of FOR training, there will be a significant increase in the number of candidates who will successfully complete the Basic Officer Training Course (BOTC) compared with previous years.

It was expected that, because FOR training leads to more accurate rating, the board members would be better able to assign representative MP ratings to candidate files. As a result of the increased rating accuracy and because the MP rating is supposed to reflect the prospective candidate's ability to successfully complete the Basic Officer Training Course (BOTC), there should be a significant relationship between MP scores and BOTC training success. Accordingly, the following two hypotheses are forwarded:

H_4) For phase I and II applicants who are selected, the relationship between MP rating and BOTC success will be significantly higher when the post-training MP (phase II) scores are used as the predictor compared to when the pre-training (phase I) scores are used as the predictor.

H_5) For phase III selected applicants there will be a significant relationship between the military potential (MP) score assigned by the participants and BOTC success.

METHOD

Participants

For all phases of the study there were eight (8) participants. Six (6) participants were Canadian Forces Recruit Unit Personnel Selection Officers and the remaining two (2) were Military Career Counselors as designated by Canadian Forces Recruiting, Education and Training System headquarters. All participants volunteered to be members of the ROTP Selection Board. Fifty percent of the participant pool was comprised of females. Participants had served an average of 15 years (SD=6.6) in the organization, with an average of eight years (SD=3.7) as a Military Career Counsellor at a Canadian Forces Recruiting Centre. In addition, participants had written ($\bar{x}$=17.0, SD=4.6) recruiter reports for past competitions. Using a five point Likert-type scale (1(not very familiar) to 5 (very familiar)) the participants indicated that they were very familiar with the recruiter reports and its contents ($\bar{x}$=4.6, SD=.5). Consequently, they were quite familiar with the structure and content of the recruiter's assessment report.

Following the participants' final review and assessment of all applicant files from the current competition, offers were made to those who had received the highest Military Potential (MP) ratings and had indicated preferences for military occupations currently available. Once the offers were accepted by the selected applicants they attended the Basic Officer Training Course (BOTC) in

the July - August 1998 time-frame. The number of successful BOTC candidates from this course were then compared with BOTC course success rates from the previous four years (1997, 1996, 1995 and 1994).

Stimulus Materials

The stimulus materials for phases I and II of the experiment consisted of a stratified (by MP rating) random sample of 44 Regular Officer Training Plan (ROTP) applicant files from the 1998 competition. Because the files varied with respect to informational detail, those that were retained for the sample were based on the following criterion; that the files contained information relevant to all of the assessment dimensions identified by a focus group (see focus group below). As a result, any randomly selected file which failed to provide this level of detail was discarded and a new file was randomly selected. In addition, of these 44 files, a subset of 15 files was selected and rated by two subject matter experts, to establish "target scores" for these files (based on the FOR assessment categories identified earlier by the focus group). During each of the first two phases, the 44 files were rated by each of the eight participants.

For phase III, 863 (this number represented all applicant files for the 1998 competition) files were assessed by the participants. Although the recruiter's report was the primary document used by the Selection Board to make their assessments, each applicant file consisted of a standardized list of items in addition to the recruiter report; (a) General Classification Test 2 or

Canadian Forces Aptitude Test results (these tests provide a measure of intelligence); (b) two standardized letters of reference; (c) academic transcripts (a minimum of 3 years); and (d) the employment application blank. The aforementioned documents were only used by the Selection Board to confirm and/or clarify information contained in the recruiter report. Further, the information contained on the report served as the basis from which the military potential (MP) rating was formed. All other documentation on the applicant's file was used only to clarify or confirm information contained in the report.

Focus Group Phase

Given that the centralized Selection Board convenes only once a year it was necessary to hold a focus group (which consisted of 4 Recruit Unit Personnel Selection Officers, the Recruitment Personnel Selection Officer , and the Director of Personnel Policy) to identify the critical characteristics of the performance factors (Appendix A) used during BOTC, and their relation to the applicant information contained in the recruiter report. During BOTC, officer candidates are assessed on 11 separate performance factors (e.g., Accepted Responsibility and Duties, Applied Job Knowledge and Skills, Made Plans and Prepares, etc.) throughout their training. The focus group then examined the recruiter report form and its contents, as outlined in the Canadian Forces Recruiter's Handbook (1997). In accordance with the handbook the report is comprised of seven distinct categories. These categories are labeled by

convention; (a) General, (b) Education, (c) Work Experience, (d) Personal Circumstances, (e) Activities, (f) Motivation, and (g) Summary and Assessment. Assessors are required to examine these seven categories of the report, and evaluate performance on 18 distinct assessment factors (e.g., assumes responsibility, team-work, etc). Appendix B presents a matrix which specifies the assessment factors and indicates which sections of the CF 283 are likely to contain information pertinent to the individual assessment factors. The focus group determined that all of the 18 assessment factors were relevant, insofar as they map onto BOTC performance factors.

The next task for the focus group was to identify the key behaviours for each of the assessment factors and produce narrative examples for these behaviours. These key behaviours constituted the performance standards which would be used by the raters to assess performance levels for each applicant on each of the 18 assessment factors. To facilitate the development of a complete assessment matrix the focus group agreed to provide behaviourally based narrative statements (called "word pictures") that would correspond to both very low military potential (MP) ratings (MP 4) and very high MP ratings (MP 9) for each of the assessment factors as they pertained to each section or category in the recruiter report where that information might be located (e.g., for the assessment factor Assumes Responsibility, pertinent information might be found in the education, work experience, activities and

NOTE TO USERS

Page(s) not included in the original manuscript are unavailable from the author or university. The manuscript was microfilmed as received.

24

This reproduction is the best copy available.

UMI

Although the focus group developed "low" (MP 4) and "high" (MP 9) word pictures for each of the nine assessment factors it was necessary to also develop word pictures for the remaining scale points (i.e., 5 through 8). Word picture development for the remaining scale points was accomplished by conducting a qualitative analysis to identify appropriate MP sample narrative statements. Unlike the sample narrative statements derived by the focus group, these narrative statements were extracted from a stratified (by MP rating) random sample of recruiter reports (n=36) taken from the 1997 competition. Each of the reports were broken down, so that each individual statement within a given report could be extracted and placed into one of the assessment dimensions and assigned a rating level. It was these sample statements that formed the basis from which the word pictures for the MP "5" to "8" ratings were developed. Moreover, the information contained in these narrative statements was also used to revise and enhance the word picture development by the focus group (i.e., for MP ratings from "4" and "9"). Word pictures (behaviourally based descriptive statements) are an integral part of FOR training, as they provide the rating anchors that the assessors use to more accurately assess an applicant's file. In summary, several word pictures that were representative of each of the key behaviours relative to a specific assessment factor were developed for each of the three categories of the recruiter report.

What resulted from the word picture development was an immense, cumbersome and unwieldy three dimensional matrix that consisted of assessment factors, report categories and word pictures for each MP rating from four to nine. If unmodified this would have generated a extremely large number of word pictures that would have to be reviewed for any one file assessment. Consequently, it was decided to reduce the MP ratings that word pictures would be generated for from six to three (odd numbered MP ratings (5,7,and 9) were dropped). This facilitated the development of a manageable word picture document that would still enable the board members to accurately assess applicant files on the full range of MP ratings (from 4 to 9). For example, if a statement on an applicant's recruiter report was assessed as below a MP rating of 6, but still above a rating of 4, then it would be rated a MP 5. In addition, because an applicant's assessed MP rating is not based on any single statement, but rather a preponderance of the evidence, it makes sense to derive specific ratings by combining information and making inferences about specific levels of performance (cf. Smith & Kendall, 1984). Figure 1 provides a sample word picture for an individual assessment factor for the category leadership.

Figure 1

Sample ROTP Word Picture (Assessment Factor: Assumes Responsibility).

Recruiter Report Assessment Category: Leadership

4	6	8
- Occasional part-time work - No full-time work during the summer. - Places blame on others (supervisor) for poor performance or tasking assignments not completed - Limited involvement in school/community activities, or at least not in anything significant.	- Part-time employment over a consistent period of time - Some full or part-time experience (paid or volunteer) - Some volunteer experience (participate, assisted and/or organized) - Charged with responsibility within a small business (oversees the work of 1 or 2 staff)	- Full-time work over summers (at least 2) - Regular coaching/ instructing/volunteering/ official referee work - Long-term (3 or 4 years) or part-time work - Takes responsibility for weak performance and took corrective action - Seeks out responsibility supervisory role (more than 2 people)

Development of Target Scores

As noted above, fifteen files from the original random sample of 44 were selected and "target scores" were generated for these files. "Target score" development followed procedures recommended by Sulsky and Balzer (1988). The scores were obtained by conducting a thorough assessment of each file using the frames-of-reference (i.e., word pictures) developed for this study on each of the assessment factors. The assessment was carried out by two

subject matter experts (SME) who were both Personnel Selection Officers (PSO) currently serving in the Canadian Forces. They had an average of 9 years of experience as PSOs in the Canadian Forces. Overall, there was substantial agreement between them - there was initial agreement for 93 percent of the total number of "target scores" generated. Where there was initial disagreement, the two parties reached consensus such that a single set of "target scores" emerged for subsequent analysis.

Rating Scales

The Military Potential (MP) rating scale was utilized for the rating task because it was typically employed by the recruiters for candidate assessment (it consisted of 9 performance categories). The performance levels ranged from 1 "Substantially Below Average" to 9 "Substantially Above Average". Because the files were pre-screened by the individual recruiting centres, it was necessary to use a truncated version of the MP rating scale (four to nine). For the 44 files chosen for phase I and II, there was a wide range of scores assigned across the recruiting centres (MP ratings from 4 to 9). Consequently, the participants in the study were provided with a diverse group of applicants for assessment (as per the recruiters' evaluations).

Phase I Training

I conducted phase I training, and it was modeled so that it would be consistent with the pre-Selection Board training from previous years. This

training consisted of a one hour rater error training session that covered the most common types of rater errors, which could affect the board members' ratings (e.g., halo error). Then, the board conducted an applicant file calibration process. That is, the participants reviewed two randomly selected applicant files from the current competition and attained a consensus on the file MP ratings. As previously mentioned an applicant's file contained; a recruiter's written report, academic transcripts (for a minimum of three years), two standardized letters of reference, the employment application blank, and past course reports from the reserves or cadets. After consensus had been reached the files were then re-inserted with the others files for the competition. The participants then reviewed the sample of 44 files from the 1998 competition and assigned a military potential (MP) rating for each file. For the file rating process the participants worked in pairs, although they each recorded the assessed MP rating on separate assessment forms (similar to those used by previous Selection Boards, see Appendix C). Participants worked in pairs because that is the actual process for the Selection Board. Consistent with Selection Board procedures, each pair of board members discussed their ratings to reach agreement on each assessed file (which constituted a consensus meeting) before passing the file on to the next pair of participants. It should be noted that although the participants worked in pairs to arrive at a consensus MP score, it was the individual assessments provided by each participant (prior to the consensus meeting) which was the focus of analysis.

Phase II (FOR) Training

I conducted the FOR training, which took place over two days. The training comprised several components including: an introduction to FOR training (its history in performance appraisal), a review of previous research, how FOR training could enhance a selection process, and word picture development. Finally, the training session directly linked the recruiter's interview assessment factors with the word pictures for the full possible range of scores for each assessment factor across the three categories of the recruiter report.

The procedure for FOR training followed those developed by Pulakos (1984, 1986). Participants were informed that they would reevaluate 44 Regular Officer Training Plan (ROTP) applicant files on nine separate assessment factors. The assessment factors and accompanying word pictures were provided to each participant and they were instructed to read along with the trainer, as the performance dimensions and scale anchors were read aloud. Participants then had an opportunity to ask any questions they might have had about the dimensions and scale anchors (e.g., what information in the file was relevant to particular dimensions). This process took approximately two hours, after which the participants were instructed to review each of the word pictures to ensure that it was consistent with rating dimension and scale. Participant review of the word pictures was done as a "homework" assignment at the end of the day.

The next morning all word pictures were reviewed and edited where necessary. This revision process, while time consuming (four hours), ensured that from the perspective of the participants, the word pictures accurately reflected the performance categories and rating scales. Hence, this provided participants with the opportunity to modify the theory of performance developed by the focus group. The revisions were then consolidated onto a single word picture document. During the consolidation of the word pictures, the participants and I developed a file review form that enabled the board members to record and score pertinent information from an applicant's file (Appendix D). Unlike the previous rating forms, where a single whole MP rating was assessed, the revised review form was included four distinct sections. These sections reflected the primary categories identified by the focus group and confirmed by the participants (Leadership, Motivation, Activities). Also, the "General" category was included on the review form and was used only if adverse information (drug use, criminal activity, racist/sexist behaviour) was present in the applicant's file, which might affect the overall suitability of the applicant (e.g., membership in a racist organization would cause an otherwise suitable file not to be considered in the competition). The remaining three categories were weighted in accordance with their perceived importance in predicting success at BOTC (as specified by the focus group) to determine a total MP score.

Two "practice" file assessments were then conducted. The files used were from the 1997 ROTP competition. Each participant reviewed the file on his/her own, rating each critical assessment factor using the consolidated word pictures and revised file review form. The assessed factor scores were then recorded and any discrepancies between raters was discussed. The average assessment factor scores for each category were then used to derive the assessment category score (e.g., for "Leadership"). The averaged weighted category scores was the MP rating for that file. Participants were then given a second practice file and asked to evaluate the applicant and determine the appropriate military potential (MP) rating. Again, the aforementioned process was carried out and any questions or concerns discussed and resolved to the participants' satisfaction. The participants then reviewed and reassessed the 44 applicant files from the 1998 ROTP competition (Appendix F). An identical procedure described above was carried out whereby participants worked in pairs and ultimately derived a "consensus" rating for each file. Identical to phase I, however, it must be noted that it was each individual's rating which served as the focus for analysis (as opposed to the consensus ratings).

The participants were given two anonymous questionnaires to complete. The first (Appendix G) was a general demographic questionnaire that provided valuable information about each participant's employment history in the Canadian Forces, and more specifically as a recruiter. This enabled additional analyses to be conducted regarding the potential influence of rater experience

and gender. The second questionnaire was an attitude measure (Appendix H) administered on two separate occasions, immediately following the conclusion of FOR file assessment in phase II, and then again on completion of the 1998 ROTP competition Selection Board (phase III). This questionnaire provided some insight into whether the participants believed that FOR was useful and/or enabled them to more accurately and confidently assess the ROTP applicant files. Measures assessing perceptions of training effectiveness can provide important insight into both the participant's affective reactions and utility judgements regarding the training (Alliger, Tannenbaum, Bennett, Traver & Shotland, 1997; Goldstein, 1983; Wexley & Latham 1991). The questionnaire was administered twice to garner (a) trainee's initial reactions to the training, and (b) their reactions once they had the opportunity to perform the phase III rating task. Assessments taken immediately following the training could be inflated by the participant's enthusiasm for the trainers, working with old/new acquaintances and/or shared experiences (Wexley & Latham 1991). Thus, re-administering the questionnaire following phase III was done to mitigate against possible rating inflation.

Phase III

Phase III was comprised of the actual 1998 ROTP Selection Board. Consequently, for phase III all 1998 ROTP applicant files (n = 863) were assessed by the Selection Board participants utilizing the theory of

performance imparted by the FOR training in phase II, and employing standard Selection Board procedures (those procedures used in previous years). That is, board members worked in pairs, rating each file separately, comparing MP ratings, and arriving at a consensus score before assessing the next file. This consensus rating was used in determining applicant selection for the available military occupation openings and subsequent attendance at officer training course. In essence, this changed the unit of analysis from the individual rater to rater pairs, which allowed the results of the 1998 ROTP Selection Board to be compared with those of previous years. Table 2 summarizes the key components for each of the three research phases.

Table 2

Summary of Events in Experimental Phases

Phase	Event
I	Participants complete Background questionnaire. Participants receive RET. Participants assess 44 applicant files.
II	Participants receive and review "word pictures". "Word pictures" are amended as necessary (based on participant recommendations). Two practice files are assessed by participants. Participants assess the same 44 applicant files used in phase I. Participants complete Utility and Benefit questionnaire.
III	Participants assess all 1998 ROTP applicant files (n = 863). Participants complete Utility and Benefit questionnaire.

Criterion Measures

Reaction measure. To measure ratee reactions to the FOR training, a multi-item questionnaire was developed and administered immediately following the second rating of the sample of 44 applicant files, and then again after the completion of the ROTP Selection Board (n = 863 files). The questionnaire consisted of ten items, asking the participants to rate both the utility of FOR training as part of the selection process, and providing them with an opportunity to identify aspects of FOR training that they perceived as irrelevant or unnecessary. Seven of the items on the questionnaire used a 5 point Likert-type scale from 1 (Disagree) to 5 (Agree). The Cronbach alpha for these seven items at time 1 was $\alpha = .83$ and at time 2 was $\alpha = .77$. The remaining three items on the questionnaire were qualitative in nature and provided the participants an opportunity to comment on the FOR training and its perceived utility.

Training course success. Basic officer training course (BOTC) success was defined in two ways. First, success was defined as the proportion of candidates who achieve a passing grade in all aspects of BOTC training as recognized by their course report. Second, upon completion of the BOTC each candidate received an overall rating from a single instructor on a five point continuous scale, which reflected the candidate's level of success during the

course (ratings between three and five were considered a pass). This continuous score served as the second criterion for BOTC success.

Rating accuracy. To assess rating accuracy, an overall measure of distance accuracy was used, which examines the "closeness" of the rater derived MP ratings and the "target scores" (Sulsky & Balzer, 1988). Because rating accuracy was computed for only a single dimension of overall performance (i.e., the MP rating), Cronbach (1955) component accuracy scores could not be computed. Last, it is important to note that the operationalization of accuracy employed here is not consistent with the conceptual definition of psychometric accuracy. That is, psychometric accuracy requires the existence of true scores serving as the standard for comparison. From a psychometric perspective rating accuracy refers to how "close" a set of ratings are to the true score ratings. In this study, however, the standard for comparison was a set of rating derived by subject matter experts (i.e., the "target scores"). Thus, the term "rating accuracy" employed here refers to how close a rater's ratings are to the subject matter expert generated target ratings.

Data Analysis

One of the concerns with this type of study is restriction of range. This concern arises because applicants given a low personnel selection assessment

rating will, in all likelihood, not receive an offer of employment (Schmitt & Klimoski, 1991). In the context of this study, this implies a correction for restriction of range was necessary to estimate the "target" predictive validity of the MP ratings. This is because, in general, the applicants with relatively low MP scores were not selected for BOTC training. The only exception to this was the unusual circumstance where an applicant with a low MP score was selected over another applicant with a relatively higher MP score because of occupational openings (see discussion of this point above).

RESULTS

To test the first hypothesis, predicting that rating accuracy would be significantly higher for phase II assessments compared to the phase I assessments, a dependent-samples t-test was conducted, comparing the mean distance accuracy scores between phase I and phase II (computed across assessors). These accuracy scores were computed by employing the "target scores" generated from the random sample of 1998 ROTP competition files (n = 15). The analysis revealed a statistically significant difference in the mean accuracy of the participants between phase I and phase II (t (7) = 7.06, p < .01). Inspection of the means in Table 2 indicates a significant improvement in the rating accuracy of the participants post FOR training. These results support hypothesis 1.

Table 3

Distance Accuracy Scores
for Each Rater in Phase I and Phase II

Rater	Phase I	Phase II
1	.86	.52
2	.71	.42
3	.76	.37
4	.91	.40
5	.59	.34
6	.44	.34
7	.95	.41
8	.81	.41
Mean	.75	.40
SD	.17	.05

Note: Lower scores denote higher levels of rating accuracy.

To test the second hypothesis, predicting there will be greater interrater agreement associated with the phase II assessments compared to the phase I assessments, two intraclass correlations were computed. Recall that each of the 44 ROTP files were evaluated by each of the four pairs of participants. Accordingly, the intraclass correlation coefficients were computed to assess agreement across the eight assessors on their MP ratings for the 44 files. The first coefficient was computed following the phase I assessments and the second following the phase II assessments. For phase 1, the pre-FOR condition, there was a moderate level of interrater agreement (intraclass $r = .70$, $p < .01$). Following FOR training, however, interrater agreement increased (intraclass $r = .82$, $p < .01$). Testing the difference between these two agreement indices indicated that there was a statistically

significant increase in interrater agreement post-FOR training ($z = 3.98$, $p < .01$), supporting hypothesis two.

To test hypothesis three, predicting that the success rate of Regular Officer Training Plan (ROTP) candidates at the Basic Officer Training Course (BOTC) for the 1998 ROTP competition (those selected after board members had received FOR training) would be significantly higher than the previous four years (no FOR training) a chi square test was conducted. This analysis revealed that there was a significant difference (χ^2 (4) = 20.57, $p > .01$; (χ^2_{crit}) = 13.28) among the 1998 and previous four year BOTC success rates. However, visual inspection of the raw data revealed that both the 1998 (87%) and 1997 (91%) ROTP competition years produced an overall improvement in BOTC completion rates compared to previous years. Completion rates for 1994, 1995 and 1996 were 85%, 84% and 82% respectively. Therefore, hypothesis 3 was only partially supported.

Hypothesis 4 predicted that the relationship between MP rating and BOTC success for phase I and II selected applicants will be significantly higher when the post-training MP (phase II) scores are used as the predictor compared to when the pre-training (phase I) scores are used as the predictor. To test this hypothesis the correlation between MP rating and the continuous measure of BOTC success was computed. Again, it must be noted that a correction for a restriction of range was carried out before the correlations were computed (Schmitt & Klimoski, 1991) . This correction was necessary because the Selection Board, by selecting the best candidates from the

applicant file, culled out the "worst" files. Only 25 of the 44 ROTP files from the sample were selected to attend BOTC.

Table 4

Uncorrected and Corrected Correlations Between Participant Phase I MP Ratings and BOTC Results

	R1	R2	R3	R4	R5	R6	R7	R8	BOTC
R1	1.0								
R2	.74*	1.0							
R3	.85*	.82*	1.0						
R4	.72*	.55*	.70*	1.0					
R5	.76*	.72*	.79*	.85*	1.0				
R6	.81*	.78*	.83*	.66*	.75*	1.0			
R7	.70*	.60*	.71*	.92*	.87*	.67*	1.0		
R8	.68*	.63*	.71*	.90*	.85*	.70*	.87*	1.0	
BOTC	-.24	.16	.06	-.40*	-.29	-.12	-.31	-.36*	1.0
BOTC RRC	-.26	.18	.07	-.47	-.40	-.14	-.38	-.47	1.0

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

Notes: 1. R1 - R8 represent the individual assessors.

2. BOTC RRC - The BOTC correlations corrected for range restriction.

3. The continuous measure of BOTC success was used as the criterion of success.

The individual MP ratings were used in these analyses (as opposed to the "consensus" ratings) because the focus was on the effects of FOR training at the individual level in enhancing predictive validities. Thus, I computed these correlations separately for each individual assessor. As can be readily

identified by examining Table 3 there was no significant relationship between the board's initial MP rating (pre FOR training) of the candidates and their achieved results on BOTC, nor was this relationship any stronger for phase II MP scores (Table 4). Not surprisingly, a dependent-samples t-test comparing the correlations (pre versus post-FOR training) across participants (with the correlations transformed to z-scores) did not yield a statistically significant result ($t(7) = -.85$, $p > .05$).

Table 5

Uncorrected and Corrected Correlations Between Participant Phase II MP Ratings and BOTC Results

	R1	R2	R3	R4	R5	R6	R7	R8	BOTC
R1	1.0								
R2	.86*	1.0							
R3	.78*	.88*	1.0						
R4	.84*	.90*	.86*	1.0					
R5	.85*	.95*	.89*	.88*	1.0				
R6	.82*	.91*	.85*	.85*	.95*	1.0			
R7	.80*	.88*	.86*	.84*	.88*	.83*	1.0		
R8	.81*	.88*	.92*	.89*	.91*	.87*	.88*	1.0	
BOTC	-.07	-.10	.03	-.24	-.07	-.10	.12	-.13	1.0
BOTC RRC	-.18	-.11	-.04	-.25	-.09	-.13	.13	-.15	1.0

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

Notes: 1. R1 - R8 represent the individual assessors.

2. BOTC RRC - The BOTC correlations corrected for range restriction.

3. The continuous measure of BOTC success was used as the criterion of success.

The final hypothesis predicted that for phase III there would be a significant relationship between the military potential score assigned by the participants and the final results achieved by the selected ROTP candidates at BOTC (n = 341). A pearson-product moment correlation for 1998 ROTP selected candidates revealed a statistically significant relationship between the Selection Board assigned MP assessment and the candidate's level of achievement at BOTC ($r = .31$, $p < .01$). This correlation was corrected for a restriction of range (Schmitt & Klimoski, 1991) in the predictor measure , because only those candidates with the highest MP ratings for an available position were selected to attend BOTC (uncorrected $r = .28$, $p < .01$).

Additional Analyzes

The attitude towards training questionnaire was administered on two separate occasions during the study (immediately following the phase II and again after phase III) and provided the participants with an opportunity to provide their perceptions regarding the FOR training, content and design. The questionnaire used a five point Likert-type scale (1 - Disagree, 5 - Agree) for questions one to nine. The participants indicated that overall the FOR training did assist them in assessing ROTP applicant files (phase II, $\bar{x} = 3.63$, SD = .92; phase III, $\bar{x} = 3.88$, SD = .64). As well, the participants believed that FOR training had improved their rating accuracy (phase II, $\bar{x} = 3.63$, SD = .74; phase III, $\bar{x} = 3.88$, SD = .64). The participants also indicated that

improvements could be made in the word picture organization (phase II, $\bar{x}$ = 2.63, SD = 1.06; phase III, $\bar{x}$ = 2.63, SD = 1.3) and content (phase II, $\bar{x}$ = 2.75, SD = .7; phase III, $\bar{x}$ = 2.75, SD = .7). When commenting on the content of the word pictures the most common statement (n = 6) indicated that because the statements were extracted from the recruiter reports, they were too specific. This specificity made it more difficult to apply the word pictures to different situations, as a result the participants stated the word pictures should be written so that they are more generic (e.g., Leadership MP 6: She babysits 2 to 3 times per week. Change to: Is employed part-time, works 2 to 3 times per week). The last question asked participants to indicate their perceptions on the overall utility of FOR training in Selection Boards (on 10 point Likert-type scale (from 1- not Useful, to 10 - Very Useful)). The results indicate that overall the FOR training was perceived to be "very useful" in the Selection Board process (phase II, $\bar{x}$ = 8.63, SD = 1.06; phase III, $\bar{x}$ = 8.5, SD = 1.07). That is, they felt it was useful, insofar as it increased their ability to more accurately assess the ROTP applicant files. These results indicate that participant's perception of FOR training and its utility in the Selection Board process were positive.

Next, I computed the correlation between total scores on the reaction measure from the second administration (i.e., phase III) and rating accuracy for the "target score" files. Alliger, et.al. (1997) suggest that participants who view the training more positively should obtain greater benefit (in this case - more

accurate ratings) than those participants who view the training negatively. However, results indicated that there was a non-significant relationship between rating accuracy and perceived utility for FOR training ($r = .16$, $p > .05$).

DISCUSSION

The primary purpose of this study was to examine the effects of FOR training within the context of a centralized personnel Selection Board. This was accomplished in three ways. First, a within-subjects experimental design was employed to determine if FOR training would increase rater accuracy and interrater agreement. Second, a between years comparison was made using success at basic officer training as the criterion measure. Finally, the validity of the FOR training was examined, comparing the selected applicant files assigned MP scores (post-FOR training), and the level of success the applicant achieved ($n = 341$) at basic officer training.

Previous research has suggested that rater accuracy may be increased by providing raters with common frames-of-reference when making their assessments (Athey & McIntyre, 1987; Bernardin & Buckley, 1981; Cardy & Keefe, 1994; Day & Sulsky, 1995; Hauenstein & Foti, 1989; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day, 1992, 1994; Woehr & Huffcutt, 1994). In this study rater accuracy was examined using distance accuracy as described by Pulakos (1984) and Sulsky and Balzer (1988).

The first hypothesis, predicting that rating accuracy would be significantly higher for phase II ROTP file assessments, received support. Following FOR training the MP ratings assigned by participants showed a statistically significant improvement in distant accuracy compared to the MP ratings derived during phase I. There are several possible explanations for this improvement in rating accuracy.

Previous empirical research has suggested that several factors could account for the improved effectiveness of FOR training. First, Bernardin and Beatty (1981) stated that FOR training eliminates idiosyncracies that would otherwise cause ratings to be incongruent. The FOR training process employed in this study clearly provided the participants with common agreed upon standards for all assessed levels of performance for each of the critical dimensions. Because the participants were actively involved in the development of the "word pictures" and the rating form, one might infer that the participants developed a shared schemata (or shared impressions) that more accurately represented the various levels-of-performance which might be expected from the ROTP applicant files (Sulsky & Day, 1992, 1994).

In addition, participating in the development of the performance standards likely fostered an increased level of acceptance for the information imparted during training. This acceptance is a critical precursor to training effectiveness (Scheicher & Day, 1998). That is not to say that the FOR training categorically eliminated all rater idiosyncracies. As some studies have shown,

eight to fifteen percent of those who receive FOR training may remain idiosyncratic (Hauenstein & Foti, 1989; Sulsky & Day, 1992).

There may, however, be other motivational and/or ability factors that contribute to the improvement demonstrated by FOR trained raters (Sulsky & Day, 1992). These "other" factors may be linked to the level of participant involvement and duration of the FOR training that the raters experienced. This would be consistent with Athey and McIntyre (1987) who stated that levels-of-processing theory might explain why FOR trained raters could provide more accurate ratings. Athey and McIntyre found that the retention of information, in this case FOR training, requires greater cognitive elaboration and consequently the training material is more effectively remembered.

Given the findings of this study, it is impossible to state what factor or group of factors accounted for the improvement in file ratings. This is because participant levels-of-processing with regard to training material was not tested, and previous rater schemas of ROTP applicant performance were not measured. However, given that there was a substantial increase in the level-of-involvement for participants when they underwent the FOR training, and unanimous consensus was obtained on all "word pictures" used to evaluate the ROTP files, there is at least some anecdotal support for these two possible explanations of FOR training effectiveness.

In the additional analysis section above, I reported the results for the utility and benefit questionnaires completed by the participants, which clearly

indicated that they believed the FOR training had substantially increased their ability to more accurately rate the ROTP applicant files. Certainly, if this perception was correct, increases in rater ability would certainly help explain the improvement in rating accuracy post-FOR training.

Beyond rating accuracy, another indicator of FOR training effectiveness is interrater agreement. The data analysis from this study supported the second hypothesis that FOR training would increase interrater agreement. Comparing the participant assigned MP ratings for the sample files from phase I and II (FOR training), the analysis revealed a statistically significant increase in interrater agreement following FOR training. This is not surprising, given that improvements in rating accuracy were also found in the study. As one might expect, if all raters are being more accurate, then by extension there would be greater agreement between the raters on the assessed military potential (MP) scores for the assessed files.

One of the fundamental differences between this and previous FOR research is that it was conducted in a field setting with expert raters. Previous FOR research has almost entirely been conducted in laboratory settings with novice raters (people inexperienced in the job or personnel management), and only a limited number of assessment factors (two or three) and a similar number of performance levels (Athey & McIntyre, 1987; Bernardin & Buckley, 1981; Cardy & Keefe, 1994; Day & Sulsky, 1995; McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986; Stamoulis & Hauenstein, 1993; Sulsky & Day,

1992, 1994; Woehr & Huffcutt, 1994). Consequently, the overall external validity of the aforementioned studies can be questioned. This study, however, represents an extension of FOR training research because (a) it was conducted in a field setting with expert raters, and (b) incorporated six levels of performance over nine separated assessment factors into the rating task.

Assessing or selecting individuals based on numerous assessment factors and multiple levels-of-performance, and completed by someone with an intimate knowledge of the job and/or assessment/selecting procedures is more representative of what one might expect in a typical workplace environment. Although it must be recognized that a limited number of positions are filled using a Selection Board process, the within-subject phase of this study (ratings assessed by the individual participants) would also suggest that FOR training could be beneficial in organizations where hiring decisions are made by a single human resource person.

Another approach for determining FOR training effectiveness was to examine the completion rates of the selected ROTP candidates at the basic officer training course (BOTC) for the 1998 competition (i.e., those selected with FOR training) compared to completion rates for previous years (i.e., no FOR training). The analysis only partially supported the third hypothesis, predicting a significant increase in the number of ROTP candidates who would successfully complete BOTC compared to previous years. That is, although there was significant difference in the completion rates across the five years

span, the success rates were not significantly different between the 1998 (i.e., when FOR training was provided) and the 1997 selection years. There are at least two possible explanations for the lack of difference in completion rates between the last two years.

First, the Canadian Forces had developed a training programme for the 1997 Selection Board. This was done in an effort to counter average failure rates at BOTC of approximately 18 percent. The training consisted of four files being reviewed by the Selection Board members and arriving at a consensus on the MP ratings for each of the files before reviewing the applicant files for that year. That year (1997) the failure rate at BOTC for the selected applicants fell to nine percent. It was this initial success in reducing the BOTC failure rate that led the organization to recognize that a more formalized FOR training programme might benefit their Selection Board process. This, of course, led to the current study and development of the FOR training material for the Selection Board. It should be noted that the completion rate for the 1998 competition did rise by four percent from the 1997 selection year.

A second and more plausible explanation for the non-significant increase in the completion rates is that the actual BOTC training programme was modified for the 1998 selection year. Specifically, many of the training segments that had been conducted in the classroom were done as programed instruction (self-study) packages. This change in the BOTC training package likely had a major impact on the completion rate because it involved major

changes to the leadership portion of the training programme (which constitutes the largest and most important component of BOTC training). In addition, the changes to training proved not only difficult for the selected applicants, but the training staff also demonstrated difficulty in assessing candidates to new training standards. In summary, a number of participants may have dropped out of the BOTC training partly due to the greater demands placed upon them compared to previous years. Although this is admittedly speculative, it underscores the difficulty in making completion rate comparisons across years when the training programme itself is not static.

The fourth hypothesis predicted that the relationship between MP rating and BOTC success would be significantly higher when phase II (post-FOR training) MP rating were used as the predictor compared to the phase I (pre-FOR training) MP ratings. However, this hypothesis was not supported, using the individual assessors as the unit of analysis.

There were several factors which could account for the lack of support. First, although the correlations were corrected for range restriction, the small sample size of selected files (n = 25) from the original set of 44 applicant files raises concerns about the stability of the individual correlations. In addition, the statistical corrections could not account for the fact that all 44 files represented a restricted set of files based upon pre-screening at the recruiting centres. Therefore, the level of range restriction may have been prohibitive. Moreover, visual inspection of the BOTC scores and standard deviation for the

25 selected applicants provided by the instructors supports the idea that the levels of range restriction were substantial on the BOTC scores (SD = .69). Last, some applicants may have been selected based in part upon occupational selection; thus, it was not simply a matter of just selecting the applicants with the highest MP scores. This could have had the effect of attenuating validity. However, inspection of the MP scores for the selected files revealed that this was likely not an issue; rather, range restriction appears to be the plausible alternative.

The final hypothesis predicted that for all phase III selected applicants (n = 341) there would be a statistically significant relationship between MP score assigned by the participants and their level of achievement at BOTC (i.e., the global rating given at the end of BOTC training). A statistically significant correlation was found between the Selection Board MP assessments and the candidate's level of achievement at BOTC ($r = .31$, $p < .01$), thus supporting the fifth hypothesis. This indicates that the ROTP Selection Board MP ratings can be a useful predictor of BOTC instructor ratings.

Although statistically significant, the practical significance of the correlation is still modest. It is important to note that the military career counselors (MCCs) at the recruiting centres are given general training in how to write a recruiter reports; however, they are not provided with any formal guidance in what specific information is required by the Selection Board so accurate assessments can be made. That is to say, there are no common

frames-of-reference established between these two distinct assessment levels. As well, no formal studies have been conducted linking the content of the predictor domain with the criterion domain of the BOTC training programme. Until such time as recruiters receive FOR training and the Selection Board assessments are more formally linked to BOTC training, it might not be reasonable to expect more than low to moderate correlations between the Selection Board and the level of success achieved by the selected applicants.

One question immediately arises upon examination of the results for hypotheses four and five. Why was a statistically significant correlation obtained in phase III, yet the correlations for phase II failed to reach significance? There are several possible explanations for this apparent inconsistency in the data analyses. First, because phase III consisted of additional files (and a larger file set) the participants received considerable "practice" employing the "word pictures" during phase III. Thus, assessors rating skills may have increased over time. Second, following normal Selection Board procedure, participants worked in pairs and used consensus rating for each assessed file. These consensus MP ratings may have been more reliable than the individually derived MP scores from phase II. As well, there was a substantially larger number of ROTP files assessed in phase III, which may have provided greater stability, and a broader range of BOTC scores.

Last, the additional analyses revealed a non-statistically significant relationship between reactions toward the FOR training programme and rating

accuracy. Alliger, et.al. (1997) stated that training success should be positively correlated with the individual participant's perceptions of the training they received. Therefore, in this context, if the participants viewed the training as positive, their subsequent ROTP candidate MP assessments would be expected to be more accurate. However, when examining the raw data it is readily apparent that even those participants who viewed the FOR training more negatively did so only in relative terms. Therefore, in absolute terms the overall assessment of the FOR training by the participants was high to very high, with no scores below the mid-point of the 5 point scale. Even though some participants identified some shortcomings in the FOR training and materials (and made several valuable recommendations for any subsequent use of the training), they indicated that their perception of the training was very positive, and that it improved their ability to assess and select prospective ROTP candidates for employment in the Canadian Forces.

Practical Implications

The theoretical implications of the this study are fourfold. First, this study has demonstrated that FOR training can be successfully taken from the laboratory into a field setting with positive results. Second, raters are capable of using a more complex assessment matrix incorporating a higher number of performance dimensions (nine) and levels of performance (six) compared to what has been used in previous research. Third, expert raters can also benefit from FOR training. Past research has almost entirely used novice

(undergraduate) raters with limited or no management/assessment experience. Raters for this study were all experienced members of the Canadian Forces who had complete numerous personnel selection assessments. Last, but not least, this study suggests that in addition to performance appraisal, FOR training can be a valuable tool in the domain of personnel selection.

Practical Limitations

Because this study was conducted in a field setting, there were several threats to its internal validity. The study had to be conducted within the constraints determined by the organization. As a result, the total number of participants was limited to eight and the research method was primarily limited to a within-subjects experimental design. Consequently, a practice effect may have been experienced by the participants from phases I and II. This, of course, makes it more difficult to unambiguously interpret the results of the data analysis comparing phases I and II. As well, because the file rating process was very time consuming, the within-subject phases of the study were limited to assessing a total of 44 files. However, for phase III the total number of files assessed was substantially increased (n = 863), so that participants did have a reasonable file population with which to use the FOR training and "word pictures".

Another limitation was the fact that the information provided to the Selection Board was not standardized and recruiters may not have been consistent in the quality of the information included in the recruiter report. In

summary, they did not receive FOR training! To the extent that the quality of the information was compromised, this would attenuate the validity of the Selection Board assessments, whether or not the board members received training. Also, the links between the predictor and criterion (i.e., BOTC) domains were ill-specified, particularly given the changes which occurred to the substance of BOTC training itself. Without a clear mapping of predictor to the criterion, validity will be necessarily attenuated.

My role as trainer may have also influenced some of the research findings. Because I personally knew some of the Selection Board members it is possible that assessors reported "positive" reactions to the FOR training simply to appease the experimenter. However, responses to the reaction measures were anonymous, thus making social desirability effects less likely. A second issue is that I personally knew some of the focus group members. This may have allowed me to influence the development of the word pictures. However, every effort was made to ensure that my participation would be solely as a facilitator to the focus group and a trainer for phases I and II. Consequently, all FOR training material ("word pictures", and scoring sheet and protocol) were developed directly by the subject matter experts in the focus group, and modified by those who would assess the officer applicant files.

There are a couple of additional limitations to this experiment that have to be mentioned in the context of generalizability, including the applicant population, and the use of Selection Boards as a hiring process. First, it must

be recognized that Selection Boards are not commonly used in business human resource departments. It would not be cost effective for organizations to hold Selection Boards every time a junior/entry level position becomes vacant. Second, the applicant population was homogenous in that it was comprised almost entirely of people graduating high school in June of the competition year, they wanted to be employed in the Canadian Forces and obtain a university degree at a military college. The homogeneity of the applicant population is not unusual even outside a military context. When examining other organizations that use Selection Boards (i.e., graduate schools, police forces, fire fighters) they all appear to attract homogeneous applicant pools. So, while both of these limitations can negatively effect the generalizability of the findings of this experiment to other forms of selection practices, the findings may still generalize Selection Boards in other organizations, and certainly to other Selection Boards conducted in the Canadian Forces.

Notwithstanding the aforementioned limitations, it must be remembered that first and foremost this was an experiment, which was conducted in a field setting. Consequently, experimental controls did exist within constraints imposed by the organization and an independent variable was manipulated. Therefore, the findings of this study can still be interpreted with some confidence and the generalizability of the findings are enhanced inasmuch as the study was conducted in a field setting.

Future Research

This study found that FOR training could make a positive and significant impact in a personnel selection process. However, while rating accuracy and interrater agreement were both improved following FOR training, the role the FOR training may play in enhancing rating validity needs to be examined further. First, future research should standardize the recruiter report and examine the utility of FOR training for these initial assessors. Second, this study should be replicated with a greater number of files employed in phases I and II. This would increase the number of selected applicants in phase II and provide a stronger test of the contribution of the training for enhancing rating validity. Third, research should examine whether assessors employ the materials imparted during training in subsequent years and the extent to which assessors retain the information over time. Fourth, research needs to clearly link assessment factors examined by the Selection Board and particular components of BOTC training. By forming conceptual links between specific predictor and criterion components, the validity of the Selection Board assessments can be examined in a more integrative and theoretically driven manner - and this may in turn reveal which components of the applicant files best predict success in specific areas of BOTC training. Last, future field research in the area of FOR training should employ a between subjects design. This would allow researchers to counter any practice effects that may occur in a within-subjects design.

REFERENCES

Alliger, G.M., Tannenbaum, S.I., Bennett Jr., W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. Personnel Psychology, 50, 341-358.

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.

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

Canadian Forces Recruiter's Handbook (revised 1997). Ottawa, ON: The Department of National Defence.

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

Cook, M. (1993). Personnel Selection and Productivity. John Wiley & Sons: NY.

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

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

Goldstein, I.L. (1993). Training in Organizations (3rd ed.). Belmont, CA: Brooks/Cole Publishing Company.

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

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. Journal of Applied Psychology, 77, 3-14.

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

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

Latham, G.P., Saari, L.M., Pursell, E.D. and Campion, M.A. (1980). The situational interview. Journal of Applied Psychology, 65, 422-427.

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

Muchinsky, P.M. (1996). Psychology Applied to Work (5th ed.). Pacific Grove, CA: Brooks Cole.

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

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

Schmitt, N.W., & Klimoski, R.J. (1991). Research methods in human resource management. Cincinnati, OH: South-Western Publishing Co.

Schein, E.H. (1990). Organizational culture. American Psychologist, 45 109-118.

Schleicher, D.J., & Day, D.V. (1998). A cognitive evaluation of frame-of-reference training: Content and process issues. Organizational Behavior and Human Decision Processes, 73(1), 76-101.

Smith, P.C., & Kendall, L.M. (1964). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology , 47, 149-155.

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

Stone, C.H., & Kendall, W.E. (1964). Effective personnel selection procedures. Englewood Cliffs, New Jersey: Prentice-Hall, inc.

Sulsky, L.M., & Balzer, W.K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73(4), 497-506.

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

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

Ungerson, B. (1975). Introduction: The scientific method in selection. In B. Ungerson (Ed.), Recruitment Handbook (pp 3-10). Essex, Great Britain: Gower Press limited.

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

Wexley, K.N., & Latham, G.P. (1991). Developing and Training human Resources in Organizations. New York, NY: HarperCollins Publishers Inc.

Appendix A

Focus Group identified assessment factors (AFs).

Evaluation Report - Officers. The PFs are:

a. Assumes responsibility.

b. Plans and prepares.

c. Teamwork/cooperation.

d. Direct and organize.

e. Problem solving.

f. Fitness (physical ability).

g. Communication.

h. Technical ability.

i. Conformity to rules.

j. Accepts criticism.

k. Handle stress.

l. Integrity/values.

m. Action oriented/energy expended

n. Motivation towards organization.

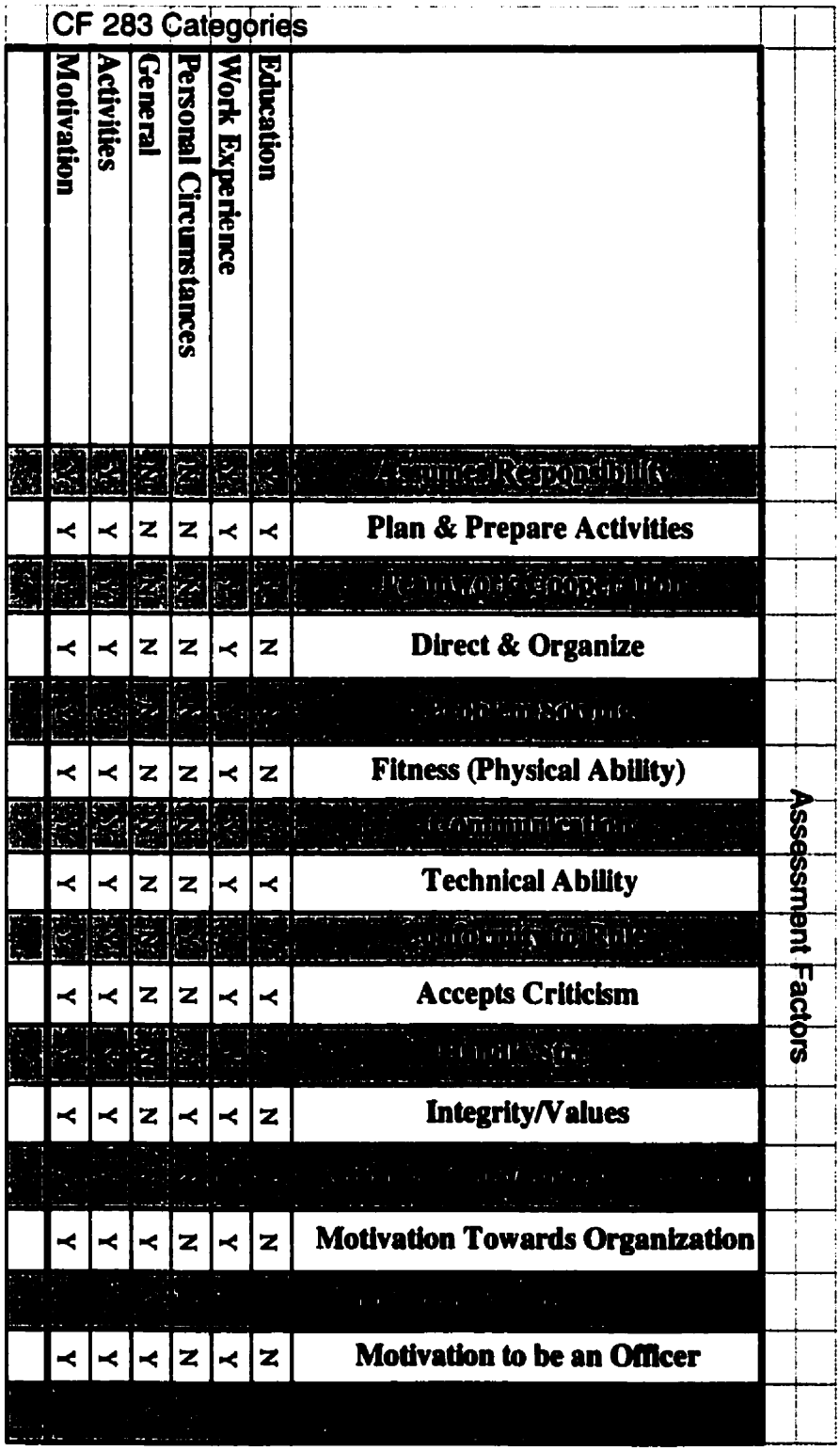
o. Motivation towards occupation.

p. Motivation to be an officer.

q. Perseverence.

r. Initiative.

Focus Group Assessment Factor Matrix



Appendix C

Participant MP Assessments for Phase I Applicant Files

File	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Selected
1	8.00	7.00	7.00	7.00	7.00	8.00	7.00	6.00	S(P)
2	4.00	3.00	4.00	5.00	5.00	5.00	4.00	4.00	
3	4.00	3.00	3.00	3.00	4.00	4.00	4.00	3.00	
4	5.00	4.00	5.00	6.00	4.00	5.00	5.00	5.00	
5	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	S(P)
6	7.00	5.00	5.00	8.00	6.00	7.00	6.00	7.00	S(P)
7	7.00	8.00	8.00	9.00	9.00	9.00	9.00	8.00	
8	6.00	5.00	6.00	5.00	5.00	6.00	6.00	5.00	
9	6.00	6.00	6.00	5.00	7.00	7.00	6.00	7.00	S(P)
10	6.00	6.00	5.00	6.00	6.00	6.00	6.00	5.00	S(P)
11	5.00	6.00	6.00	7.00	6.00	6.00	6.00	6.00	S(P)
12	5.00	4.00	4.00	5.00	6.00	5.00	5.00	4.00	
13	6.00	6.00	6.00	7.00	7.00	7.00	6.00	5.00	
14	6.00	4.00	5.00	7.00	6.00	7.00	7.00	7.00	
15	5.00	6.00	5.00	6.00	5.00	6.00	5.00	6.00	S(P)
16	7.00	6.00	7.00	8.00	7.00	7.00	8.00	6.00	S(P)
17	4.00	6.00	4.00	4.00	4.50	6.00	5.00	3.00	
18	7.00	8.00	7.00	8.00	7.00	7.00	8.00	7.00	S(P)
19	5.00	5.00	4.00	5.00	5.00	6.00	6.00	5.00	S(P)
20	6.00	7.00	6.00	5.00	7.00	8.00	6.00	6.00	S(P)
21	7.00	7.00	8.00	9.00	8.00	8.00	8.00	6.00	S(P)
22	5.00	7.00	6.00	5.00	6.00	7.00	5.00	6.00	S(P)
23	4.00	4.00	3.00	3.00	4.00	3.00	3.00	3.00	
24	6.00	6.00	7.00	7.00	7.00	7.00	6.00	6.00	
25	7.00	7.00	7.00	7.00	8.00	8.00	7.00	7.00	
26	7.00	6.00	7.00	9.00	8.00	8.00	9.00	7.00	S(P)
27	6.00	6.00	6.00	7.00	7.00	7.00	7.00	7.00	S(F)
28	5.00	6.00	4.00	4.00	5.00	5.00	4.00	4.00	S(P)
29	7.00	8.00	8.00	8.00	8.00	8.00	8.00	7.00	S(P)
30	4.00	4.00	4.00	5.00	4.00	5.00	5.00	4.00	
31	5.00	3.00	5.00	4.00	5.00	5.00	5.00	3.00	S(P)
32	6.00	5.00	6.00	6.00	6.00	7.00	5.00	6.00	
33	8.00	9.00	9.00	8.00	9.00	8.00	8.00	8.00	
34	5.00	3.00	3.00	5.00	4.00	5.00	5.00	3.00	
35	6.00	5.00	6.00	6.00	6.00	5.00	6.00	5.00	S(P)
36	6.00	6.00	6.00	6.00	5.00	7.00	6.00	6.00	
37	7.00	7.00	7.00	8.00	8.00	9.00	8.00	8.00	S(P)
38	6.00	6.00	7.00	8.00	6.00	7.00	8.00	7.00	S(P)
30	5.00	5.00	6.00	6.00	6.00	6.00	6.00	6.00	S(P)
40	6.00	7.00	7.00	6.00	7.00	7.00	8.00	6.00	S(P)
41	5.00	4.00	6.00	5.00	5.00	6.00	5.00	5.00	
42	4.00	4.00	4.00	4.00	5.00	6.00	4.00	4.00	S(P)
43	7.00	7.00	7.00	8.00	8.00	7.00	8.00	7.00	S(P)
44	4.00	4.00	4.00	4.00	5.00	5.00	4.00	4.00	

Note: The last column indicates only those applicants selected (S) for officer training, and whether they passed (P) or failed (F).

Appendix D

Pre-1998 ROTP File Rating Form

ROTP FILE REVIEW FORM (APR 98)

SN_____ NAME _____

Military/Para-military Experience

Cadets: type, years, rank, trg, staff
Reserves: type, years, rank, trg
Regular: type, years, rank, trg

Fitness

School Rep
School Intra-mural/recreational
Community sports
Fitness programme
Other sort participation

Organizational/Social Involvement

School Organizations (Student council, etc)
Youth Groups
Community Involvement
Other Organizations

Leadership

Team Capt
Executive Positions
Work supervisory roles (tutor, counsellor, mgr, etc)
Leadership courses/camps

Work

Part-time (number of hours, etc)
Summer employment

Motivation

Leadership knowledge
MOC knowledge
Rural Disadvantage
CF knowledge
Adjustment

Comments

Assessment: ____ Adjusted Assessment: ___Reviewer: _____

Appendix E

1998 ROTP File Rating Form

ROTP FILE REVIEW FORM (APR 98)

SN_____ APPLICANT NAME_____

LEADERSHIP _____ (Category Score)
Formula: Score Total / # of Dimensions (.45) = Category Score

Assessment Dimensions
Assumes Responsibility ____
Plans & Prepares ____
Directs & Organizes ____
Teamwork & Cooperation ____

ACTIVITY LEVEL _____ (Category Score)
Formula: Score Total / # of Dimensions (.35) = Category Score

Assessment Dimensions
Fitness (physical Ability) ____
Involvement in Activities ____

MOTIVATION _____ (Category Score)
Formula: Score Total / # of Dimensions (.2) = Category Score

**** NOTE: MUST SCORE AT LEAST 1 ASSESSMENT DIMENSION ****

Assessment Dimensions
Towards CF/CMC ____
Towards MOC ____
Officer Career ____

GENERAL

Rural Disadvantage
Conformity to Rules
Drugs
Racism
Criminal
Adjustment Difficulties

COMMENTS

OVERALL ASSESSMENT ____ **RATER IDENTIFICATION**

ADJUSTED ASSESSMENT ____ **(IF RECONCILIATION IS REQUIRED)**

Appendix F

Participant MP Assessments for Phase II Applicant Files

File	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Selected
1	7.80	7.80	7.40	7.80	7.80	7.90	8.50	7.60	S(P)
2	4.90	5.50	5.90	6.00	5.90	6.10	6.20	5.60	
3	4.30	4.50	4.60	4.90	4.80	4.60	5.60	4.70	
4	5.70	5.60	5.30	5.40	5.20	4.80	6.10	5.10	
5	6.40	6.30	6.70	6.80	6.10	6.80	7.00	6.00	S(P)
6	6.70	6.20	6.30	7.20	6.70	6.90	6.80	6.60	S(P)
7	7.90	8.80	7.40	8.10	8.80	8.70	8.30	8.20	
8	5.70	6.10	5.80	5.90	6.10	6.30	6.20	6.30	
9	5.90	7.20	6.60	7.20	7.20	6.90	7.00	6.00	S(P)
10	7.30	6.60	5.90	7.40	6.50	7.00	6.50	6.20	S(P)
11	6.90	6.40	6.80	7.00	6.20	7.00	7.10	6.90	S(P)
12	6.10	5.20	5.50	5.10	5.60	5.60	6.00	5.50	
13	6.40	6.90	5.90	7.00	6.60	7.40	7.00	5.90	
14	5.70	5.80	6.20	7.10	6.20	6.90	6.40	6.30	
15	5.90	6.20	6.00	6.20	6.60	6.70	7.00	6.40	S(P)
16	7.30	6.50	6.00	8.20	6.60	6.00	7.50	6.30	S(P)
17	5.60	5.60	7.20	6.60	6.50	6.70	6.40	7.00	
18	8.00	8.30	7.60	8.40	7.60	7.90	7.40	7.30	S(P)
19	5.50	6.50	6.70	7.00	6.60	6.60	6.50	6.00	S(P)
20	6.70	6.60	6.50	7.60	7.00	6.70	6.80	6.60	S(P)
21	7.10	7.00	6.20	7.50	7.40	7.60	7.00	7.00	S(P)
22	6.60	7.10	6.40	7.00	7.30	7.40	7.20	6.40	S(P)
23	5.00	4.60	4.40	5.70	4.20	4.30	5.50	4.80	
24	6.00	5.80	5.20	6.50	6.60	7.00	6.20	5.50	
25	7.60	7.60	7.20	7.50	7.90	8.20	7.40	7.20	
26	7.50	7.80	7.50	8.00	7.70	8.00	7.80	7.00	S(P)
27	6.30	7.70	6.60	7.90	7.00	7.70	7.20	7.00	S(F)
28	6.00	5.20	5.40	5.20	5.40	5.50	5.80	4.80	S(P)
29	6.40	6.90	6.60	7.30	7.10	7.40	7.00	6.90	S(P)
30	4.90	5.00	4.40	5.00	4.60	4.90	5.20	5.20	
32	5.10	4.70	5.00	4.70	4.90	4.60	5.40	4.50	S(P)
32	5.70	6.20	5.40	5.70	6.30	6.60	5.70	5.10	
33	8.20	8.30	7.70	8.50	8.30	8.80	8.10	8.00	
34	5.00	4.70	4.30	5.00	4.60	5.20	5.20	4.30	
35	5.90	5.60	5.40	5.60	5.80	6.20	6.10	4.80	S(P)
36	6.40	6.40	6.10	6.80	6.50	6.40	6.60	5.90	
37	7.20	8.00	7.40	8.00	7.70	7.60	8.50	7.40	S(P)
38	6.80	7.50	7.20	7.80	6.90	6.60	6.60	6.90	S(P)
39	5.70	5.60	5.30	5.40	5.90	6.40	7.00	5.30	S(P)
40	5.70	6.60	6.60	7.00	6.60	6.60	7.80	6.60	S(P)
41	6.00	5.80	6.40	6.40	6.20	6.00	6.90	6.30	
42	4.70	5.40	5.40	5.70	5.30	5.50	6.00	5.10	S(P)
43	7.20	8.10	7.30	7.60	7.60	7.40	7.90	7.20	S(P)
44	5.60	5.40	4.80	5.80	5.20	5.50	5.40	4.50	

Note: The last column indicates only those applicants selected (S) for officer training, and whether they passed (P) or failed (F).

Appendix G

Number __ Selection Boards and Frame-of-Reference Training
Participant Background Questionnaire

This participant background questionnaire will provide important information, which will facilitate the data analysis for this study. Answer the questions as accurately as possible, and if have any questions or are unsure of what is being requested please raise your hand and a researcher will assist you.

1. What is your current MOC? _____

2. Gender (circle one). M/F

3. What is the total number of years you have served in the CF? _____ (Count all Regular Force and Reserve time.)

4. What is the total number of years you have served as a PSO/MCC? ____

5. What is the total number of years you have served as a Base/Wing /Garrison/ Reserve PSO, in any? _____

6. What is the total number of years you have served in a CFRC (MCC/UPSO/ZPSA)? _____

7. Have you ever completed a CF 283 for a ROTP competition? (circle one) YES/NO

If yes (circle one) 1-5 6-10 11-15 16-20 >20

For the following please circle one .

8. How would you rate your familiarity with the form CF 283 and its composition for ROTP.

Not at All Familiar Somewhat Very Familiar
1----------2----------3----------4----------5

9. How would you describe your current understanding of FOR?

Poor Somewhat Excellent
1----------2----------3----------4----------5

10. How would you describe your current understanding of the ROTP Selection Board process?

Poor		Somewhat		Excellent
1	2	3	4	5

11. How would you describe your current understanding of ROTP relevant assessment factors?

Poor		Somewhat		Excellent
1	2	3	4	5

12. How would you describe your current understanding of ROTP assessment criteria at BOTC?

Poor		Somewhat		Excellent
1	2	3	4	5

Appendix H

Number ___ Selection Boards and Frame-of-Reference Training
Utility and Benefit Survey

This survey will provide important information regarding your perceptions of FOR and what benefits you believe it provided for you during the assessment of ROTP candidate files.

For the following please circle one.

1. Was the training logically organized?

Disagree		Neither Agree nor Disagree		Agree
1	2	3	4	5

2. Did you find that FOR assisted with the assessment of applicant files?

Disagree		Neither Agree nor Disagree		Agree
1	2	3	4	5

3. Do you think that your rating accuracy improved following FOR?

Disagree		Neither Agree nor Disagree		Agree
1	2	3	4	5

4. Were you able to process applicant files more quickly following FOR?

Disagree		Neither Agree nor Disagree		Agree
1	2	3	4	5

5. Were the word pictures organized in a logical and coherent manner?

Disagree		Neither Agree nor Disagree		Agree
1	2	3	4	5

6. Were the word pictures easy to comprehend?

Disagree		Neither Agree nor Disagree		Agree
1	2	3	4	5

7. Did you find it necessary to use the work picture throughout the entire board process? YES/NO

If no, at what point during the selection board did you stop?

8. If you were to be a members of the 1999 ROTP selection board, do you think FOR refresher training would be necessary? YES/NO

9. Are there other areas where you believe FOR might be of benefit? YES/NO If yes where?

10. On a scale from one to ten, how would you rate the overall utility of FOR in selection boards.

Not Useful Very Useful
1-----2-----3-----4-----5-----6-----7-----8-----9-----10