A CRITERION VALIDATION OF THE NEW ZEALAND ARMY OFFICER SELECTION BOARD

A thesis presented in partial fulfilment of the requirements for the degree of
Master of Science in Psychology at Massey University,
Palmerston North, New Zealand.

Kathryn Benjamin

2006

Abstract

The purpose of this research was to assess the incremental validity of selection measures employed on the New Zealand Army Officer Selection Board (OSB) over and above measures of cognitive ability. The study assessed whether the use of measures of personality, cognitive ability, peer assessment ratings, and observer competency gradings, could predict future training performance and job performance. Criterion measures of training and job performance included Officer Cadet School (OCS) performance results, supervisor ratings, and annual reporting documents. The sample population consisted of 72 New Zealand Army officers. Of these participants 15 were female and 57 were male. The average age of the participants was 27.5 years. It was hypothesised that individual elements of the assessment centre (observer ratings, psychologist ratings, and peer assessment ratings) would provide incremental validity over cognitive ability testing. It was also hypothesised that elements of the Eysenck Personality Questionnaire (EPQ-R) and the Gordon Personal Profile-Inventory (GPP-I) would be positively correlated with measures of training performance and job performance. Lastly, it was hypothesised that increased time since commissioning would be positively correlated with higher job performance. The results demonstrate that no linear combination of predictors was able to predict future training performance or job performance. Only the last hypothesis was supported and the results are discussed in light of methodological shortcomings.

Acknowledgements

I would like to acknowledge the invaluable assistance of the following people in producing this thesis:

My Supervisor, Dr Fiona Alpass for her knowledge, support, and unfaltering calm in the face of research hurdles. I would like to acknowledge the assistance provided in data collection and statistical analysis. Thank you for your time, effort and honest advice.

My Co-Supervisor, Major Helen Horn (Senior Psychologist Army) for her help and support in gaining access to resources, her commitment to completing this research, and her never-ending optimism and patience.

Lieutenant Colonel Paul King (Commandant of the Officer Cadet School) and his staff, for their willingness and assistance in gaining access to records and information.

Nicole Frost, for your patience and honesty whilst proof-reading my work.

My friend Carolyn Freeman, who motivated, empathised and laughed along with me on our journey to thesis completion.

Finally, I would like to thank the employees of the New Zealand Army who participated in this research. Your involvement is greatly appreciated.

Contents

Appendices

Statistical Appendices

List of Figures

Figures... ... ...

List of Tables

Glossary of Abbreviations

AC	Assessment Centre
ADFA	Australian Defence Force Academy
AIB	Admiralty Interview Board
APS	Army Psychology Service
B90	ACER Advanced Test B90
BARS	Behavioural Anchored Rating Scales
BP	Board President
CSSB	British Civil Service Selection Board
DBP	Deputy Board President
DCA	Deputy Chief of Army
DV	Dependant Variable
EFL	Entry Fitness Level
EPQ-R	Eysenck Personality Questionnaire Revised
FOR	Frame of Reference
FOSB	Final Officer Selection Board (Navy)
GMA	General Mental Ability
GPP-I	Gordon Personality Profile Inventory
IOT	Initial Officer Training
IV	Independent Variable
KASO's	Knowledge, Abilities, Skills, Other attributes
LO	Liaison Officer
MS	Military Secretary
MTO	Military Testing Officer
NCO	Non-Commissioned Officer
NZCC	New Zealand Commissioning Course
OAR	Overall Assessment Rating
OCS (NZ)	Officer Cadet School
OIC	Officer In Charge
OSB	Officer Selection Board
RATEL	Radio Telephone Procedure
Raven's APM	Raven's Advanced Progressive Matrices

RCB	Regular Commissions Board
RFL	Required Fitness Level
RMAS	Royal Military Academy Sandhurst
RMAS YO	Royal Military Academy Sandhurst Young Officer
RNZAF	Royal New Zealand Air Force
RNZN	Royal New Zealand Navy
TCS	Training Criterion Score
TEWT	Tactical Exercise Without Troops
TFMS	Territorial Force Military Secretary
TTCP	The Technical Cooperation Program
WOSB	War Office Selection Board
YO	Young Officer

Chapter 1
Introduction

1.1 Overview of the Issues

The aim of this research is to analyse the procedures used in the New Zealand Army Officer Selection Board (OSB) with a view to assessing the current criterion validity of selection decisions. The intent is to assess the incremental validity of the measures employed in the OSB over and above that of cognitive ability testing. To assess criterion validity, candidate performance data collected by the Army Psychology Service (APS) during the OSB, will be compared against candidate training performance at the New Zealand Officer Cadet School (OCS (NZ)) and job performance as indicated by annual personnel reports and immediate supervisor ratings.

The unique working environment offered by the military means the emphasis placed on ensuring a good fit between the person and organisation is more salient. Due to the specificity of the Army Officer Job, successful OSB candidates are required to attend a rigorous training course in order to familiarise themselves with the customs, traditions and discipline of the military and to train and equip them with the leadership and management skills necessary to effectively perform their job. The focus of the OSB is to select candidates capable of passing this training who have the potential to develop into sound military leaders. It is the responsibility of the OCS (NZ) to train the candidates to the necessary standard for commissioning (Carston, 2002). In accordance with this reasoning, this study has used two levels of criterion; performance at OCS (NZ), and performance as a junior officer.

The selection of competent and effective officers is vital to the military profession (Hughes, 2000). The officer fulfils the essential role of command. Officers within the New Zealand Army are responsible for the leadership and management of soldiers within their command. The position is one of great responsibility and comes with both moral and ethical obligations. Curran (2000) states that successful leadership is defined as "behaviour that conforms to societal moral and ethical standards, the law

and professional military practice" (p. 8). He claims that successful military leaders of the past developed competencies of aggression, determination, physical courage, endurance, mental toughness, and the ability to manage chaos and adapt to change.

Nowadays the selection of the Army's leaders is based on identifying key behavioural attributes deemed necessary to lead and command soldiers of the future (Snider, 2003). There are many core qualities deemed essential for the modern officer; integrity, professionalism, honour, initiative, loyalty, and courage (Curran, 2000), but perhaps the most important skills are those associated with the ability to perform the management and leadership functions of command.

Leadership is a deciding factor on the battlefield, and no one leadership style or trait is sufficient to be universally effective in all situations (Harper & Hayward, 2003). Therefore junior officers need to be able to adapt their command style to best suit the dynamic nature of the environment (Romaine, 2004). A successful officer must be able to perform these functions in both the stable non-operational environment, and whilst on operational deployment. The ability to perform efficiently and effectively under conditions of extreme mental and physical stress is crucial. Junior officers face the challenge of inspiring and motivating soldiers in increasingly complex and ambiguous circumstances (Romaine, 2004). Personality and trait research has consistently shown that scores on scales of adjustment are related to effective task performance in the operational environment (Driskell & Salas, 1991).

Romaine (2004), discusses the need for leaders who are 'emotionally intelligent', who can interact with and influence subordinates within their span of command, whilst still providing mutual respect and discipline. This is supported by Hughes (2000), who reported a lack of trust and mutual respect between young officers and their leaders in a study conducted with the New Zealand Army. Hughes suggests that effective leadership is being lost as commanders' focus on functional aspects of leadership rather than relational aspects. Romaine also suggests that military leaders need to adopt a systems approach to thinking and problem solving. Pech (1998) categorizes this ability as Conceptual Skill, defining it as a "general analytical ability, logical thinking, proficiency in concept formation and conceptualisation of complex and

ambiguous relationships, ability to analyse events and perceive trends, anticipate changes and recognise opportunities and potential problems." (pp. 90-91).

It is the responsibility of the New Zealand Army then to ensure that the most suitable applicants are selected for training, and that these applicants receive thorough and comprehensive training to prepare them for commissioned service and ultimately operational service. "It is the true test of military preparedness that military systems and personnel operate efficiently under the stress of combat" (Driskell & Salas, 1991, p. 190).

A soldier in the New Zealand Army typically follows one of two career paths; enlistment at the rank of Private to progress through the Non-Commissioned Officer (NCO) ranks, or enlistment as an Officer Cadet to become a commissioned officer. To become a Regular Force General List Officer a candidate must first be selected at an OSB. After selection the candidate will attend one of three commissioning courses; the New Zealand Commissioning Course (NZCC), the Kippenberger Scheme, or the Australian Defence Force Academy (ADFA). All candidates complete a 7 week Initial Officer Training (IOT) aimed at providing the candidate with the basic military skills required in the New Zealand Army. NZCC candidates then attend a nine month course in Waiouru before graduating from OCS (NZ), whereas Kippenberger candidates are required to complete a three year degree at Massey University followed by the nine month NZCC course in Waiouru. ADFA candidates complete their four year course in Australia and graduate from Duntroon Military Academy[1] (Taylor, 2005). At the completion of officer training, candidates graduate into the Officer Corps and are placed in command of others. This process is displayed in the flowchart in Fig 1.

[1] These schemes apply for Regular Force candidates only. Different schemes apply for Territorial Force candidates.

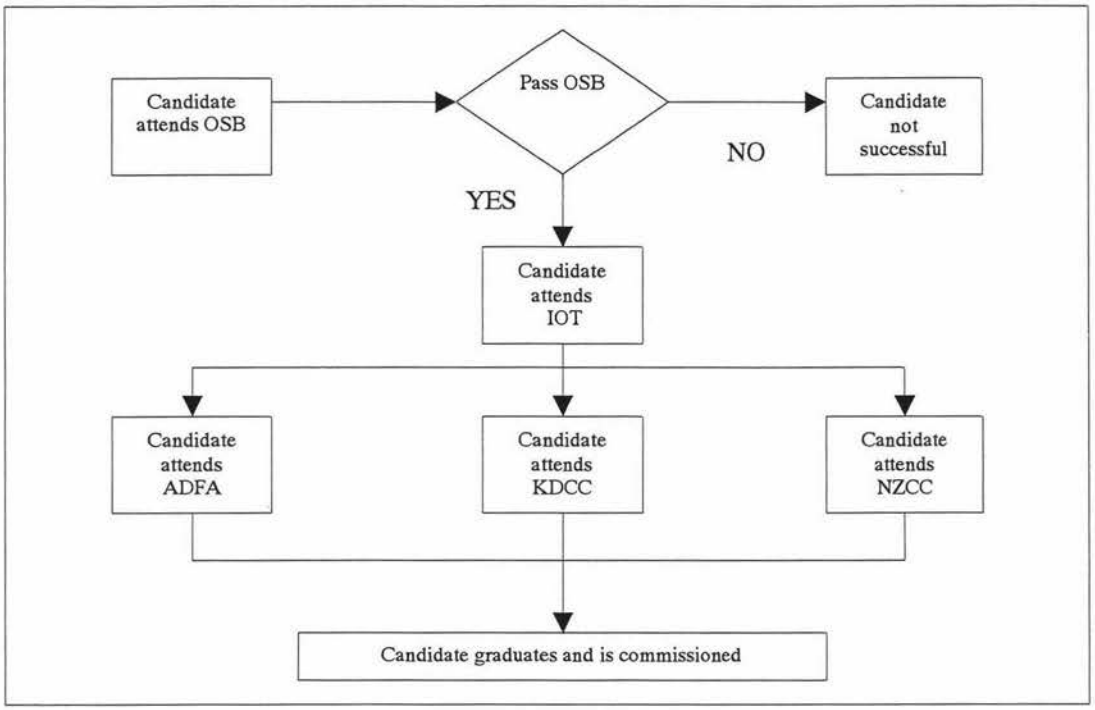


Figure 1. Flowchart of the Officer Training Process

The training conducted at OCS (NZ) in Waiouru is both mentally and physically challenging. Candidates are taken out of their comfort zones and are pushed to excel. The course currently has nine different field exercises, totalling over 17 weeks spent in the field (Preece, 2005). Typically speaking, at least one of these components is conducted overseas in locations such as Brunei, Fiji or Australia. Due to the cost both to the military and to personnel selected, it is crucial to the New Zealand Army that only those candidates assessed as having the potential to train and work as an officer are selected. Additionally, for both ethical and legal reasons it is important that the OSB is assessed as providing both valid and reliable grounds for selection.

In order to ascertain whether the OSB is successfully selecting suitable candidates it is necessary to conduct a criterion-related validation study. This involves determining the relationship between scores on a predictor and scores on some criterion. Criterion-related validation is deemed appropriate when the objective of the validation is to determine whether a selection measure or system provides suitable grounds to make predictions (Aguinis, Henle, & Ostroff, 2001). There are two criterion used in this research; the first relates to candidate training performance, and the second relates to candidate job performance.

Schmitt and Robertson (1990) suggest a five-factor model to assess criterion-related validity as follows:

1. Job Analysis
2. Criterion development and measurement
3. Predictor development and measurement
4. Evaluation of the relationship between the predictor and criterion
5. Consideration of the practical utility and social, legal or organisational implications of the selection strategy.

Previous research conducted by the APS has investigated both criterion and predictor development and measurement (Bennett, 1990; Carston, 2002; Eagar, 2004; Gracie, 2004), whilst the New Zealand Defence Force Competency Review and selection competency research are contributing the review on the Junior Officer Job Analysis. The scope of the present research has focused in particular on evaluating the relationship between predictor and criterion.

The purpose of this research is to investigate the predictive validity of assessment outcomes utilising two criterion measures. The results of this investigation will enable improvements to be made to the OSB to enhance the selection of junior officers in the New Zealand Army. This study is a partial replication of the research conducted by Kelly (1994) on the OSB in 1989. This research investigated the incremental validity of the selection board using OSB scores as the predictor and graduation status and job performance rating as given by the Army's annual reporting document for officers (MD68) as the criterion. This research was conducted on data collected from selection boards held over the period 1986-1988. Additionally the current study has used junior officer performance ratings provided by supervisor ratings as a distant criterion subsequent to training performance at OCS (NZ).

The introductory chapters of this thesis encompass the literature review, which is broken into three key areas. Firstly, the review has focussed on important issues related to the theory of validation and the validity of selection methods providing the theoretical foundations of this study. This is followed by an investigation into Assessment Centre (AC) validity and its use in selection decisions. Finally the OSB is discussed, past validation studies on the OSB are presented, and the current selection

process and the measures used are investigated. These chapters are intended to familiarise the reader with the theoretical background and rationale for this validation study.

1.2 Literature Review

The purpose of a selection system is to predict the best candidate for the job from a pool of applicants. Selection of the "right people for the right jobs constitutes a source of competitive advantage" (Salgado, Viswesvaran, & Ones, 2001, p. 165). The optimal selection and placement of employees' impacts on the financial health and well-being of an organisation. The fundamental property of a selection system is predictive validity (Schmidt & Hunter, 1998). The foundation for the prediction of performance is that the assessed characteristics will generalise to the work environment and will endure long enough to enable useful predictions (Guion, 1987). Predictive validity is increased when selection decisions can accurately discriminate between high performance and low performance as measured by the job criteria. The personnel selection process consists of five key functions; analysis of the job, determining desirable job behaviours and outcomes, development of assessment procedures, making inferences and predictions on applicant performance and, evaluating individual performance by a criterion (Binning & Barrett, 1989).

Traditional validation studies relied on Trinitarian views of validity, focusing on the construct, content and criterion (predictive) validation efforts independent from each other. These studies looked at the validity of the predictor and the coefficient of correlation, and evaluated the construct validity of the measures employed, rather than the selection system as a whole. In contrast, many modern validation studies are now based on a Unitarian framework of validity (Guion, 1980), and are conducted to provide "evidence for determining the inferences that can be made from scores on a selection measure" (Gatewood & Feild, 2001, p. 163).

The Unitarian framework concentrates on the interpretations made from statistical scores, and aims to combine the outcomes of Trinitarian validation to produce a more holistic view. The term 'Construct validity' is employed to encompass both construct, content and criterion validity, and has become the focus for most validation efforts in industrial selection (Guion, 1980). This is reiterated by Steege and Fritscher (1991) who state that; "today's emphasis is on construct validity and the realisation that interpretation (inferences) and the use of the test score is the proper subject of validation" (p. 25).

The importance of validation studies is reiterated by research conducted by the British Military (Elshaw, Abram, & Weston-Lovelock, 1997). Elshaw et al. suggest that the process of validation allows for the production of quantifiable data to support selection decisions, improving selection efficiency, and ensuring a non-discriminatory and fair process. Validation also allows the user to justify the use of the selection system, and helps to identify which elements provide the greatest predictive ability. Organisations don't want to use complicated and expensive techniques, when simpler, less costly substitutes are just as effective (Byham & Thornton, 1986). Jones, Herriot, Long and Drakeley (1991), suggest that there may be a significant amount of redundancy in AC procedures, and that additional exercises do not contribute to unique explained variances. Therefore, it may be possible to reduce the procedures without reducing the predictive validity of the AC.

Recent research on employee selection practices employed by large organisations in New Zealand reported 100% of surveyed organisations (n=100) used interviews in their selection process', with 99% using past information, training and experience, and 98% using references. The study also showed that 47% of organisations used some form of cognitive ability test, this was compared to 46% who used personality tests, and only 10% who used some form of AC (Keatly, 1998). Research on the use of the AC is well documented and will be discussed later in this chapter. Firstly, we will look at the Unitarian framework of validity used as the foundation of this research.

1.2.1 Conceptual Analysis of Validity

The present validation study is based on the theoretical model of validity proposed by Binning and Barrett (1989) who claim that "demonstrating the validity of decisions based on psychological assessment is of fundamental importance to personnel and other applied psychologists" (p. 478). It is essential to recognise that validity is not a feature of an assessment procedure or a test, but is a characteristic of the inferences made from information provided by these means. The process of validation involves determining the formal, logical correctness of some proposition or conclusion (Reber & Reber, 2001). Therefore, when we examine validity, we are actually examining the accuracy of our inferences and conclusions (Gatewood & Feild, 2001; Guion, 1987, 2002; Steege & Fritscher, 1991).

A nomological framework of construct validity was presented by prominent theorist Nunnally (1978). This framework identified four components of validity as; the predictor construct, the predictor measure that operationalises the construct, the criterion construct and the criterion measure (Fig 2).

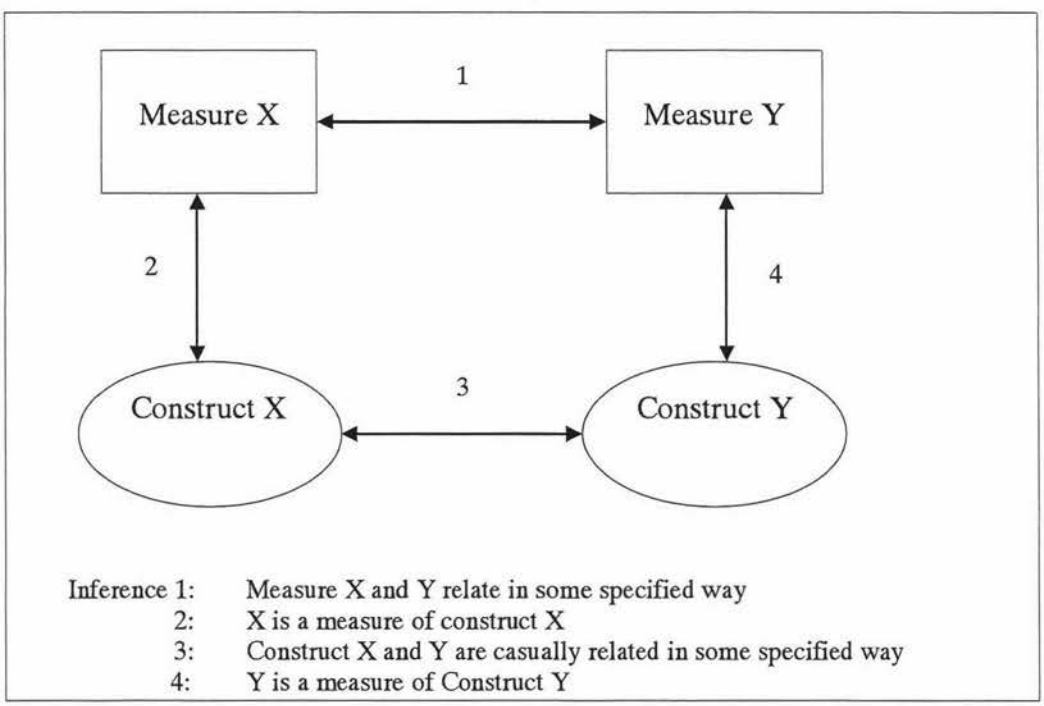


Figure 2. Nomological framework proposed by Nunnally (1978)

In order to validate the inferential linkages in this framework we need to gain evidence to support at least three of the four inferences. As we can only gain direct empirical evidence to support inference 1, the validity of two of the three remaining inferences is assumed based on logical or qualitative evidence (Nunnally, 1978).

This nomological framework can be utilised in applied psychological research, and is a useful foundation for personnel selection theory-building. Binning and Barrett (1989) built upon this theory by adapting it for personnel selection purposes, highlighting additional linkages between the components (Fig 3). This linkage is also reported in the work by Schmidt, Hunter & Outerbridge (1986) and Gatewood & Feild (2001).

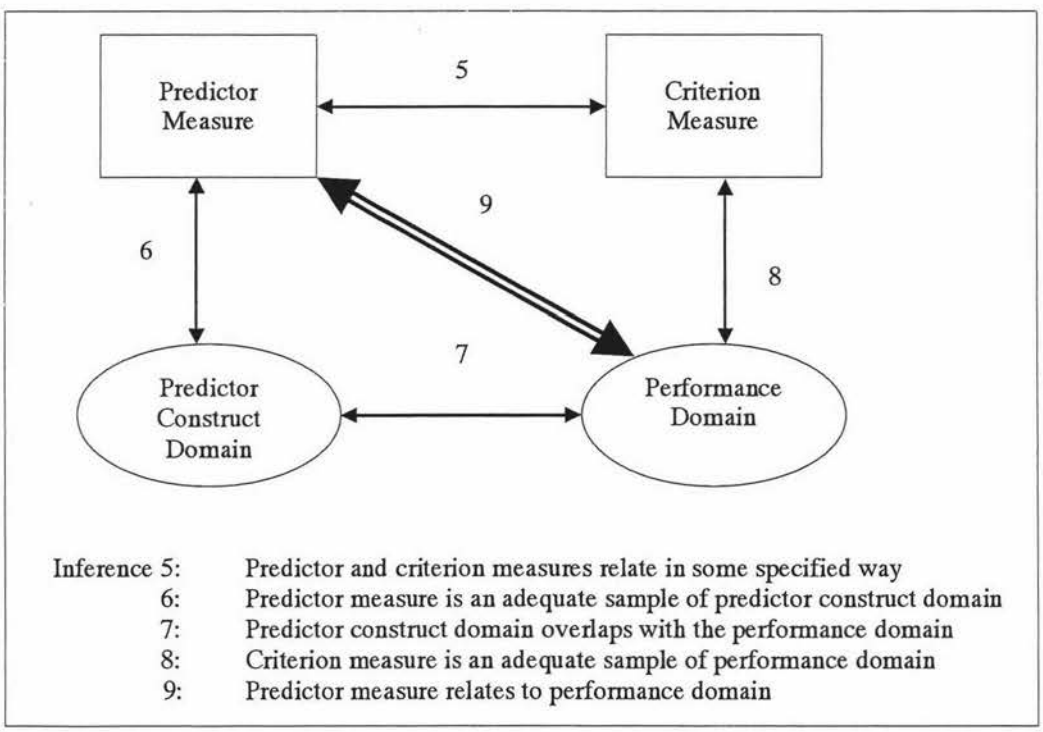


Figure 3. Conceptual Framework for personnel selection proposed by Binning and Barrett (1989)

Gatewood and Field (2001) identified a link between the performance domain and the predictor measure, and suggest "if a predictor is correlated with job-relevant criteria, then we can draw inferences from scores on the measure about individual future job performance in terms of these criteria" (p. 161). This is supported by Smith (1994),

who describes the essence of selection as the identification of the measures that are able to predict job success. Additionally, there must be justified links between the predictor measure and criterion measure, and the criterion measure and performance domain to achieve high validity (Braun, Wiegand, & Aschenbrenner, 1991). In particular we rely on Inference 8 to be strong in order to support the conclusions reached about the validity of other inferences (Binning & Barrett, 1989).

Within this framework, Inference 9 is our primary area of interest. This inference is the extent to which the predictor measure will allow accurate predictions about individual job performance. In order to justify Inference 9 we can empirically validate the predictor and criterion (Inference 5), provided we are confident the criterion accurately measures the performance domain (Inference 8). This method is referred to as criterion validation. Alternatively, we can justify Inference 9 if we have strong qualitative evidence that our predictor measure efficiently taps the construct domain (Inference 6) and that our construct domain is a essential component of job performance (Inference 7). This method is referred to as construct validation. The predictor – performance link can also be justified isomorphically using content validation, when the predictor represents the actual performance domain such as in a work sample.

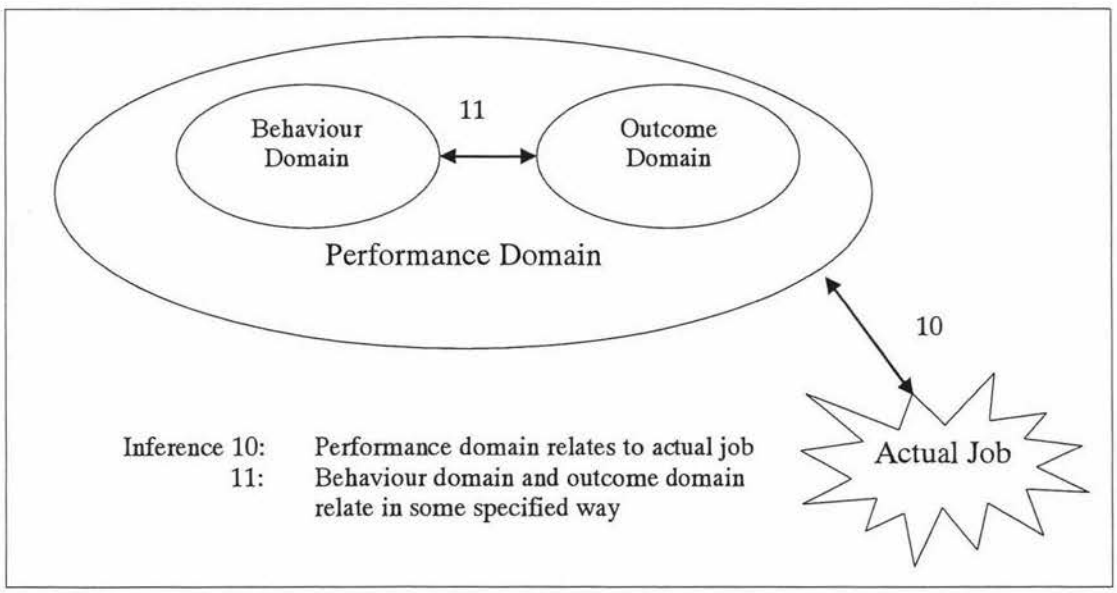


Figure 4. Conceptualisation of the Performance domain. Adopted from Binning and Barrett (1989)

We can further develop this framework to extend our understanding of the performance domain as shown in Fig 4. The performance domain can be conceptualised through a collection of overt job behaviours, or by a prescription of outcomes. This process is achieved through job analysis. Inference 10 represents the accuracy to which actual job demands are reflected in the performance domain.

The work of Binning and Barrett (1989) is supported by Smith (1994), who continues to build upon the components of criterion validation by focussing on the predictor measure – performance domain link (Inference 9). Smith proposes that "the observed validity of a predictor is a function of the domains of characteristics they cover and the accuracy with which the domain is measured" (p. 15). Smith's theory on the domain of human characteristics proposes four separate domains as:

a. Characteristics irrelevant to work performance
b. Characteristics relevant to all work, called Universals
c. Characteristics relevant to particular jobs, called Occupationals
d. Characteristics relevant to the way a person relates to a particular work setting, called Relationals

Whereby: *Observed Validity* $\int\sum$ *characteristic domains X measurement accuracy*

Work performance or the performance domain depends on the latter three of these domains of human characteristics. When a measure accurately reflects the degree that Universals, Occupationals and Relationals are represented in work performance (Inference 9) then validity is expected to be high. Selection validity therefore, is a function of the accuracy of the selection criteria to represent the job Knowledge, Attributes, Skills, and Other Abilities (KASOs) required and the accuracy the measures used to assess those criteria. Correct identification of the characteristics of the performance domain as criteria, and accurate measurement of these criteria will increase selection validity.

Perhaps the most relevant of these domains to selectors are the Universals. Universals enable effective performance in 90% of jobs and display a stable linear relationship to job performance (Smith, 1994). The Universals include; cognitive ability, which determines the quality and speed of learning on the job, vitality, which encompasses

both mental and physical energy, and work importance, which includes work ethic, centrality and job involvement.

Occupationals are specific knowledge, skills, abilities or personality traits that may be required in a job. Jobs that are more technical in nature or require more specialised workers will have a higher occupational emphasis. Lastly, Smith (1994) discusses Relationals. Relationals concern Person-Environment fit and are characteristics that make a person and job compatible. These characteristics may include company culture, working hours and employment conditions, level of autonomy or inter-personal skill requirements. The validity relationship can be represented as follows:

$$V = \int\sum U\,x\,W_u\,x\,S_u\,x\,B_u + O\,x\,W_o\,x\,S_o\,x\,B_o + R\,x\,W_r\,x\,S_r\,x\,B_r$$

Where: V - Validity
U – Universals
W – Weighting
S – Sampling
B – Objectivity
O – Occupationals
R – Relationals

Given that there are three key domains of work performance (Universals, Occupationals, and Relationals) it makes sense that different predictors vary in their success to target each of these domains. Interviews are more likely to target Relationals when they are unstructured, but well structured interviews can target Occupationals and Universals also. Ability tests generally target the Universal domain, whereas a work sample tests focuses primarily on Occupationals.

The final model presented by Binning and Barrett (1989) extends Fig 3. to include convergent and discriminate evidence of criterion validity, offering an idiographic framework for selection. In Fig 5. additional inferences show how Inferences 5, 6, 7 and 8 can be strengthened.

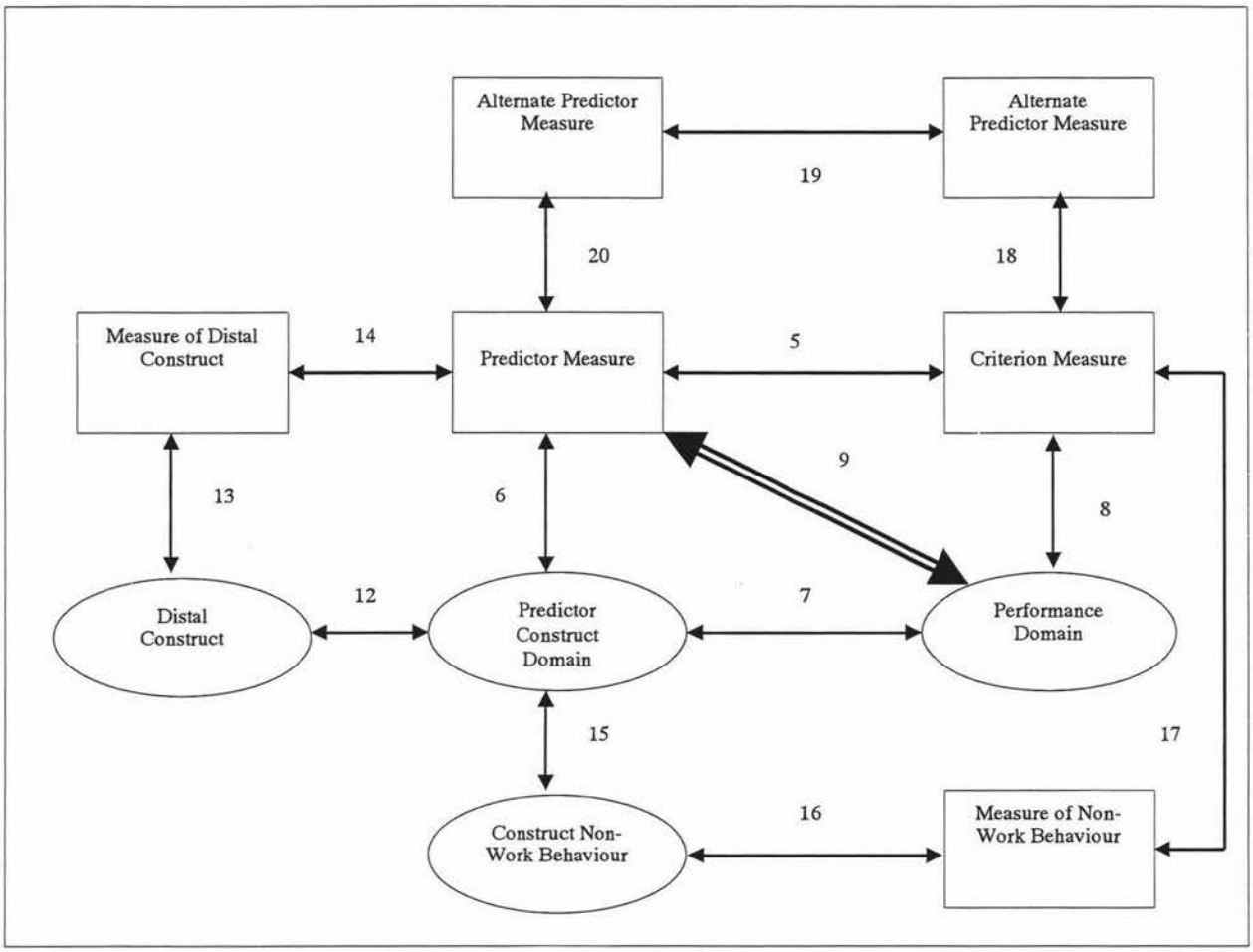


Figure 5. Idiographic Framework for Selection proposed by Binning and Barrett (1989)

When the predictor and the criteria reflect congruent domains, observed validity will increase. Where a mismatch occurs validities will not be as high. For example, we would expect a trade test (high in Occupationals) to correlate higher with a work sample, than with a supervisor rating (high in Relationals). However, we would expect peer ratings (also high in Relationals) to correlate highly with supervisor ratings. Additionally, the more criteria we predict for, the greater the likelihood that we will sample all the relevant domains. Increasing the range and number of predictor measures used in a selection process, increases the likelihood of capturing a greater component of the performance domain, resulting in a greater level of explained variance. This is the reasoning behind the use of the Assessment Centre.

1.2.2 The Assessment Centre

This chapter provides a definition of the Assessment Centre (AC). It covers the history and development of the AC and its components over the last century. This is followed by discussion into AC validity and the problems encountered with its use in the organisational setting. Next, the use of competency-based frameworks is looked at, and finally, some of the popular components employed in the AC are discussed.

An AC consists of a standardised evaluation of behaviour in which multiple assessment techniques are employed to assess individual performance (Byham & Thornton, 1986; Spychalski, Quinones, Gaugler, & Pohley, 1997). It is a predictor of performance in the sense that it is a method used to make predictions about the future job performance of applicants (Arthur, Day, McNelly, & Edens, 2003). The use of the AC grew from the realisation that individual techniques of assessment such as the interview and biodata were no longer sufficient for identifying the domains of work performance in a holistic manner (Carston, 2002).

The popularity of the AC has increased markedly over the last two decades. Recent research provides data suggesting that between 45 and 85 percent of large organisations in the United States and United Kingdom now use the AC for the selection of management personnel (Carston, 2002). Results from New Zealand studies indicate that AC use in non-management roles increased by 27% between the years 1991 and 1998, however this was matched by a 28.5% decrease in the use of AC techniques in management roles. Interestingly, there was a 107% increase in the use of cognitive ability tests, and a 135% increase in the use of personality tests in non-management roles over this period, and a 380% increase in the use of cognitive ability tests, and a 48% increase in the use of personality tests in management roles (Keatly, 1998).

The development of the AC can be traced back to the 1920's when the German Military developed the techniques for their officer selection. Two of the key activities developed by the German psychologists were the 'Rundgespräch' or Leaderless Group Discussions, and the 'Führerprobe' or Command Task. The Command Task was further developed by the Hungarian Army in 1936 (Jones, 1991), to include

multiple assessors and a structured manner of assessing planning ability that is still remnant in the New Zealand Army OSB. This was succeeded by the War Office Selection Board (WOSB) designed by the British post WWII, and Intelligence Officer Selection designed by the Office of Strategic Studies in the United States. The first civilian application of an integrated assessment process was the British Civil Service Selection Board (CSSB), and followed shortly thereafter in America with Bray's work at the Chicago Bell Cooperation (Zaal, 1998).

In the early 1970's concerns were raised over the inappropriate conduct of some AC's threatening the reputation of the AC process. As a consequence, during the third International Congress on the AC, certain ethical guidelines were established for the use of the AC. These guidelines have since been revised under the International Taskforce on Assessment Centres in 1979, 1989, (Howard, 1997) and more recently in 2000. In order for a selection process to be considered an AC it must meet a series of criteria as laid out in the guidelines (Taskforce of Assessment Centre Guidelines, 2000). These criteria are as follows:

1. The selection process must be based around a thorough job analysis. This involves identifying the relevant KASOs that accurately reflect the job performance domain.

2. Through this process behaviours relevant to the KASOs are identified and classified into selection competencies.

3. Assessment techniques are then determined on their potential to assess the selection competencies.

4. Multiple assessment techniques are used.

5. Job-relevant simulations are included in the selection process.

6. Multiple assessors are used for assessment ratings, and assessors are trained in observation and assessment techniques.

7. Behavioural observations are recorded which form the basis for report production.

8. Selection decisions are made on the basis of integrated data collected during the assessment period.

The first step in this process involves the analysis of the job. However the evaluation of management performance is problematical due to the nature of the management role. It is often difficult to define management jobs precisely due to the large variation and complexity in job tasks and behaviours; each management job is unique to its organisation, and department (Byham & Thornton, 1986). This can lead to inaccuracies in the job analysis process and thus affect the development of the assessment process.

The products of job analysis are the person specification and the job description. These two outcomes allow AC developers to determine what dimensions are essential for good performance on the job. From this point they are able to select assessment techniques to tap into these dimensions. These techniques can range from interviews and psychometric tests through to situational tests and work samples.

ACs typically place individuals into syndicates, which enables group assessment to take place. This allows assessors opportunity to observe peer interactions (Byham & Thornton, 1986), group communications skills and individual ability to influence group thinking. Research by Kleinmann (1993) suggests that high performing candidates in AC generally possess increased levels of intelligence, social skills, achievement motivation, authoritarianism and high self-esteem.

Assessment Centre Validity

Numerous studies have been conducted to assess the predictive validity of the AC method and its generalisability (Chan, 1996; Damitz, Manzey, Kleinmann, & Severin, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Klimoski & Brickner, 1987; Schmidt & Hunter, 1998; Zaal, 1998). Research shows that as the specificity of the predictors used increases so too does the validity of the inference (Braun et al., 1991),

however this causes generalisability to decrease. The reverse is also true, as the generalisability of an inference increases the specificity and validity of the predictor is decreased.

Smith (1994), proposes that when selecting applicants for long term career progression where the final job varies greatly from the entry level position, then predictors that focus on Universals will be more beneficial than those that focus on the Occupationals for the entry level job. This is supported by Guion (1987) who suggests that if your purpose is to hire generally good people then a global predictor is useful. Focussing on Universals will increase the generalisability of the selection system, but may also cause specificity and validity to be decreased as stated above. Additionally, Braun et al. (1991) claim that most procedures used in military assessment are better predictors of training performance than of job performance. This is based on an accumulation of military research, and may actually reflect inaccuracies in criterion measurement as a lack of standardised job analysis.

Gaugler et al. (1987) report a corrected average validity for AC of 0.37. However, this meta-analysis was based on a large range of validation studies. Research suggests that where AC are used strictly for selection or development purposes rather than for promotion purposes, higher levels of validity are expected (0.41 and 0.46 respectively) (Zaal, 1998). Jones (1991) reviewed the predictive validity of the British Navy's Admiralty Interview Board (AIB) (n=2144), after corrections were made for range restriction the results showed a correlation of r=.39 with training performance.

In a meta-analysis conducted by Arthur et al. (2003) on the criterion-related validity of the AC at the competency level, the authors found an Overall Assessment Rating (OAR) validity rating of r=.36. However, when the research was analysed at a competency level three competencies were found to have validities higher than that of the OAR (Influencing others, Organising and planning, and Problem solving), suggesting that certain competencies are better predictors of performance than others. Using the top four competencies, the authors yielded a correlation of R=.45 and were able to explain 20% of the variation in job performance, in comparison to Gaugler et al. (1987) who explained 14% with the OAR. These findings are supported by Jones (1991) who suggests that there is a degree of statistical redundancy within the AC.

The variation of measures, levels of generalisability, and difference in both assessment tools and methods, makes it difficult to assess the overall validity of the AC. Additionally, many AC are conducted without proper reference to the Task Force guidelines (2000) and lack the precautions that provide the robustness to the AC. Walsh, Weinberg and Fairfield (1987) found that the validity of the AC varied as a function of the characteristics of the candidates, assessors, the validation efforts undertaken and the AC itself. It is because of this that many validity findings conducted from meta-analysis may actually be more conservative than necessary.

Problems in the Assessment Centre

The AC has consistently been shown to have high criterion validity and content validity, however, it has frequently been criticised as lacking in construct validity (Arthur et al., 2003; Arthur, Woehr, & Maldegen, 2000; Kleinmann, 1993; Lievens, 1998). From a Unitarian perspective of validity this lack of construct validity is paradoxical, as the three forms of validity should be interrelated as shown in Fig 5. One cause of this lack of construct validity may be due to the use of the OAR in the validation of selection decisions. Arthur et al. (2003) claim that the use of the OAR results in a loss of construct level information, as separate constructs are collapsed into a single rating and it becomes conceptually unclear as to exactly what the OAR is measuring.

The reliability of the criterion and predictor is also a key concern, where measurement error exists in both measures it will be compounded, and decrease the validity of predictions (Gatewood & Feild, 2001). Schleicher, Day, Mayes and Riggio (2002) discuss one component of measurement error referred to as the 'exercise effect'. Exercise effect occurs when ratings on the same dimension across exercises correlates lowly, and ratings on different dimensions within the same exercise correlate highly. This leads to low convergent validity in the first instance and low discriminant validity in the second (Lievens, 1998).

Exercise effect is a violation of the AC construct validity, where the predictor measure is the assessors rating given on several dimensions. Numerous studies have shown that these ratings reflect method variance rather than dimensions, where the

assessor rating is reflecting candidate performance on a task rather than the hypothesised trait (Atkins & Wood, 2002; Fleenor, 1996; Kleinmann, 1993; Russell & Domm, 1995). Some studies suggest that exercise effects are actually a representation of cross-situational specificity, where variation in assessor ratings is actually due to true variation in applicant performance across exercises, and therefore a valid source of variation in AC performance (Lance, Newbolt, Gatewood, Foster, French, & Smith, 2000).

Comparative analysis of exercise effect is complicated due to the large variance in AC structure and dimension measurement (Lievens & Conway, 2001). This point is emphasised by research conducted by Keatly (1998), which showed that of 100 large NZ organisations only 16 employed some form of AC. Of these, eight organisations said they rated their applicants against set competencies, with five reporting no formal method of rating applicants. Three of the 16 organisations said they assigned an OAR while informally referring to competencies, and only seven claimed to provide some form of assessor training.

There are numerous design characteristics that can help to increase dimension variance and reduce the effects of exercise variance. One such category of characteristics involves limiting the cognitive overload assessment for assessors. These interventions can include; providing adequate assessor training, limiting the number of dimensions rated, use of behavioural checklists, and making the dimensions known to applicants (Lievens & Conway, 2001). Schleicher et al. (2002) propose the use of Frame of Reference (FOR) techniques. FOR training emphasises the distinctiveness of each of the assessed dimensions, and uses behavioural descriptors to encourage assessors to report on exercise behaviours (Atkins & Wood, 2002; Lievens, 1998).

Bernadin and Buckley originally developed FOR training in 1981 (Schleicher et al., 2002). The primary focus is to reduce 'idiosyncratic standards' held by assessors and to substitute these with a common frame of reference for assessment. This is achieved by:

1. Operationally defining selection dimensions and providing behavioural examples.

2. Discussing the dimensions, the various levels of performance associated with each, and the indicative behaviours of those levels.
3. Providing opportunity for assessors to practice the use of the FOR.
4. Providing feedback on assessor evaluations.

Another intervention involves the use of expert assessors in the AC. A four year validation study conducted by Tziner, Ronen and Hacohen (1993) found a positive relationship between AC ratings and subsequent job performance, presenting a .40 correlation coefficient between the OAR and criterions of advancement and salary. This study looked at the predictive validity of behaviour dimensions, and the use of psychologists' vs. managers as assessors. The findings suggested that trained psychologists were able to produce ratings with higher discriminant and convergent validity, while managers' ratings produced a higher correlation with overall performance excellence. These results support Lievens and Conway's (2001) and Lievens (1998) proposition that both psychologists and managers are employed in assessor roles.

In a comparison study of the 215 US organisations Spychalski, et al. (1997) report that AC predictive validity increases with the use of peer assessment, psychologist assessors, the number of evaluation devices used and an increase in female assessors. This is supported by research conducted by Walsh et al. (1987), who investigated the interaction effects of assessor and applicant gender. They found a significant main effect with female applicants being rated higher than male applicants when an all male assessor group was used. The authors speculate that this is due to "male assessors being more lenient with their evaluations of females because they perceive that female assessees have already surmounted a number of obstacles in being seriously considered for the male-oriented position" (p. 308). However there were no significant gender differences in ratings when the assessor group included females.

Validity levels can also be affected by sampling error. When the sample used in the validation is too small to be representative of the population or to meet the criteria of psychometric assessment, then the results can not be reliably interpreted (Gatewood & Feild, 2001). Samples can also be affected by range restriction. Range restriction refers to the "artificial reduction in the variance in one or more of the variables under

consideration" (Aguinis et al., 2001, p. 39). Range restriction involves the violation of the assumption of variances of scores, causing the variance to be reduced. This effect is seen when pre-selection is conducted, as the selection pool is already restricted by removing the unsuitable candidates. It is also observed when there is criterion restriction. Criterion restriction is usually due to turnover, when employees leave the organisation prior to the validation data being collected, resulting in missing criterion data (Gatewood & Feild, 2001).

Direct criterion contamination can also affect the validity findings for AC validation studies (Tziner et al., 1993). Criterion contamination can occur in a number of ways, but is most commonly due to either; disclosure of an applicants AC performance to their supervisor, whereby their performance on the AC bias' the supervisor's future actions of promotion or job opportunity, or due to assessors not adhering to AC competencies and rating the applicant's performance in accordance with perceptions of promotional ability. Research conducted by Klimoski and Strickland (1977) investigated criterion contamination and suggested that AC ratings were prescient rather than making valid evaluations, and were simply duplicating the assessment of promotion decisions. Tziner et al. (1993) recommend the use of multiple criterion measures to help combat the effects of criterion contamination. Additionally, they suggest not using an OAR, as the collapsing of criteria may introduce elements of assessor bias, especially when non-mechanical or judgemental methods are used to derive the OAR (Dunnette & Borman, 1979).

Other research suggests that it is the model of the AC that is at fault, and that organisations should look towards using task specific-models rather than dimension-specific models (Russell & Domm, 1995). Task-specific models do not use a set of competencies for assessors to rate against. Instead assessors simply provide a rating on candidate performance at the task. These tasks models need to be high in job relevancy and are therefore difficult to employ when a high degree of job training is required (Lowry, 1997).

Competency-Based Frameworks

The OSB uses a competency-based framework for selection. The competencies used on the OSB are outlined in Appendix 1. The use of competency-based frameworks is relatively common within military selection systems particularly in New Zealand and the United Kingdom. Due to the high level of training provided to candidates in both leadership and military skills, it is not necessary to only select candidates who currently display good leadership, but instead to select those who exhibit the potential to develop good leadership through training. The competency framework is developed around leadership traits that are deemed as contributing factors to leadership potential.

The Regular Commissions Board (RCB) uses a framework based on eight different competencies, seven of which are assessed using the Behavioural Anchored Rating Scales (BARS) (Drive and Determination, Personality and Character, Impact, Problem Solving, Physical, Analysis and Planning, and Oral Communication) and one (Intelligence) which is assessed through psychometric testing (Elshaw et al., 1997). These competencies are similar to those suggested by Bass (1990) as the ten top leadership traits.

The number of competencies used affects the convergent validity of the assessment, and impacts on the cognitive loading of the assessor (Bycio, Alvares, & Hahn, 1987). Many AC utilise over ten different dimensions for assessment, however studies show that assessors frequently use only three (Arthur et al., 2000; Gaugler & Thornton, 1989). This is supported by factor analytic studies of final dimension ratings (Gatewood, Thornton, & Hennessey, 1990). Research suggests that assessors possess a limited capacity to process information and that reducing the number of competencies to observe for helps to reduce this loading and increase assessor accuracy (Arthur et al., 2000; Bycio et al., 1987; Lievens, 1998; Lievens & Conway, 2001).

Recent meta-analytic studies of AC dimensions have aggregated a variety of dimensions down to six (Level 1) dimensions as follows: Influencing Others, Organising and Planning, Problem Solving, Consideration/Awareness of Others, Communication, Drive, and Tolerance for Stress (Arthur et al., 2003). Four of these

six dimensions accounted for the predictive validity of the AC (Problem Solving 15%, Influencing Others 3%, Communication 1%, Organisation and Planning 1%), with the first three achieving higher predictive validities than the OAR reported in Gaugler et al. (Gaugler et al., 1987) meta-analysis.

Research conducted by Kleinmann (1993) on the transparency of AC competencies suggested that the accurate identification of the assessed competencies produced better performance on behalf of the candidate. These findings were supported by Lievens and Conway (2001), who claim to "highly recommend the disclosure of dimensions" (p. 144) in developmental ACs and suggest it as acceptable practice for selection oriented ACs.

Job Relevant Simulations

One of the key advantages of the AC over other selection techniques is its use of job relevant simulations. Work samples and situational tests are two forms of job relevant simulations. The work sample involves a test situation in which the applicant performs one or more tasks from the actual job; this type of simulation is particularly useful when the applicant is required to have the necessary practical skills for task completion at the time of selection. Schmidt and Hunter (1998) reported an incremental validity gain of .12 (24%) when the work sample was used in combination with a test of General Mental Ability (GMA), however subsequent studies have failed to achieve this validity level (Salgado et al., 2001).

Situational tests use a context rich environment to expose the applicant to a variety of items and events that may be encountered during the job. The situational test is better suited to allow candidates to display behavioural attributes as well as physical skills (e.g. problem solving ability, stress tolerance) (Siegel, 1986). This method allows the candidate to display real skills and abilities, rather than having to base decisions on a candidate's ability to describe their level skill through oral expression as used in an interview (Asher & Sciarrino, 1974; Byham & Thornton, 1986; Gaugler et al., 1987; Zaal, 1998). Research evidence suggests that job relevant simulations assist in producing higher predictive validities, increase face validity and reduce adverse impact (Robertson & Kandola, 1982).

The Interview

The interview is the most frequently employed method of personnel selection; however unstructured interviewing is still far more prevalent than structured interviewing (Salgado et al., 2001). For a period of time now the interview has reported high predictive validity when employed in the correct manner. Research by Howard (1997) now places the validity of the interview at .40. This is supported by meta-analytic research conducted by Huffcutt, Roth and McDaniel (cited in Salgado et al., 2001)), who found a corrected average correlation of .40 between interviews and GMA.

When the interview is based on structured questioning techniques and related back to the job analysis it can achieve relatively high validity coefficients. Panel interviews (Guion, 1987) and high structure interviews (e.g. behaviour description interview and situational interview) report high validity and appear to be more reliable and valid than unstructured interviews (Lievens, Harris, VanKeer, & Bisqueret, 2003). Wiesner and Cronshaw's (1988) meta-analysis of interview structure found a validity correlation of .62 for the structured interview. Performance in structured interviews also correlates highly with both job experience (r = .71) and job knowledge (r =.53) (Salgado et al., 2001).

Structured interviewing helps to improve both the accuracy and consistency of the information collected (Gatewood & Feild, 2001). Research shows that the concepts best evaluated in the interview are Personal Relations (sociability and verbal fluency) and Good Citizenship Behaviour (motivation) (Jennings, 1998).

Measures of Cognitive Ability

Generally speaking the best predictor of future performance is past performance. Often though, it is not viable to base the selection of applicants on past performance as applicants may be new to the job or work environment and we may discriminate against those who lack the knowledge base or skill familiarity necessary to achieve high performance. Research shows that when hiring employees without previous job

experience the best predictor of future performance is GMA, r =.51 (Braun et al., 1991; Schmidt & Hunter, 1998).

GMA has also been shown to be a consistent predictor of job performance across a wide variety of jobs, especially when the job is complex (Anderson, Born, & Cunningham-Snell, 2001; Lievens et al., 2003). "The predictiveness of *g* varies systematically as a function of job complexity" (p. 225), where job complexity is increased so too is the validity of *g* (Ree, Caretta, & Steindl, 2001). Additionally, GMA has been linked as a key determinant for training (Ree & Earles, 1991), where research has shown training performance as more a function of GMA than of specific aptitude factors (Ree et al., 2001).

In a United Kingdom study, Bertua, Anderson and Salgado (Bertua, Anderson, & Salgado, 2005) found operational validities ranging from .38 to .47 for GMA and overall job performance, and validities ranging from .54 to .62 for GMA and training success. GMA was shown as the best predictor of performance in training in a study conducted on pilot and navigator training in the US Air Force, and was strongly related to trainee learning ability (Olea & Ree, 1994).

Path analysis conducted by Schmidt, et al. (1986), produced model and path coefficients of .13 between work sample performance and GMA, .56 between work sample performance and job knowledge, and .55 between GMA and job knowledge. The authors concluded that "the major causal impact of ability is on the acquisition of job knowledge, which in turn has a major impact on performance capabilities as assessed by work sample measures" (p. 433).

When GMA tests are combined with other methods as part of the AC there is a 4% increase in validity (R =.53) (Schmidt & Hunter, 1998). GMA has also shown an average correlation of r =.60 with leadership qualifications, and has been strongly and consistently shown to be related with leadership effectiveness (Braun et al., 1991). The employment of multiple assessment techniques has had successful results in increasing the predictive validity of selection decisions.

Personality Questionnaires

Personality traits, (in particular conscientiousness) have been shown to provide incremental validity in predicting future job performance over and above that of GMA (Dayan, Kasten, & Fox, 2002; Ree et al., 2001; Salgado et al., 2001). Research into personality testing shows conscientiousness to be one of the most consistent predictors across validity studies (Anderson et al., 2001). Despite this, personality traits have returned generally weak correlations with job performance ranging from r =.15 to r =.2 (Braun et al., 1991). An examination of the ability of the Five Factors to predict job performance produced correlations close to zero, with the exception of conscientiousness (Hough & Ones, 2001).

A study conducted by Sumer, Sumer, Demirutku and Cifci (2001), used content analysis, followed by Principle Components Analysis and Confirmatory Factor Analysis to determine officer selection constructs in the Turkish Armed Forces, five key dimensions emerged as follows: conscientiousness (self discipline), self confidence, agreeableness/extraversion, leadership, and military factor (pride, respect for command, commitment). Interestingly, conscientiousness was able to explain more than two thirds of the variance in the factor analysis.

Results from the use of personality testing in military selection have been less than conclusive. Generally weak predictive validities are reported for personality tests and job performance (Braun et al., 1991). Hough and Ones (2001) suggest that this is because personality constructs allow us to tap factors of job performance that are related to interpersonal constructs such as Organisational Citizenship Behaviours and social awareness. Other assessment techniques specific to this research design will be discussed later in the design section of this research.

1.2.3 The New Zealand Officer Selection Board

The objective of the OSB is to select candidates who are deemed suitable for officer training and who have the potential to develop into sound military leaders (Russell, 2005). The OSB is conducted at Trentham Military Camp in Upper Hutt, Wellington. It is held three times a year in June, September and December and caters for up to 64 candidates applying for the Regular Force (RF), Territorial Force (TF) and Specialist Officer positions. It is conducted over a period of five days, and includes a variety of assessed and non-assessed activities, that allow the candidate a realistic job preview of the military environment (Horn, 2005).

Past Validation

Byham and Thornton (1986) suggest that validation research should attempt to evaluate the contribution of the following relationships:

1. The reliability of the observers' judgements
2. The correlation of the techniques with the OAR
3. The correlation of techniques with subsequent criteria of management success
4. The unique contribution of the technique over and above others.

Although previous research investigating the validity of the OSB has attempted to achieve this, it has been somewhat piecemeal, focussing on elements of the selection process rather than the outcomes of training. As a result, few studies have investigated the ability of the OSB to discriminate between successful and unsuccessful junior officers.

This lack of validation literature is not uncommon. Since its establishment in 1949 until 1989 the RCB used by the British Army has only undertaken three validation studies; one each in 1952, 1957 and 1967. The result of this most recent validation study reported a moderate correlation of r =.31 with the OAR and performance as a junior officer (Dobson & Williams, 1989). The difficulty in conducting this type of research project is in finding appropriate criteria against which selection decisions can be validated (Bennett, 1990).

The most comprehensive validation study to be completed was by Kelly (1994), this study used seven predictor measures (prelim psychologists ratings, B40 scores, Purdue scores, 2 x peer assessment ratings, final Military Testing Officer (MTO) rating, and final rating) and two criterion measures (graduation from OCS, and job performance as provided by the MD68). After corrections were made for both range restriction and criterion unreliability the results revealed that while several of the predictors were strongly correlated with each other, no linear combination of predictors was able to predict graduation status from OCS (NZ). In particular, the study found a correlation between B40 score and Mean Performance Appraisal rating ($r = .19$, $p < .05$).

A pilot study conducted by the APS in 1988 examined the outcome of selection decisions made by the psychologists and MTOs. The study compared the OSB gradings against training outcomes at OCS (NZ). Despite the small sample size ($n=116$) and range restriction problems, the results highlighted some concerns about the validity of the selection decisions made (Bennett, 1990). The results show that psychologists were able to better predict training success than MTOs (average of 72% correctly classified over seven OCS (NZ) classes, compared to 66%). MTOs consistently made more false positive selections than psychologists, who were better able to discriminate between successful applicants and not successful. However, psychologists had a higher rate of false negatives when compared to MTOs.

More recently, research into the OSB components has focused on the use of the ACER Advanced Test B90 (B90). The B90 is a measure of intellect or cognitive ability utilised during the pre-selection phase of officer selection. This research has consistently identified a gender bias with males scoring significantly higher than females (Allen, 2000b; Allen & Mirfin, 1999; Barker, 1997; Gracie, 2004; Lane, 1998). This is similar to previous findings as reported by Cullen (1995) and the test publishers (Reid & Croft, 1991).

A suitability comparison study between the B90 and the Raven's Advanced Progressive Matrices (Raven's APM), found that both tests were positively related to MTO gradings, with the strongest relationship existing between B90 scores and

ratings of planning ability. However, when this relationship was controlled for, both the B90 and the Raven's APM ceased to predict OSB success (Gracie, 2004).

In 1995 small changes were implemented in to the OSB process. Firstly, the pre-selection was changed to allow recruiters to conduct interviews and testing sessions (APS, 1995). More recently, the OSB has included an introductory lesson on leadership with a command task example provided to candidates to help 'level the playing field'. Additionally, some elements of assessment including the Physical Training Exercise have been modified to reflect more job relevancy (Carston, 2002).

In 2005, a partial validation study was undertaken by the Royal New Zealand Navy (RNZN) to validate their Final Officer Selection Board (FOSB) (Harrison, 2005; Ratcliffe, 2005). The aim of this research was to assess the validity of the assessment tools employed in the FOSB at predicting individual performance under training. The results were mixed, largely due to sample size restrictions resulting in a loss of power in the study. Despite this the researchers were able to show that the FOSB does show predictive validity, with good face validity.

Leadership potential and personal qualities were both found to correlate significantly with measures of training ($r = .46$). The researchers used a Training Criterion Score (TCS) which was a composite score derived from eight criterion competencies from junior officer training, however, when placed into the regression equation only leadership potential was able to explain a significant portion of the variance in TCS. Investigation into the contribution of individual competencies revealed that the competency 'Oral Communication' is seen to be an essential element to leadership, with a correlation coefficient of $r = .45$ to measures of leadership potential.

Research conducted as part of the Officer Selection Review Board for Army 2005 (Allen, 2000c) investigated the comparability of the selection methods employed on the OSB against other Technical Cooperation Program (TTCP) nations. The findings concluded that OSB selection methods were equivalent to those used in other countries ($n=15$), and that the process and systems employed were similar to those used in Australia and Canada.

Despite this research, the OSB process is not without its critics. Curran (2000) claims that the current selection process is not based on a modern job analysis and fails to identify specific leadership traits and behaviours. Additionally, it fails to define leadership potential. Despite this the OSB is based on relevant literature collected from other military research, and does present high face validity. Overall, most assessments of the OSB have been positive, with results suggesting it to be a fair and robust process.

Pre-Selection Process

Prior to attending an OSB all candidates are pre-selected through their local Army Recruiting Office. The Army recruiters identify potential candidates based on candidates qualities of; Organisational Fit, Job Fit, Intellect, Motivation, and Fitness. This process involves meeting a series of eligibility criteria based on age, citizenship, security, medical, fitness and academic elements (ARRC, 2005). It is at this stage that candidates are administered the B90 on which they must achieve a required standard in order to progress forward. This process is outlined in more detail in Appendix 2. The important factor to note here is that the selection system has already had an impact during pre-selection, whereby candidates who fail to meet the necessary criteria for attendance on the OSB do not progress, creating a degree of range restriction within the OSB candidate population.

Current Selection Process

The OSB adheres to a principle of negative selection. Basically, there are two ways of accepting and rejecting candidates; these are termed positive and negative selection. A positive selection method is based on the premise that the objective of the selection system is to select the best candidates for the job. This is used when there is a large selection pool and relatively few positions available. A negative selection method is based on identifying those candidates that do not possess the necessary criteria for selection (Turner, 1988). This method is used when there are many positions available to a smaller selection pool and is designed to maximise selection in a high demand versus low supply environment (Carston, 2002). Zaal (1998) suggests that method of selection has a resultant effect on test utility. Where positive selection methods are

used to select the best candidates, test utility is increased, however when negative selection methods are used, test utility can be decreased. This is because many tests are designed to discriminate top performance, rather than testing for a minimum standard. By selecting at the low end of the tests' range, we do not make full use of its ability to discriminate.

Key Functional Groups

The New Zealand Army places a large emphasis on its officer selection, and the OSB is very manpower intensive involving 22 high ranking officers and over 40 military personnel in total. There are three key groups of personnel involved in the OSB; the Administration Team, the candidates and the Board.

The Administration Team involves the:

1. Officer in Charge (OIC) and the Liaison Officers (LO) who help the candidates with daily administration and ensure they arrive at activities on time,
2. The Psychology Clerks who administer the planning task and psychometric testing, as well as marking and grading all psychometric documents; and
3. The Support Team who work behind the scenes to ensure syndicate rooms and exercises are prepared (Horn, 2005).

The candidates are all those applying for positions on the NZCC, Kippenberger Scheme, ADFA, TFCC, Malone Scheme or Specialist roles. The candidates are split into syndicates of seven or eight members. The syndicates are generally age related, however there is an effort made to ensure that candidates are not disadvantaged by gender or ethnicity (Horn, 2005). Candidates are provided with military clothing and wet weather gear to wear during most of the OSB.

The Board assumes the role of assessment of the candidates. There are five elements to the Board, these include; the Senior Board, an Education Officer, Military Testing Officers (MTOs), Observers, and Syndicate Psychologists. The latter three make up the Syndicate Board.

The Senior Board is the most experienced group within the OSB, and consists of the:

1. Board President (BP), who is of Brigadier rank,
2. Two Deputy Presidents (DP), who are of Colonel rank,
3. The Military Secretary (MS), who is of Lieutenant Colonel rank, and
4. The Territorial Force Military Secretary (TFMS), who is also of Lieutenant Colonel rank (Carston, 2002).

The responsibilities of each of these members are outlined in Table 1.

Table 1.

Responsibilities of Senior Board Members taken from the OSB Folder

Senior Board Member	Responsibilities
Board President (BP)	Responsible for the general conduct of the board
Deputy President (x2) (DP)	Interviews candidates and is responsible for the daily running of the board
Military Secretary (MS)	Interviews candidates and acts as an advisor and assistant to the Board President
Territorial Force Military Secretary (TFMS)	As above for Military Secretary

The Board also has a specialist Education Officer who provides advice to the Board on academic qualifications achieved by the candidates, and who also provides expert assessment on written communication gradings (Horn, 2005). Finally, each syndicate has allocated a MTO (Lieutenant Colonel rank), an Observer (Lieutenant Colonel/Major rank), and a Syndicate Psychologist. In order to be eligible to act in the role of MTO the officer must have participated in three previous boards in the role of Observer (Bennett, 1990) and have met certain selection criteria as laid out in the OSB folder (Lowry, 1993). The MTO and Observer work separately from the Psychologist in order to achieve independent and objective assessments of candidate behaviour (Carston, 2002). The responsibilities of each of these members are outlined in Table 2.

Table 2.

Responsibilities of Syndicate Board Members taken from the OSB folder

Syndicate Board Member	Responsibilities
Military Testing Officer (MTO)	Responsible for assessment of candidates in syndicate. Observe all tasks and activities, and submit an individual report on each candidate at the completion of the OSB
Observer	Learn the process and provide relevant information to the Senior Board as required.
Syndicate Psychologist	Acts as an independent assessor to provide specialist advice on candidate aptitude, motivation, temperament and intellect. Interviews candidates in syndicate and observes all tasks and activities.

This layered approach to assessment allows for a more thorough observation of candidate behaviour. Byham and Thornton (1986) propose that the assessor to assessee ratio should be as low as possible, near 1:3 for the AC to be more successful. The calculated ratio for the OSB (excluding the Education Officer and using six Syndicate Psychologists) is 1:2.56, and well within the suggested ratio levels. The experience level of the assessors used also adds to the robustness of the OSB (Bennett, 1990). The majority of the assessors have held senior command or unit command positions, effectively making them subject matter experts (Horn, 2005).

The Board meets twice throughout the duration of the OSB. The first meeting, known as the Preliminary Board is aimed to identify marginal candidates, and allow the Senior Board to identify which candidates need closer observation (Kelly, 1994). At the conclusion of the assessment period the Board reconvenes for the final board meeting where each candidates is assigned an 'IN' or 'OUT' grading to represent their success or otherwise at the board . Each candidate is individually informed of their final grade by the BP, and then debriefed by their syndicate MTO and Observer. Candidates are given the opportunity to discuss any questions about their performance with their Syndicate Psychologist (Horn, 2005).

Assessment Criteria and Techniques

The OSB uses a Multi-trait Multi-method approach to assessment. Throughout the OSB, candidates are assessed on group functioning and on individual performance (Turner, 1988). This is based on four general criteria; written, verbal, practical and social (Carston, 2002). The criteria are further expanded into the competencies as listed in Appendix 1. These competencies were initially developed through analysis of a job specification of the Royal Military Academy Sandhurst Young Officer (RMAS YO) (RMAS, 1985) and criteria used in the Australian Army Officer Selection process (Turner, 1988). Guion (1987) suggests that "where it is not feasible to prove a job related criterion one can chose predictors that have been shown generalisability by the accumulation of prior research." (p. 211).

The RMAS YO job specification identifies very similar tasks, responsibilities and duties as those required of a New Zealand Army Officer. Additionally, the identified conditions under which the job is performed are very similar. The RMAS YO job specification identifies sixteen Personal Qualities as required for carrying out the duties of the job. The Australian Army identified ten personal qualities criteria associated with the officer role. It is from these Personal Qualities that the NZ Army OSB criteria were first developed (Turner, 1988). A number of assessment techniques are employed to assess the competencies, with a variety of activities conducted to align with these criteria.

There are five key assessment techniques employed on the OSB . These are as follows:

Peer Assessment: Throughout the selection process the candidates are asked to conduct a peer assessment on the members of their syndicate. These assessments help to provide insight into a candidate's acceptability within their syndicate.

Written Essays: Candidates are given an essay topic and are required to write an essay during the selection board. The essay is used to assess a candidate's abilities of written expression.

Interviews: Each candidate participates in three interviews. Firstly, they are interviewed by the BP and either the MS or the TFMS, then by one of the DP's, and finally by their Syndicate Psychologist. The interviews are semi-structured in that the interviewers utilise a set of guidelines for questioning, but do allow for variation in questioning technique and exploration of answers. These interviews are designed to help the Senior Board and the Syndicate Psychologist to assess the candidate's motivations and military aspirations (Russell, 2005).

Psychometric Testing: Candidates complete the B90 prior to attending the OSB, as part of the recruiting process. In addition to this, candidates complete the Raven's APM and two personality measures, the Gordon Personality Profile Inventory (GPP-I) and the Eysenck Personality Questionnaire-Revised (EPQ-R) whilst at the OSB.

Behavioural Observation: A wide variety of exercises are observed and candidates are rated by their MTOs against the 12 competencies. Each of the competencies are rated against a five point behaviourally anchored rating scale (Appendix 1). The technique of behavioural observation is by far the largest component of the OSB. Whereas psychometric testing provides insight into candidate personality profiles and predicted behaviour, it is through systematic behavioural observation that these profiles are reinforced (Horn, 2005).

There are six different Behavioural Specific Exercises conducted during the OSB. These include; physical exercises, leaderless group activities, planning exercises, individual and group problem solving activities, public speaking activities and command exercises . These activities are described in more detail in Appendix 3.

Many of these activities are designed specifically to align with activities that a candidate could expect to perform whilst at OCS (NZ) or in the Junior Officer Role. They have been designed to target the competencies that have been identified for the Junior Officer Role, and to allow a candidate multiple opportunities to display those competencies (APS, 2004).

During each OSB, prior to the commencement of assessed activities, behavioural observation training is conducted for the MTOs in order to ensure accuracy in ratings. This training is conducted by one of the Syndicate Psychologists and highlights a number of techniques to help assessors avoid bias in their ratings. This training is attended by all MTOs and Observers. The training aims to introduce seven key areas of assessor observation as shown below :

1. Focus on the behaviour.
2. Focus on the observation of behaviour.
3. Focus on the interpretation.
4. Focus on the analysis and decisions made about the behaviour.
5. Feedback is provided to assessors.
6. Methods of recording behaviour.
7. Methods of reporting behaviour.

The emphasis of the training is on objective observation of behaviour and reporting observable candidate behaviour, rather than making judgements or assumptions. The behavioural descriptions of each competency and the behavioural anchored rating scale is also discussed (Carston, 2002).

Physical Testing

Candidates are assessed on their physical ability whilst on the OSB through a variety of physically oriented tasks. In addition to this candidates need to achieve a required standard on the Entry Fitness Level (EFL) which is a similar test to the Required Fitness Level (RFL) test that all enlisted personnel must pass bi-annually (ARRC, 2005). Physical testing is predominately restricted to physically demanding occupations such as the military, police and fire service (Salgado et al., 2001). Tests of physical ability are usually highly intercorrelated and do not tend to show any incremental predictive validity over that of GMA (Schmidt & Hunter, 1998).

1.3 The Present Research

The aim of the present study is to investigate the incremental validity of the components employed in the OSB over and above that of cognitive ability testing. Additionally, this research aims to assess the construct validity of OSB through comparison of selection dimensions against those identified as necessary for good performance as a commissioned officer. This can be delineated into two separate objectives. The first objective is to assess the relationship between OSB component scores and performance during officer training at OCS (NZ) as measured by OCS (NZ) reports. The second objective is to assess the relationship between OSB component scores and performance as a junior officer as measured by annual personnel reports and immediate supervisor ratings. The results of this research will enable suggestions to assist in streamlining the OSB for effectiveness and efficiency. The following section provides the hypotheses for the outcomes of this research.

1.4 Hypotheses

Lievens et al. (2003) maintain that aside from assessing a predictor's validity, it is both theoretically and practically essential to examine the predictive validity of individual predictors over and above each other. Doing so from a theoretical standpoint allows us to determine what constructs a predictor is tapping into; from a practical point of view, investigating incremental validity allows us to determine the additional explained variance that each predictor can provide, and therefore its utility in the AC. In order to assess the incremental validity of the components of the OSB a number of hypotheses are proposed. Firstly, it is important to determine whether the AC is explaining more variance in candidate performance than a simple test of cognitive ability is able to achieve. To do so we will examine the influence of both combined MTO ratings and individual MTO ratings against each of the criterion variables (training performance and job performance).

Hypothesis 1: *MTO combined ratings will provide incremental predictive validity to the selection of candidates over cognitive ability testing.*

Hypothesis 2: *Individual MTO ratings will provide incremental predictive validity to the selection of candidates over cognitive ability testing.*

Secondly, we want to determine whether the use of psychologists in assessment is assisting in achieving higher predictive validity, and whether combined or individual psychologist ratings are more predictive.

Hypothesis 3: *Psychologists combined ratings will provide incremental predictive validity to the selection of candidates over cognitive ability testing.*

Hypothesis 4: *Psychologists individual ratings of temperamental suitability, intellectual ability, educational ability, and motivation will provide incremental predictive validity to the selection of candidates over cognitive ability testing.*

Next we want to identify the value of Peer Assessments in the AC, and determine whether the Peer Assessment rating increases predictive validity.

Hypothesis 5: *Peer Assessment ratings of Likeability and Leadership will provide incremental predictive validity to the selection of candidates over cognitive ability testing.*

Finally, we want to look at the relationships that exist between the personality measures employed on the OSB and performance outcomes both on the job and during training.

Hypothesis 6: *Elements of the EPQ-R will return positive relationships with both training performance and job performance measures.*

Hypothesis 7: *Elements of the GPP-I will return positive relationships with both training performance and job performance measures.*

Additional to these hypotheses, it is also expected that increased time since commissioning will be correlated with increased performance. Therefore Hypothesis 8 will investigate this interaction.

Hypothesis 8: *Increased time since commissioning will be positively correlated with higher job performance.*

The following chapter will discuss the design of this research, and covers the sample selection and description. The predictor measures and criterion measures are discussed in more detail and the method of data collection is described. This is followed by a discussion on the statistical analysis process.

Chapter 2
Design

2.1 Sample

The Sample was selected from the general population of junior officers (ranking from Second Lieutenant to Captain) in the New Zealand Army. The sample was not randomly selected, instead participants had to meet a variety of criteria in order to be included. This sample consisted of all officers who had graduated from OCS (NZ) between the years 1996 and 2004, and were still currently serving on a Regular Force engagement (n = 183). Information sheets and consent forms inviting participants to partake in the research were sent to all participants. From this sample, 20 personnel were operationally deployed overseas, 12 personnel were away on non-operational deployments overseas (overseas exchanges, postings or promotion courses), and nine personnel were on courses held in New Zealand.

A total of 76 participants responded to the invitation to partake in the research. Of this 4 declined and 72 participants agreed to provide consent for performance data to be collected, resulting in a 41.5% response rate (39.3% consent rate). There are three reasons for this low response rate; firstly, the consent distribution happened to coincide with the changeover period between Operation Crib 5 and Operation Crib 6 (deployment to Afghanistan), resulting in more people being away from work due to pre-deployment training, actual deployment or post deployment leave. Secondly, the research collection period also coincided with an annual exchange with the British Army called Exercise Longlook, resulting in a number of participants being away. Lastly, the distribution period coincided with a promotion course that a number of participants were attending, resulting in a number of participants not receiving the information due to mail not being forwarded.

2.2 Predictor Measures

The predictor measures were all collected as part of the Officer Selection Board (OSB) process. They included measures of cognitive ability, personality, MTO ratings, psychologist's ratings and peer assessment ratings. These measures are designed to assess the candidate against competencies identified through job analysis processes conducted in the UK. They are thought to accurately tap into the predictor construct domain, ratifying Binning and Barrett's (1989) Inference 6 (see Fig 3.) It is also anticipated that these measures relate to the performance domains of training performance and job performance, in order to ratify Inference 9 (Horn, 2005).

2.2.1 Psychometric Tests

Psychometric tests have been shown to be valid and reliable if administered in a standardised fashion and have been carefully evaluated to determine their ability to successfully predict performance (Atkinson, 2003). Despite this, psychometric tests have frequently been criticised as they only allow for assessment of behaviour as indicated by the applicant, rather than actual displayed behaviour, requiring a certain degree of self awareness and leaving the results open to impression management (Damitz et al., 2003). For this reason, psychologists observing on the OSB use both test results and candidate observation to assess candidate performance (Horn, 2005). The OSB uses both cognitive ability and personality tests to assess candidates during the selection process.

Cognitive Ability Testing

In the literature review we discussed the predictive ability of tests of General Mental Ability (GMA). Research has consistently shown GMA to be a reliable predictor of both job performance and training ability (Anderson et al., 2001; Chan, 1996; Ree et al., 2001; Ree & Earles, 1991). Sixteen percent of organisations in the USA utilise tests of cognitive ability for management selection, whilst 42% use some form of specific aptitude test (Gowing and Slivinski, 1994 cited in Salgado et al., 2001). The OSB conducts two tests of cognitive ability, the B90 and the Raven's APM.

ACER Advanced Test B90 (B90)

The B90 is a test of general intellect for individuals of relatively superior intellectual ability. The B90 was introduced into New Zealand Army and Royal New Zealand Air Force (RNZAF) officer selection systems in 1994 (Atkinson, 2003). All officer applicants sit the B90 at their local recruitment office; it is administered by their local Recruitment Officer (Gracie, 2004) who is trained in administration by the APS. The B90 is used as a screening test, with applicants needing to achieve a score within a selected range in order to progress on to the OSB.

The B90 is a group test consisting of a 70 item form designed to measure general intellectual ability (Rust, 1991). Its intended use is to assess the intellectual abilities of students planning to undertake tertiary study, and for adults in a selection or training context (Reid & Croft, 1991). It was developed from two Australian tests, the ACER Advanced Test B40 (1983) and the ACER Test of Cognitive Ability (1983). Items from these two tests were modified and rewritten to suit a New Zealand market. The test authors claim to have maximised the use of items that are judged to provide an assessment of GMA (Reid & Croft, 1991). There are issues surrounding the method of standardisation and sampling used to create the B90 (Rust, 1991), however, the New Zealand Army uses its own norms (n=1045) a factor that helps to negate these issues (Bowden, 1999).

The B90 reports adequate reliability coefficients based on the research conducted on its two founding tests, the B40 and Test of Cognitive Ability. Research conducted by the APS has investigated the effects of age, gender, education and ethnicity on B90 scores. The results indicated normal distribution across the sample with no main effects for age (Allen & Mirfin, 1999). However significant differences were found between males and females with males scoring significantly higher (Allen, 2000b), and between New Zealand Maori and New Zealand European/Pakeha, with Maori applicants scoring significantly lower (Allen, 2000a). A significant relationship was also found between B90 scores and Year 12 results (Allen & Mirfin, 1999).

Recent research on the use of the B90 on the OSB found a significant, moderate, positive relationship with one of the competencies related to planning ability (r (322)

=.39, $p<.001$). A number of significant, weak, positive correlations were also found between B90 score and assessment competencies (Gracie, 2004).

Raven's Advance Progressive Matrices (Raven's APM)

The Raven's APM is a test of GMA (Bertua et al., 2005) and was based on the Raven's Standard Progressive Matrices (SPM) (Raven, 2000). The APM was developed for persons over the age of 11, and was designed to spread the top 20% of the population. It assesses high-level eductive ability and indicates a respondent's speed for accurate intellectual work. The test contains 36 items with a ten item familiarisation test. Flanagan, Genshaft and Harrison (1997) claim it to be an almost pure measure of *g*. The Raven's progressive matrices series (includes the Advanced, Standard and Coloured versions) have consistently shown significant correlations with scores on other intelligence tests (Raven, 2000). Perhaps the most important element of the Raven's progressive matrices is evidence supporting its use as a gender and culturally unbiased assessment tool (Irvine & Berry, 1983; Jensen, 1980).

The Raven's APM has only been used on the OSB since 2001; hence very little validation research has been conducted by the APS. Research conducted by Gracie (2004) (n = 335) found no significant differences between Raven's APM scores and gender, but did find a significant difference for ethnicity; however this difference did not emerge clearly between the two groups. Significant, weak, positive relationships were found between Raven's APM test performance and GPPI Original thinking ($r(333) = .10$, $p = .047$) and GPPI Personal Relations ($r(333) = .13$, $p = .014$), suggesting that higher levels of these two traits resulted in higher Raven's APM scores.

Personality Testing

The two measures of personality used on the OSB are the GPP-I and the EPQ-R. The EPQ-R is used to determine how well applicants cope with stress, whilst the GPP-I is used to determine work oriented aspects of personality (APS, 2004). The EPQ-R is a self report ipsative questionnaire. Ipsative tests require participants to rate themselves in relation to other constructs (within-person ranking) and usually utilise forced

choice questions (Hough & Ones, 2001). Although the GPP-I does use force choice questions it is not purely ipsative as the question format uses high and low preference items paired together in a tetrad, and the respondent must choose the statement most and least like them from the four statements in the tetrad. This also allows to scales to remain independent of each other (Bowden, 1999).

Research conducted within both the New Zealand Army and the RNZAF shows that GPP-I traits of personal relations, vigour, ascendancy, reliability, and sociability, are strong indicators of success at selection boards. Additionally, this research showed that within the RNZAF high scores on scales of Extraversion and Neuroticism were strong indicators of successful performance on the selection board (Atkinson, 2003).

Eysenck Personality Questionnaire (Revised) (EPQ-R)

The Eysenck Personality Questionnaire (Revised) (EPQ-R, 1985) is a revision of the EPQ developed by Eysenck and Eysenck in 1975. It is based on a research foundation consisting of over forty years of development, and 'many hundreds of psychometric and experimental studies' (Eysenck & Eysenck, 1991, p. 1). Subsequent studies of the EPQ identified three major faults with the Psychoticism scale which impacted on gender comparisons, hence the scale was revised and the EPQ-R was produced (Eysenck & Eysenck, 1991). Reliability and validity data presented for the EPQ show internal consistencies above .70, with test-retest reliabilities above .70. The EPQ has also shown high factor analytic validity, with the three main factors clearly separating (Kline, 1993). Reliability scores for the revised Psychoticism scale are .78 for males and .76 for females.

In developing the scales used in the EPQ-R, Eysenck argued for the psycho-physiological basis of personality (Furnham, 1992). The EPQ-R helps to provide the OSB psychologists with an assessment of emotional stability based on the three 'Super Traits' proposed by Eysenck (1970) as follows (Atkinson, 2003);

P. Psychoticism (tough mindedness)

E. Extraversion (outgoing nature)

N. Neuroticism (level of anxiety)

These three traits are each connected to physiological processes and are based on the 'Big Five' (Hough & Ones, 2001). The EPQ-R has three additional scales measuring criminality (C), addiction (A), and intentional distortion or the lie scale (L). In addition to detecting those who are "faking good" the EPQ-R L scale can also detect a need for approval or "conforming factor" (Bowden, 1999). The OSB only utilises the first three, (P,E,N) and the lie scale (L) in their assessment of candidate suitability and uses their own normative data (n=1045) .

Gordon Personal Profile Inventory (GPP-I)

The GPP-I was constructed to assess eight important factors in the personality domain (Gordon, 1993). The questionnaire is a combination of two personality questionnaires; the Gordon Personal Profile (GPP) and the Gordon Personal Inventory (GPI). The GPP-I is not based on any particular theory but on constructs that have been defined through factor analyses (Guion, 1998), and is an 'example of an empirically derived questionnaire designed to be used in industrial settings' (Bowden, 1999, p. 57).

The GPP-I measures respondent tendencies against eight scales. These are Ascendancy (A), Responsibility (R), Emotional Stability (E), Sociability (S), Cautiousness (C), Original Thinking (O), Personal Relations (P), and Vigour (V). The GPP-I has demonstrated adequate reliability and validity, returning coefficient alpha scores above .80 in all scales. Test-retest reliabilities show the GPP-I to be fairly stable across time and reliable (Guion, 1998). The OSB utilises all eight scales in their assessment of candidate suitability and uses previous OSB normative data (n=1045) .

Intentional Distortion and Impression Management

There has been much debate over the effects of intentional distortion (faking, impression management) on the validity of test scores. Research conducted by Hough (1998, cited in Hough & Ones, 2001) and Ones, Viswesvaran and Schmidt (1993) claim that the setting moderates the effects of distortion on validity. When job applicants were directed to distort their scores, low validity resulted. However, when job applicants were directed not to distort their scores, the validities results were similar to those obtained from job incumbents. The findings suggested that intentional

distortion has limited to moderate effects on test validity levels (Hough & Ones, 2001). This is supported by research that asserts that where participants are warned not to distort or fake their answers distortion is decreased (Dwight & Donovan, 1998 cited in Bartram, 1995; Hough & Ones, 2001).

Results on impression management studies for the EPQ-R suggest that it is open to faking (Bowden, 1999). In a study conducted by Elliot, Lawty-Jones and Jackson (1996), three groups of respondents were assigned (two experimental groups; stockbroker and librarian, and a control group) and asked to complete the EPQ-R. The two experimental groups were asked to imagine they were going for a job interview for their respective professions, whilst the control group was asked to answer honestly. The authors found that Neuroticism scores were lower for the experimental groups than for the control, and that the Lie scale could not discriminate between respondents that answered honestly and those that were faking. Studies conducted on response distortion with the GPP-I prove it to be relatively resistant to faking (Bowden, 1999). The author claims that while some responses on scales can be distorted the magnitude of distortion was small (Gordon, 1993).

2.2.2 Rating Scales

A rating is a category-oriented judgement (Braun et al., 1991), and as such is subject to a variety of bias' and contamination factors. There are four main types of ratings used in validation and selection research; peer ratings, supervisor ratings, expert ratings, and self ratings. Of these, supervisor ratings are most common. The reliability and validity of ratings can be improved through the conduct of assessor training and the use of Behaviourally Anchored Rating Scales (BARS). BARS are designed to illicit performance ratings on employees on relevant job dimensions, they are particularly effective when based on a thorough job analysis, and anchored by unambiguous definitions (Jacobs, 1986).

In a meta-analysis conducted by Harris and Schaubroeck (1988) relatively high correlations were found between peer and supervisor ratings (p=.62), and moderate correlations were found between self and peer ratings (p=.36), and peer and supervisor ratings (p=.35). However, the authors advise that these findings should be

interpreted with caution as there has been much inconsistency across studies conducted on rating to rating correlations. Three rating measures were used in this study; MTO ratings, Psychologist ratings, and Peer Assessment ratings.

Military Testing Officer Ratings

The MTOs assess candidates against twelve competencies, and are provided with BARS for each of these competencies as outlined in Appendix 1. Each candidate is awarded a grade for each behavioural exercise completed, and then the MTOs use these grades to determine an overall grade for each competency based on observed behaviour. There are five degrees of aptitude for each competency; Strong (S), Good (G), Adequate (A), Limited (L), and Weak (W). Each grading is accompanied by a descriptor to maintain consistency between MTOs (APS, 2004; Gracie, 2004). For the purposes of this study, these gradings were coded and mechanically combined to provide an OAR for each participant. Tziner, Meir, Dahan, and Birati (1994) suggest that "a mechanical, standardised process should be used to generate an overall rating which combines the dimensional assessment data both accurately and exhaustively" (p. 231).

Psychologists Ratings

The psychologists provide independent ratings on candidates against competencies of temperamental suitability, intellectual ability, educational ability, and motivation (Gracie, 2004). These competencies are graded in a similar fashion to MTO ratings using a five degree scale of competence with scores of S, G, A, L, and W . The ratings are based on a variety of information available to the psychologist including test scores, school results, behavioural observations and interview observations (Horn, 2005). These gradings were also coded and mechanically combined to provide an OAR for each participant.

Peer Assessment Ratings

The use of peer assessment ratings in the AC has been shown to help to improve AC validity (Dayan et al., 2002; Spychalski et al., 1997). Research conducted by Shore, Shore and Thornton (1992) shows strong support for the use of peer assessments, they state that 'peer assessments are most useful when they focus on dimensions for which the participants have greater amounts of behavioural information on which to base their judgements' (p. 50). Hough and Ones (2001) state that ratings collected through peer assessment have higher validity than those of self assessment. This is supported by Chao, Dobson, Tziner and Dolan who present average validities ranging from .32 to .50 (cited in Jones, 1991). Clearly, validity scores vary according to the time groups have spent together, the format of the peer assessment, and the definition of the peer assessment constructs.

Eagar (2004) conducted a study investigating the use of peer assessments within the OSB. This research recommended the use of a peer ranking system within syndicates and suggested a revision of the existing peer assessment exercise to include ratings of effort and persistence. This research also suggested that individual candidates should rate themselves independent of their peers using a social comparison ranking, in order to avoid leniency bias.

Candidates rank each member in their syndicate (each syndicate has 7 or 8 members) against three dimensions; likeability, leadership, and effort/persistence. These rankings are combined by the Psychology Clerk to give each candidate an overall rank within the syndicate for each dimension (Eagar, 2004). The effort/persistence ranking was recently introduced into the OSB in 2004, therefore, this study has utilised only the Likeability and Leadership rankings.

2.3 Criterion Measures

The definition of the criterion is one of the most important aspects of the selection system, and one of the most difficult. In this study performance in the junior officer role is the primary criterion of validation. However, the data for this criterion has been collated at least eighteen months after the predictor data, (in some cases up to ten years), additionally all participants have attended a rigorous year of training at OCS. Dover (1991) raises concerns about using a distant criterion, as there is occasionally an observed decline in validity the more distant the criterion is from the predictor. To help combat these concerns an interim criterion for training performance whilst at OCS has been added to the study. The use of an interim criterion is suggested by Dover provided there are linkages between the interim criterion and the predictor, and the interim criterion is representative of the final criterion, this suggestion is supported by Smith (1994).

Novick (1985) suggests that there are six key concerns to address when determining criterion measures in a validation study. Criterion measures need to be related to the purposes of the research and must be available for collection. The measures should represent important work behaviours or outputs, and be reliable. The possibility of bias in the criterion should be considered, and if it is necessary to combine criterion measures then a rationale is should be provided. Minimising these concerns will help to align the criterion measure to the performance domain.

2.3.1 Performance during Training

Criterion measures of training performance were collected from OCS (NZ) personal report files archived in Waiouru. These reports date between 1996 and 2004. Over this period of time there has been substantial changes made in the training conducted at OCS (NZ), and there exists a wide discrepancy in the type of reports filed into the personal report file. More recently OCS (NZ) has reverted to a more qualitative form of reporting, using word descriptions rather than standardised scores (Kearney, 2005). Despite this a number of assessments have remained the same throughout this period. In particular those assessments related to staff work, education and military studies, and tactical exercises. However, there is no way to determine the degree of variation

in marking leniency or harshness throughout this period. Additionally, problems exist between mapping the criterion to the predictor domain.

The data that was collected was collapsed into three key domains:

Academic Development This domain contained scores from Junior Staff Officer's packages, military writing and essays, and communication and professional studies.

Tactical Development This domain contained scores from tactical assessments called TEWT's (Tactical Exercise Without Troops), Operation Staff Duties, and from written appreciation assessments.

Practical Development This domain contained assessments from the practical and theoretical activities conducted to assess navigation and Radio Telephone Procedure (RATEL).

Finally, participants were given an overall training performance score. This composite criterion was derived from the three key domain scores. This is a similar method to that employed by Pynes and Bernardin (1989) in their validation of police selection.

2.3.2 Performance as a Junior Officer

Components of job performance are a function of declarative knowledge, procedural knowledge and motivation. Job performance is subject to individual differences and situational variation (Daft & Noe, 2001). Path analysis research conducted by Schmidt et al. (1986) on military data sets, found correlations of r =.30 between supervisor rating and knowledge, and r =.10 between supervisor rating and work sample performance.

Supervisor ratings are subjective criteria and are subject to a number of bias' including leniency, halo effects and range restriction errors (Tziner et al., 1994).

Tziner et al. (1994) also suggest that "standardised performance ratings formulated by supervisors with no knowledge whatsoever of their subordinates performance in the AC" (p. 229) should be used to help negate these biases.

Criterion measures of performance as a junior officer were collected from two sources, the MD68 (Appendix 4), and a Supervisor Rating Questionnaire (Appendix 5). The MD68 is an annual performance reporting document completed by junior officers on their birthday. Officers are graded on 31 different competencies, using a 9 point grading scale. It is then forwarded to their immediate supervisor for performance appraisal, and to the Commanding Officer for comment. After this is it sent to the Military Secretary Branch in Army General Staff, Wellington to assist with officer career management. The MD68 data was collected by the APS and merged into the participant database once all identifying information had been removed. Two sets of MD68 data were collected, the most recent held on file and the earliest held on file. The most recent MD68 data was collected in order to obtain current performance criteria (referred to as 'Time 1' data), whilst the earliest MD68 data was collected in order to collect data from the period directly post-commissioning (referred to as 'Time 2' data).

The second criterion measure was collected via the Supervisor Rating Questionnaire (Appendix 5). This form was developed with assistance from the Senior Psychologist (Army), and is based on the 12 competencies used for behaviourally rating candidates on the OSB. The rating form is accompanied by a behaviourally anchored rating scale in order to provide standardised grading of participants. This scale was adapted from the behaviour anchored rating sale provided to MTOs on the OSB. The Supervisor Rating Questionnaire also provided for the participant to be given an OAR.

2.4 Summary of Quantitative Information

The quantitative information available for data collection is summarised in Table 3.

Table 3.

Summary of Quantitative Information

	Source	Information	No. of Variables
Predictors	**OSB Database**	B90 score	
		Raven APM score	
		EPQ-R scores	4
		GPP-I	8
		Psych Ratings	4
		MTO Ratings	12
	Composite	Psych OAR	
		MTO OAR	
Criterion	**OCS Files**	Academic OAR	
		Tactics OAR	
		Practical OAR	
	Composite	OSC OAR	
	Supervisor Rating Form	Supervisor Ratings	12
		Supervisor OAR	
	MD68	Time 1	30
		Time 1 OAR	
		Time 2	30
		Time 2 OAR	

2.5 Method

The present research involved the collection of selection data, training performance data, and job performance data from Army archives, and the collection of performance data from immediate supervisors. In order to achieve this permission was first sought from the New Zealand Army to conduct the research. Approval was granted and is included in Appendix 6. Approval was then sought from the Massey University Human Ethics Committee and received on 25 July 2005 (Appendix 7). Thereafter individual consent was requested from both participants and supervisors to access and use their records. An article was published in the Army News newspaper advertising the research and encouraging participation. In addition, the Senior Psychologist Army distributed a letter to all Formation Commanders and Unit Commanders to raise awareness of the study within the camps throughout New Zealand.

All participants in the sample were mailed a research package that contained an information sheet and consent form (Appendix 8), a letter from the Deputy Chief of Army (DCA) endorsing the research (Appendix 9), and a return envelope. Twenty-two participants who were overseas at the time of distribution were emailed the documents in addition to being mailed a copy. Six weeks after distribution a reminder email was sent via the internal email by the Senior Psychologist Army requesting additional participation.

Participants were asked to read the information sheet and consent form, and to fill out and return the consent form. The consent form asked participants to provide consent to; access their OSB data, OCS (NZ) performance data, MD68 job performance data, collect supervisor performance ratings, and to retain their data for future validation purposes (by the APS). The form required them to either agree (tick 'yes') or decline (tick 'no') to provide consent for each option. The consent form was also used to collect additional data for use as control variables. This included information on participant age, ethnicity, time since commissioning, and number of deployments.

Once consent had been received, a package containing the; Supervisor Information and Consent Form (Appendix 10), Supervisor Rating Sheet and Rating Information

(Appendix 5), a letter from the DCA endorsing the research, and a return envelope were mailed to the participant's immediate supervisor. This allowed for the collection of supervisor ratings on the same competencies assessed for at the OSB.

Finally, data was collected from the various databases held by the APS, OCS (NZ), and the Military Secretary's Branch. Each applicant that attends the OSB completes the B90, Raven's APM, EPQ-R and the GPPI. Scores from these tests, in addition to gradings given by different elements of the board (MTOs, Syndicate Psychologists, Senior Board) are recorded onto database files held by the APS. Access to this data was provided by the APS. All MTO ratings and Psychologist ratings were translated into numerical values as follows:

Strong	=	5
Good	=	4
Adequate	=	3
Limited	=	2
Weak	=	1

Each officer cadet that attends OCS (NZ) has a personal report file that contains their performance reports and other personal information relevant to their posting at OCS (NZ). Training performance data was collected from these files, predominately in the form of test results that are administered throughout the training year. There were a wide variety of test results throughout the sample, with participants ranging from having 30% to nearly 90% of scores recorded in their files. Three composite scores were produced from these results (Academic, Tactical, and Practical) the criteria for each composite score are as follows:

Academic Development: There were 16 different scores; participants had to have at least eight scores in order to produce an Academic Development composite.

Tactical Development: There were 24 different scores; participants had to have at least 12 scores in order to produce a Tactical Development composite.

Practical Development: There were nine different scores; participants had to have at least four scores in order to produce a Practical Development composite.

These scores were then combined into an OCS OAR. Once commissioned, participant personnel reporting is conducted by Unit Commanders using annual reporting documents that are retained by the Military Secretary's Branch. This data was collected by an independent Trainee Psychologist from the APS, and merged into the participant database. Participant identifying details (name, initials, and date of birth) were then removed and the database was returned to the author. This allowed for current performance data to be collected confidentially. The supervisor ratings grades returned by mail were translated in the same manner as the MTO and Psychologist ratings.

2.5.1 Statistical Analyses

A power analysis was conducted to determine the generalisability of the results and to anticipate the population effect size and power of the research. The data will then be analysed to produce demographic results to determine the representativeness of the sample to the OSB population. Following this, the data will be screened for univariate and multivariate outliers, and checked for violations of the assumptions of multiple regression. Subsequently, the data will be analysed for correlations between predictor and criterion measures, as well as within these measures to detect the amount of covariance between measures and to determine if a linear relationship exists between predictor and criterion variables. To determine the unique contribution of AC components hierarchical multiple regression analysis will be conducted. The results from these analyses are presented in the following chapter.

2.5.2 Power Analysis

Statistical power analysis relies on the relationships of four variables involved in the statistical inference: Sample size (n), risk of Type 1 error, or significance criterion (a); population effect size (ES), and power. The relationship between these variables "are such that each is a function of the other three" (Cohen, 1993, p. 156), therefore, using three of these variables allows us to compute the fourth. The power of a study refers

to "the long term probability, given the population effect size, and sample size, of rejecting the null hypothesis" (Cohen, 1993, p. 156). Ideally, researchers aim to achieve a power statistic of .80 when working in a conventional setting (Cohen, 1992). Cohen also recommends that studies testing several hypothesis should set the significance criterion (a = .01) in order to reduce experiment-wise risk.

Population effect size refers to the discrepancy resultant between H_0 and H_1 when the null hypothesis is false. Population effect size is generically referred to as small, medium or large. The ES index for multiple and multiple partial correlations are; small f^2 = .02, medium f^2= .15, and large f^2 = .35. This is based on standard F test for $df = k, N - k - 1$ (Cohen, 1992). Therefore, we can determine from the power analysis that the necessary sample size to achieve a power statistic of .80, with a medium ES in a multiple correlation equation (when a = .05) is as follows:

2 independent variables (n = 67)

3 independent variables (n = 76)

4 independent variables (n = 84)

5 independent variables (n = 91)

6 independent variables (n = 97)

7 independent variables (n = 102)

8 independent variables (n = 107)

The type of statistic computed effects the power, as different statistics produce different effect size indexes. Additionally, the degree of range restriction and reliability of the criterion will impact on the power of a study (Novick, 1985). Given the research sample size (n = 72), it is crucial to attempt to minimise variables used within the correlational equations in order to maintain acceptable power in the study.

The sample size (n = 72) has limited the research in particular by reducing the number of variables permitted in the correlational equation to maintain acceptable statistical

power. Based on this reasoning the following variables were excluded from the analysis:

Raven's APM	(n=11)
Written Expression gradings	(n=20)
MD68 Time 2 Data	(n=26)

In addition to power, the sample size should meet requirements of generalisability. The ratio between the number of observations and the number of independent variables, know as the generalisability ratio, should not fall below 5:1 (Hair, Anderson, Tatham, & Black, 1998). This is to "avoid 'over fitting' the variate to the sample" (p. 166).

2.5.3 Data Screening

Four key assumptions must be met in order for the regression analysis to be conducted. These assumptions require data to be checked for the following (Hair et al., 1998):

- Normality of the error term distribution
- Linearity of the phenomenon
- Constant variance of the error terms
- Independence of the error terms

These assumptions involve the analysis of residuals to determine the degree of linearity, independence and homogeneity of variance. This allows us to check for any univariate or multivariate outliers, as both the unregressed data and the regression variate must meet these assumptions (Hair et al., 1998). Additionally, the assumptions require the data to be checked for multicollinearity, Hair et al, (1998) state that the "impact of multicollinearity is to reduce any single variable's predictive power by the extent to which it is associated with other independent variables" (p. 42). Therefore, as collinearity increases, the unique variance explained by each of the variables will decrease, resulting in a rise in shared prediction. The results from the data screening are presented in the Chapter 3.

2.5.4 Sample and Range Restriction

Hunter and Burke (1995) propose that a major concern with conducting a criterion related validation study is that usually data is only available for those who have met some qualifying standard. A sample becomes restricted when it consists of only those individuals who have meet some form of selection criteria, whether that criteria be a cut-off score on a test (range restriction), attendance on a course, or passing some for of pre-entry assessment (Sackett & Larson, 1990). This means that the full range of scores is not available for calculating correlations, causing the possible size of the correlation to decrease. This problem is only increased as applicants are followed through the training system (Hunter & Burke, 1995).

The sample range has been restricted in a variety of ways. Firstly, the role of an officer can be considered quite complex, and this complexity impacts on the type of applicant that applies. Salgado et al. (2003) state that "the selection process results in a greater range restriction for highly complex jobs than for lower complex jobs, because applicants tend to gravitate towards jobs with a complexity similar to their cognitive abilities and skills" (p. 1072). In addition, not all those that apply are recommended to attend the OSB, and some may not achieve the required standard on the GMA tests. So the sample range is already restricted compared to the general population.

Further to this, the sample was taken from current serving, commissioned officers, as there was no means of collecting data from those outside of the regular force. This resulted in additional sample restriction due to the following steps:

1. Participants must have passed the OSB.
2. Participants must have passed training at OCS (NZ).
3. Participants must still be serving.
4. Participants must have responded to the invitation to participate in the research.

In order to establish the impact of this sample restriction, two measures will be employed: Firstly, a descriptive comparative analysis will be conducted between data provided by the APS and the sample. Secondly, B90 scores will be corrected for range restriction and analysis results will be compared to uncorrected analysis to determine if the correction was necessary. These results are presented in the results section, in Chapter 3.

2.5.5 Criterion Attenuation

Criterion attenuation can occur though a variety of factors, generally it occurs due to rater error, or error variance. This results in a degree of statistical unreliability, which in turn lowers the validity coefficient and will result in an underestimation of the true predictive validity (King, Hunter, & Schmidt, 1980). Rothstein (1990) demonstrated that supervisor performance ratings frequently have low reliabilities, which serves to limit validity levels. It is recommended that studies using supervisory ratings as criterion measures should employ statistical techniques to correct for criterion unreliability. Therefore, Supervisor composite OAR scores and MD68 Overall Assessment scores will be corrected for criterion unreliability and analysis results will be compared to uncorrected analysis to determine if the correction was necessary. The results of this analysis are reported in the results section.

Chapter 3
Results

The data was analysed using SPSS 13.0. The results of this analysis are provided below.

3.1 Demographic Characteristics

The data was analysed to produce demographic results to determine the representativeness of the sample to the OSB population. The demographic characteristics are explained in terms of gender, age, ethnicity, and B90 score. These results are presented along with statistics comparing the sample to New Zealand Army Regular Force Officers.

The New Zealand Army has 4303 current serving Regular Force personnel. 689 (16.0%) of these personnel are officers, with 341 falling between the rank bracket of Second Lieutenant to Captain (49.5% of all officers and 7.9% of Army). Demographic comparison figures between the sample and the New Zealand Army Regular Force Officers are outlined in Table 4.

Table 4

Demographic Comparison Figures Between New Zealand Army Regular Force Officers and the Sample

	New Zealand Army	Sample
Average Age	37.15 yrs	27.46 yrs
Female	15.9%	20.8%
NZ European/ New Zealander	66.0%	77.8%
NZ Maori	7.8%	8.3%
Missing (ethnicity data)	15.4%	
n	689	72

(DSHHR, 2005; Harmer, 2006)

Three groups were used to conduct comparative data analysis of the sample; data was obtained from the APS. These groups were as follows:

OSB Total	Consists of all personnel who have applied for selection at the OSB (regardless of success or otherwise) over the period 1995-2004, less the Officer Population and the Sample (N = 1465).
Officer Population	Consists of all personnel who were invited to participate in this research but did not consent to participate (n = 111).
Sample	Consists of all personnel who consented to participate in this research (n = 72).

Females made up 20.8% of the Sample, compared to 21.6% of the Officer Population and 26.9% of the OSB Total population. A Pearson Chi-Square was conducted to determine if there were any significant differences between the groups for gender. The results, $x^2(2, N = 1465) = 3.240$, $p=.198$ were not significant, indicating that there are no significant differences across groups on gender.

The mean Age for the Sample at time of consent was 27.46 yrs, with 22.5% aged between 16-24 yrs. The mean Age of the Sample during attendance at the OSB was 20.67 yrs with 88.4% aged between 16-24 yrs, compared to the Officer Population (M = 20.05 yrs) and the OSB Total population (M = 21.58). An ANOVA (see Table 5) was conducted to determine if there were any significant differences between the groups for Age. The results indicated a significant difference between the groups. A post hoc analysis using the Scheffe test was conducted to determine where the difference lay. The results revealed that the Total Population was significantly older than the Officer Population, but was not significantly different from the Sample. These results indicate that there was more variability in the Total Population than in the Officer Population, and slightly more variability than in the Sample. This result may be due to range restriction in both the Officer population and the Sample.

Table 5

One-Way Analysis of Variance in ACER B90 Score and Age by Group

Source	*df*	*SS*	*MS*	*F*	*p*
			Age		
Between Groups	2	263	131	4.416	.012
Within Groups	1456	43344	30		
Total	1458	43607			
			B90 Result		
Between Groups	2	683	341	4.235	.015
Within Groups	1442	116275	81		
Total	1444	116958			

There was not enough ethnicity data available in each of the subgroups to conduct statistical comparison. However, the ethnic breakdown of the sample by gender is presented in Table 6 (n = 69).

Table 6.

Ethnic Breakdown of the Sample by Gender

	Male		Female	
Ethnic Group	Count	%	Count	%
New Zealander	45	81.8%	11	78.6%
Maori	4	7.3%	2	14.3%
European	3	5.5%	0	
Pacific Islander	2	3.6%	1	1.8%
Asian	1	1.8%	0	
Missing	3		0	
Total *n*	55		14	

The mean B90 score of the Sample was 40.21 (*sd* = 8.451, n = 70), compared to the Officer Population (*M* = 36.79, *sd* = 9.053, n = 1267) and the OSB Total population (*M* = 37.14, *sd* = 8.223, n = 109). An ANOVA (see Table 5) was conducted to determine if there were any significant differences between the groups for B90 score.

The results indicated a significant difference between the groups. A post hoc analysis using the Scheffe test was conducted to determine where the difference lay. The results revealed that the Total Population had significantly lower B90 scores than the Sample, but was not significantly different from the Officer Population. These results indicate that there was more variability in the Total Population than in the Sample, and slightly more variability than in the Officer Population. This result may be due to range restriction in both the Officer Population and the Sample.

The final demographic variables that were investigated were Operational Experience and Time Since Commissioning. There was a wide range of Operational Experience within the sample with 68.1% having been operationally deployed at least once. 40.3% had deployed on one operational deployment, 19.4% having had two operational deployments, and 8.4% having had three or more operational deployments. 31.9% of the sample had never been operationally deployed. The mean Time Since Commissioning was 4.67 yrs (*sd* = 2.72, n = 72).

3.2 Results of Data Screening

Prior to the main analyses, the data was screened and checked against the assumptions of multivariate analysis according to Tabachnick and Fidell (2001). The data was screened for accuracy of data entry, missing values and distribution fit. This involved examining the data for normal distribution, and checking the skewness and kurtosis levels present in each of the variables. This process also enabled the detection of any univariate outliers. Eight of the variables were outside the bounds of three times the standard error for kurtosis and skew. No univariate outliers were detected.

Peer Assessment ratings for Likeability and Leadership were both positively skewed. Square root transformation was able to markedly improve skewness. Current age and EPQ-R Psychoticism scale both showed signs of positive kurtosis, with EPQ-R Psychoticism also being positively skewed. Logarithmic transformation was able to reduce both the kurtosis and skew of these variables. When descriptive statistics are provided for these variables, untransformed means and standard deviations are reported for ease of interpretation. GPP-I Responsibility, and three of the MTO

ratings showed signs of positive kurtosis, however they were left untransformed as transformation did not alter the results.

The data was checked for independence of the error terms using a correlation matrix. No bivariate correlation of predictors exceeded r = .9 (Tabachnick & Fidell, 2001, p. 83). Next the data was checked for linearity. Tabachnick and Fidell state, "the assumption of linearity if that there is a straight-line relationship between two variables (where one or both of the variables can be combinations of several variables)." (p. 77). Data screening was conducted using partial residual plots to detect a straight-line relationship. These plots indicated that no linear relationship existed between B90 score and predictor scores (Psychologist rating, Psychologist OAR, MTO ratings or MTO OAR) or between B90 score and any criterion scores (OCS composite scores, OCS OAR, Supervisor ratings, Supervisor composite OAR, MD68 Overall Assessment). This lack of relationship violates the assumption of linearity, and will be discussed further later in this chapter.

The data was screened for multivariate outliers using Mahalanobis distance. A regression analysis was conducted using $p< .001$ and critical values of Chi Square taken from Tabachnick and Fidell (2001, p. 933). No cases were identified as outliers. The data was also checked for multicollinearity by ensuring that tolerance levels in the regression analysis were not less than .1 (Tabachnick & Fidell, 2001, p. 84).

3.3 Results of Range Restriction Analysis

Only some of the necessary information required to conduct corrections for the effects of range restriction was available (Dick, 1993). The information required to conduct range restriction corrections for the criterion (due to the effects of turnover and dropout) was not available. However, it was possible to conduct analysis on the predictor variable 'B90 score'. The B90 is employed on the OSB as a pre-selection tool, in that applicants need to obtain a cut-off score in order to progress. In most instances this score is set at 24, however there are mitigating circumstances that have allowed applicants scoring below this cut-off to progress further. McNemar's formula (equation 1) was used to make corrections to investigate the effects incurred by the cut off scores for the B90 test. This correction formula relies on two main

assumptions being met; linearity and homoscedasticity between the error terms in the restricted sample and the unrestricted population (Sackett, 2000).

$$r_c = \frac{r_u\left(\frac{S_x}{s_x}\right)}{\sqrt{1 - r_u^2 + r_u^2\left(\frac{S_x}{s_x}\right)^2}}$$

Where: r_c = the corrected correlation
r_u = the uncorrected correlation
S_x = the unrestricted standard deviation
s_x = the restricted standard deviation (1)

The unrestricted range for the B90 is 0-70; by using a cut-off of 24 the test range has been restricted to 24-70. The unrestricted standard deviation used for the Total Population and the Sample were $S_x = 9.053$ and $s_x = 8.757$ respectively. The corrections for range restriction resulted in very small improvements in the bivariate correlations. These are displayed in Table 7.

Table 7

McNemar's Formula Corrections for ACER B90 Range Restriction on Main Variables

	Uncorrected Correlation	Corrected Correlation
MTO OAR	-.089	-.092
Supervisor Composite OAR	.029	.030
OCS OAR	-.031	-.032
MD68 Overall Assessment	.028	.029
MTO Motivation	-.294*	-.303
Psych Education	.189	.195
Peer Assessment Likeability	.163	.168
Peer Assessment Leadership	.179	.185
MTO Planning Ability	.213	.220

*$p<.05$.

3.4 Results of Criterion Attenuation Analysis

Bivariate correlations between criterion scores revealed a lack of inter-rater reliability on performance. Correlations between Time 1 and Time 2 MD68 scores on 'Attitude to further service', 'Assessment of future potential', and 'Overall Assessment' revealed moderate significant correlations, but nothing greater that r (46) = .649, ($p<.05$). The correlations between Supervisor composite OAR and MD68 Overall Assessment ratings were also moderate ($r(45) = .426$, $p<.01$) suggesting that there is a lack of reliability between the two ratings or that they are in fact capturing different construct domains.

Ghiselli, Campbell and Zedeck (1981) suggest that the degree of unreliability with which a variable is measured serves to veil the true relationship of the variable with other variables. To counter this unreliability, Supervisor composite OARs and MD68 Overall Assessment scores were corrected for criterion unreliability using a correction for attenuation formula (equation 2).

$$r_{x^{\infty}y^{\infty}} = \frac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}} \qquad (2)$$

Where:
- $r_{x^{\infty}y^{\infty}}$ = the corrected correlation
- r_{xy} = the uncorrected correlation
- r_{xx} = the reliability of the predictor
- r_{yy} = the reliability of the criterion

This formula allows us to "correct for the degree to which the correlation between two variables is attenuated by their unreliability" (Ghiselli et al., 1981, p. 241). Because we are concerned with criterion reliability, rather than predictor reliability, we treat the reliability of the predictor as perfect, giving it a value of 1, effectively removing it from the equation. As reliability data was not available for New Zealand Army supervisor ratings on the MD68 or the Supervisor Rating Sheet, an average inter-judge reliability estimate (r =.52) was employed (Rothstein, 1990). The corrections

for criterion unreliability resulted in very small improvements in the bivariate correlations. These are displayed in Table 8 below.

Table 8

Formula Corrections for Criterion Unreliability on Main Variables

Main Variable	Uncorrected Correlation	Corrected Correlation
	Supervisor Composite OAR	
ACER B90	.029	.040
Psych OAR	.072	.099
Peer Assessment Likeability	-.046	-.064
Peer Assessment Leadership	.072	.099
	MD68 Overall Assessment	
ACER B90	.028	.040
Psych OAR	-.022	-.031
Peer Assessment Likeability	-.015	-.021
Peer Assessment Leadership	-.003	-.004

*$p<.05$.

3.5 Descriptive Statistics

Sample descriptive statistics including means and standard deviations are presented for the main variables in Table 9.

3.6 Predictor and Criterion Correlations

The data will be analysed for correlations between predictor and criterion measures, as well as within these measures to detect the amount of covariance between measures and to determine if a linear relationship existed between predictor and criterion variables. There were no significant correlations between B90 score and any of the criterion variables (see Table 10). Additionally, the Supervisor composite OAR did not correlate significantly with Psychologist OAR, MTO OAR, or any of the individual MTO ratings. However, further investigation into the bivariate correlations

between individual MTO ratings and Supervisor composite OAR elements did reveal some significant relationships. These are reported in Statistical Appendix 1.

Table 9

Means, Standard Deviations and Case Numbers for Main Variables

Main Variable	Mean	Std dev.	*n*
		Predictors	
B90	40.21	8.45	70
Psych Intellect	3.34	0.88	72
Psych Education	3.50	0.86	72
Psych Temperament	2.71	.69	68
Psych Motivation	3.23	.64	71
Psych OAR	12.77	1.64	72
MTO OAR	38.34	4.38	72
Peer Assessment Like	3.22	1.93	69
Peer Assessment Lead	3.00	1.83	69
		Criterion	
OCS OAR	78.67	4.12	52
Supervisor composite OAR	45.36	6.83	64
MD68 Overall Assessment	6.12	1.05	49

3.7 Multiple Regression Analyses

Tabachnick and Fidell (2001) suggest some general considerations for choosing variables for regression equations. "Regression will be best when each Independent Variable (IV) is strongly correlated with the Dependant Variable (DV) but uncorrelated with other IVs. A general goal of regression, then, is to identify the fewest IVs necessary to predict a DV where each IV predicts a substantial and independent segment of the variability in the DV" (p. 116). As this research was theory driven, it was necessary to determine a bivariate linear relationship with B90 scores (as a measure of GMA) and criterion measures to progress towards multiple

Table 10

Intercorrelations Between Main Variables for Sample

Subscale	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. B90 Score	_	.890**	.189	-.082	-.145	.475**	.156	.146	.104	-.032	-.089	-.031	.029	.028
2. Psych Intellect		_	.195	-.026	-.046	.605**	.162	.188	.133	-.081	-.087	-.039	.001	.030
3. Psych Education			_	.206	-.084	.655**	.037	.297*	.137	-.124	.102	.131	.172	.072
4. Psych Temperament				_	.057	.529**	.081	.279*	-.006	.001	.046	-.117	-.034	.087
5. Psych Motivation					_	.329**	.029	-.092	.053	.090	.254*	.115	.012	-.267
6. Psych OAR						_	.147	.324**	.133	-.095	.126	.043	.072	-.022
7. Peer Assessment Likeability							_	.593**	.003	-.208	-.068	.117	-.046	-.015
8. Peer Assessment Leadership								_	.163	.061	-.327**	.021	.072	-.003
9. Time since Commissioning									_	.589**	-.195	-.087	.381**	.230
10. Number of Deployments										_	-.096	-.091	.261*	.144
11. MTO OAR											_	.159	-.108	-.009
12. OCS OAR												_	-.018	1.46
13. Supervisor Composite OAR													_	.436**
14. MD68 Overall Assessment														_

Note: Square root transformations are used for the variables Peer Assessment Likeability and Peer Assessment Leadership
*p<.05. ** p<.01.

regression analysis. Due to the lack of empirical evidence to support this relationship, the regression analyses for hypotheses one through to five were not conducted.

3.8 Hypothesis Six

A Pearson correlation matrix was produced to determine whether a positive relationship existed between elements of the EPQ-R and both training performance and job performance. The four scales of Psychoticism, Extraversion, Neuroticism and Lie were correlated with the variables OCS OAR, MD68 Overall Assessment and Supervisor composite OAR. These correlations are reported in Table 11. There were no significant correlations.

3.9 Hypothesis Seven

A second Pearson correlation analysis was conducted to determine whether a positive relationship existed between elements of the GPP-I and both training performance and job performance. The eight scales of Ascendancy, Responsibility, Emotional Stability, Sociability, Cautiousness, Original Thinking, Personal Relations and Vigour were correlated with the variables OCS OAR, MD68 Overall Assessment and Supervisor composite OAR. These correlations are also reported in Table 11. There were no significant correlations.

3.10 Hypothesis Eight

Hypothesis eight investigated whether increased Time Since Commissioning was positively correlated with higher job performance. The correlations are presented in Table 10. The number of deployments correlated moderately with time since commissioning. This is to be expected since the longer a member serves in the armed forces, the more likely it is that they will deploy on operations. Time since commissioning and number of deployments also produced small to moderate correlations with Supervisor composite OAR.

Further investigation into the elements of the MD68 revealed significant correlations between Time Since Commissioning and scores on seven of the MD68 elements,

Table 11

Intercorrelations Between Subscales of the EPQ-R and GPP-I for Training Performance and Job Performance

Subscale	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. OCS OAR	_	-.018	-.142	-.147	.097	.082	.028	.170	-.114	.030	.201	.096	-.033	.002
2. Supervisor composite OAR		_	-.053	.171	.120	.097	.043	.036	-.171	.072	-.006	-.074	.006	.155
3. EPQ-R Psychoticism			_	.374**	-.076	-.056	.216	-.154	-.074	.193	-.145	-.097	.134	-.102
4. EPQ-R Extroversion				_	-.029	-.239*	.221	-.111	-.342**	.472**	-.407**	-.142	-.024	.350**
5. EPQ-R Neuroticism					_	-.267*	-.062	-.153	-.508**	.209	.100	-.214	-.258*	-.051
6. EPQ-R Lie Scale						_	-.106	.151	.185	-.209	.217	.115	.027	.008
7. GPP-I Ascendancy							_	-.628**	-.261*	.351**	-.397**	.292*	-.117	.110
8. GPP-I Responsibility								_	.222	-.399**	.284*	.004	-.170	.328**
9. GPP-I Emotional Stability									_	-.620**	.164	.152	.229	-.173
10. GPP-I Sociability										_	-.201	.046	.108	.026
11. GPP-I Cautiousness											_	-.082	.156	-.125
12. GPP-I Original Thinking												_	.018	.107
13. GPP-I Personal Relations													_	-.354**
14. GPP-I Vigour														_

*p<.05. ** p<.01.

ranging from Power of written expression ($r(52)=.283$, $p<.05$) to Oral Expression ($r(52)=.378$, $p<.01$). Additionally, investigation into the elements of the Supervisor composite OAR revealed significant correlations between Time since commissioning and scores on seven of twelve competencies, these correlations are presented in Statistical Appendix 2.

3.11 Additional Analysis

Due to the lack of ability of the GMA measure (B90 score) to predict subsequent job performance, an investigative discriminant analysis was conducted to determine if there was a combination of variables that were able to predict group high and low performers in the criterion Supervisor composite OAR. In order to achieve this, a new criterion variable was created from the Supervisor composite OAR. This dichotomous variable separated job performance into 'High' and 'Low' performance categories. High performance was indicated by a Supervisor composite OAR of 44.5 or greater, whereas low performance was indicated by a Supervisor composite OAR of 44.4 or lower.

The predictor variables selected for entry into the discriminant analysis were MTO ratings of; Oral Expression, Planning Ability, Practical Ability, Determination, and Motivation, and Psychologist Education ratings. These variables were selected as they had exhibited at least one significant correlation ($p<.05$) with one or more of the elements of the Supervisor composite OAR. From this analysis 64 cases were determined in which both predictor and Supervisor composite OAR data were available. It was found that no linear composite of the predictors was able to predict high or low performance categories on Supervisor composite OAR at a statistically significant level ($p< .05$).

Chapter 4
Discussion

The purpose of this research was to evaluate the incremental validity of the measures employed on the Army Officer Selection Board over and above measures of General Mental Ability. The results have important implications for the use of the OSB and its approach to the selection of candidates. The following discussion aims to interpret and examine the results.

The use of the Assessment Centre (AC) as a valid and predictive selection tool for future training and job performance is strongly supported in the literature, particularly in the military environment. Schmidt and Hunter (2004) found that the overall average predictive validity for job performance was .55, with an average predictive validity of .65 for training performance. Ree and Earles (1991) support these findings, reporting AC predictive validities of .60 for Air Force pilot trainees. However, Schmidt and Hunter (2004) suggest that the AC provides limited incremental validity (.01) over GMA for job performance.

4.1 Predictive Validity of the B90

In order to evaluate incremental validity of the measures employed on the OSB over and above measures of GMA, it was critical to first establish a level of predictive validity between the test of GMA employed, and the criterion measures. The results determined that there was no significant bivariate correlation between the B90 and the Supervisor composite rating. In addition to this, the B90 returned no significant correlations with any of the individual Supervisor ratings, or ratings of training performance gained at OCS (NZ). Overall, there was a lack of evidence to support any significant relationships between the B90 and any of the criterion measures.

This result is surprising considering the results reported in previous international military studies of training performance and job performance (Anderson et al., 2001; Braun et al., 1991; Lievens et al., 2003; Ree & Earles, 1991; Schmidt & Hunter, 1998), which provide evidence supporting the use of measures of GMA as predictors

of both training performance and job performance. Since there is a multitude of research that supports the predictive ability of GMA for both training and job performance, it is more than likely that some error of measurement exists with either the predictor or criterion.

Considering that New Zealand military research conducted by Ratcliffe (2005) found only small significant correlations between B90 score and criterion measures of Leadership (r=.18, p <.05) and Adaptability (r=-.23, p <.05), these results may indicate an inability of the B90 to accurately measure GMA. Furthermore, the B90 has only shown small bivariate correlations with OSB success (Gracie, 2004), suggesting that other elements of the OSB have a greater impact on candidate selection.

4.2 Predictive Validity of MTO Ratings, Psychologist Ratings and Peer Assessments

The aim of Hypotheses 1 through to 5 was to individually investigate the incremental validity of the following predictor measures over and above that of the GMA test employed (the B90):

a. MTO combined ratings,
b. MTO individual ratings,
c. Psychologists combined ratings,
d. Psychologists individual ratings, and
e. Peer Assessment ratings of Likeability and Leadership.

These ratings are important components of the OSB AC and are the basis on which decisions for suitability for training at OCS (NZ) are determined. However, due to the lack of predictive validity of the B90, there was no empirical basis on which to conduct this analysis.

Past validation studies have shown that dimension based AC are effective selection tools when based on a thorough job analysis (Jacobs, 1986). The use of peer assessment rating and supervisor ratings is supported by Tziner et al. (1994), with high correlations reported between the two predictors (Harris and Schaubroeck,

1998). Tziner et al. (1993) used managers ratings as a predictor measurement to determine management potential and management performance. Over a four year period they found predictive validities ranging between .33 and .56 for management potential, and .20 to .21 for management performance, with managers ratings correlating higher than psychologists ratings in most instances.

Research conducted by Harrison (2005) with RNZN Officer Cadets found that observer ratings were able to predict training performance, with the AC OAR correlating significantly with the TCS (r=.42). In addition, the literature on Peer Assessments shows strong support for their use in the AC. Despite this there were no significant correlations between Peer Assessment ratings of Likeability or Leadership and any of the criterion measures.

4.3 Personality, Training Performance and Job Performance

Hypotheses 6 and 7 investigated the relationship between measures of personality and performance, both in training and on the job. Pearson's correlations were conducted to determine the degree to which elements of both the EPQ-R and the GPP-I might be correlated with Supervisor composite OAR scores and OCS OAR scores. The results reveal no significant relationships on any of the scales. These results similarly suggest that personality measures employed on the OSB are unable to predict training performance and subsequent job performance.

This result is contrary to research (Dayan et al., 2002; Ree et al., 2001; Salgado et al., 2001) supporting personality traits, in particular conscientiousness, as valid predictors of future performance. Lievens et al. (2003) found a significant relationship between openness and instructor ratings of job performance (r=.31), however no significant correlations were found between instructor ratings and extraversion and conscientiousness scores.

These results may also have been affected by the degree of range restriction present in the sample as discussed previously in the method section. It is possible that due to selection criteria and turnover effects, variations in personality type may have been reduced, producing a degree of homogeneity in the sample group. It may also be due

to the method in which personality tests are used on the OSB. On the OSB, personality measures are aimed to target any elements of personality that are considered undesirable, rather than to identify elements of conscientiousness.

4.4 Time Since Commissioning and Job Performance

Hypothesis 8 investigated the relationship between Time Since Commissioning and job performance. The results revealed a small to moderate relationship ($r=.381$, $p<.01$) with Supervisor composite OAR, however there was no significant relationship between Time Since Commissioning and MD68 Overall Assessment. Since Supervisor composite OAR and MD68 Overall Assessment have a moderate significant correlation ($r=.436$, $p<.05$), this result suggests that although the two criterion measures do overlap to some degree, they are also measuring unique performance domains independent of each other, and Time Since Commissioning has more influence on the domain measured by Supervisor composite OAR, than that measured by MD68 Overall Assessment.

These results are expected as they indicate that the length of time you have been conducting a job role, or the level of experience in that job role, affects job performance. It also suggests that increased length of time in a job leads to increased experience, which in turn leads to better job performance. However the positive correlation may also reflect a degree of rater bias. It may be that officers who have been commissioned for longer periods of time are better known by their superiors as they have had a longer working relationship, reflecting factors of familiarity (Brumback, 1969) and these factors have caused a halo effect to occur in supervisor ratings.

4.5 Subsequent Analysis

Subsequent investigations revealed no linear combination of predictor measures was able to predict high or low job performance as measured by the Supervisor composite OAR, and that there were no significant correlations between the criterion measures of training performance (OCS OAR) and job performance (Supervisor composite

OAR). This result is similar to that reported by Kelly (1994) in her previous validation of the OSB.

The lack of any significant relationship between the remaining predictor variables (other than the B90) and the criterion measures suggests that the predictor measures employed on the OSB to select officer candidates are not related to those attributes that define subsequent performance as an officer. However, it could also reinforce the stance that there is a significant degree of measurement error present in the study.

Similarly, the lack of any significant correlations between criterion measures of training performance and job performance suggests that high performance at training does not equate to high performance on the job. This is at odds with training objectives, as the primary focus of training at OCS (NZ) is to produce good officers, not good Officer Cadets. However, it is more likely that the result is a reflection of the competency domains that were used to measure training performance and job performance.

There were a number of significant bivariate correlations found between the competencies making up the elements of the MTO OAR and the competencies making up the elements of the Supervisor composite OAR. These relationships are discussed in Statistical Appendix 1.

4.6 Methodological Effects on the Data Obtained

The lack of empirical evidence supporting the predictive validity of the OSB is surprising given the relative face validity of the OSB. At face value the results suggest that the OSB fails to accurately predict both training and future job performance, and hence the use of the OSB for the selection of candidates as officers is not empirically supported. However, there are many theoretical and practical explanations that need to be explored before such conclusions should be drawn from these results. These explanations fall into four generic categories of factors relating to the:

a. Methodology of the study
b. Problems with predictor measurement
c. Problems with criterion measurement, and

d. Methodology of the selection system.

These topics are discussed in the remainder of this section.

4.6.1 Methodology of the Study

The two main methodological factors affecting the study involved sample size and range restriction. Sackett and Larsen (1990) pose the question; "under what conditions can valid generalisations be drawn about a population solely on the basis of sample observations?" (p. 431). As was seen in the power analysis (see method section) the sample size (n = 72) restricted the level of analysis that was possible. With such a small sample size and response rate (39.3% of the total population) caution must be used in interpreting the results from the study, hence given a greater sample size; more definitive conclusions could be made from these results.

More importantly, however, the small sample size resulted in quite severe range restriction, therefore, the results from this sample may not be generalisable to the total OSB population. Tabachnick and Fidell (2001) state that where there is restricted range in a sample, "sample correlations may be lower than population correlation" (p. 57). This range restriction was evident to a small degree in the sample comparisons, which were conducted based solely on B90 score. Had additional predictor or criterion related data been available for remaining population it is anticipated that variances between the populations would have been far greater.

4.6.2 Problems with Predictor Measurement

There are several levels of predictor measurement where measurement error may have been introduced in this study. Measurement error is observed when a given measure, whether it be a psychometric test, observer rating, or some other score, does not accurately reflect the construct domain it is supposed to (Klem, 2003; Licht, 2003). Measurement error is a reflection of the degree of construct validity present in the selection system.

The construct validity of a measure is a particularly important element to consider in validation studies. In a validation study we want to firstly ensure that the predictor

domain and the criterion domain are congruent, and secondly, ensure that the predictor measure is accurately measuring the predictor domain, and that the criterion measure is accurately measuring the criterion domain (Smith, 1994). Only then can we be confident that resultant associations between scores on a predictor measure and score on a criterion measure are due to some kind of relationship between the predictor construct and criterion construct.

The most prominent predictor concerns for this research involve the B90 as a measure of GMA. One of the fundamental assumptions of the study was that the B90, as a measure of GMA, would have a significant positive relationship with measures of training and job performance. The fact that no such relationship existed, begs the question of whether the B90 is able to accurately measure the domain of cognitive ability. Reliability coefficients of the B90 and Raven's APM (r=.49, p=.05) (Gracie, 2004) suggest that the B90 is only partially tapping into the domain of GMA, and thus has limited capability to discriminating between high and low GMA. This is supported by Gracie's claim that the B90 measures crystallised intelligence rather than fluid intelligence.

Secondly, we want to consider the selection competencies at the foundation of the OSB. Selection competencies should be based on a thorough job analysis, identifying the necessary KASOs required to perform in the job. Theoretically speaking, competencies should have the ability to capture the performance domain. Practically speaking this implies that the competencies identified for the OSB, should be those KASOs necessary to perform successfully as an officer, however, this does not take into consideration the element of training necessary to produce effective officers. Therefore, OSB competencies should be identified as those KASOs necessary to complete officer training at OCS (NZ). The logical progression then is that OCS (NZ) training objectives should be identified as developing the KASOs necessary to perform as an officer.

The lack of any significant relationship between the OSB competencies and both OCS OAR and MD68 Overall Assessment may be due to inconsistent alignment between the predictor domain and the criterion domains. However, it is unlikely that a lack of any significant relationship between the OSB competencies and the Supervisor

composite OAR is due to a misalignment of the construct domains, as both measures used the same competencies and very similar BARS. This does not eliminate the possibility that the competencies were not measured correctly in the first place.

Measurement error as a result of assessor bias or lack of observation training is common (Riggio, 1996). Turner (1988) states that "if the basis of assessment has not been clearly or adequately defined or is misunderstood, valid assessment is unlikely" (p. 12). Although MTOs are provided with assessor training prior to the assessment phase of the OSB, this training is only for a two-hour period and relies heavily on experiential factors. Given the wide variety of experience present in any given group of assessors, it is logical to assume that a wide range of perceptions of successful performance might also exist.

Furthermore, literature suggests that the number of competencies employed effects the quality of assessment (Arthur et al., 2000; Bycio et al., 1987; Gaugler & Thornton, 1989; Lievens & Conway, 2001). The use of twelve competencies is likely to result in cognitive overload, causing MTO to revert to pre-held biases for assessment. Combined with a variation in assessor opinion, these factors allow for the introduction of measurement error in gradings.

4.6.3 Problems with Criterion Measurement

Hunter and Burke (1995) proposes a low level of prediction may indicate problems in criterion measurement rather than a lack of validity with the predictor. There were three measures of criterion used in this research, OCS (NZ) performance, Supervisor ratings, and MD69 Annual Reports. Each of these measures is open to its' own form of measurement error, and is discussed below.

The degree of measurement error present in the data obtained from OCS (NZ) was a potentially limiting factor in gathering training performance criteria. This concerned the consistency of grades given to candidates whilst at OCS (NZ). Performance data was collected from archived paper files rather than electronic file because as each class graduates, electronic performance records are deleted, resulting in a wide discrepancy in information available. Additional inconsistency is introduced as staff

are rotated on posting, consequently it was not possible to determine the degree of measurement error present in these ratings.

There was a notable lack of standardisation in the manner in which performance results were recorded onto personnel files, resulting in a lack of consistent reporting. This is to be expected as the data was being obtained from a large time period, however, it resulted in a large amount of missing training performance data. Therefore, interpretations made from the training performance data in this research should be made with caution.

The final limiting factor involving OCS (NZ) performance records concerns the disparity between the measurement of the construct domains. Criterion information was collected from three key variable domains (Academic, Tactical, and Practical), which focused predominantly on external performance functions. In contrast, the predictor information focuses on internal performance functions, in particular personality oriented behaviours.

The second criterion measurement used in this research was the supervisor rating provided by the participant's immediate supervisor. As discussed previously, unlike the OCS OAR this criterion was aligned to the predictor construct domain through the use of the same grading criteria and scales. Despite this, supervisor rating are still susceptible to measurement error.

Supervisor ratings were generally based on a longer period of observation of job performance than MTO ratings. Additionally, given the nature of the officer role, most supervisors would have a close working relationship with their subordinates. This may have introduced an element of assessor bias into the ratings. Unlike the MTOs, supervisors were not given any form of behaviour observation training on the assessment criteria prior to rating subordinates, although MTOs have been exposed to observation training during promotional courses this training, it may have been some time since the training was last conducted. Additionally, supervisors were generally junior in rank to the majority of the MTOs, indicating that they are likely to have less experience in both supervision and assessment. These two factors may have served to reduce the reliability of the supervisor ratings.

The final criterion used was the MD68 Overall Assessment grading. Similar to OCS OAR and the Supervisor composite OAR, this criterion was susceptible to the effects of measurement error through both a misalignment of the construct domain measured, and through assessor biases and lack of training. The resultant correlation between MD68 Overall Assessment and Supervisor composite OAR ($r = .436$, $p=<.01$), does imply that they are measuring the same domain to some degree, however, it also suggests that there is a large amount of disparity between the two performance domains. This is direct reflection of the military performance appraisal system not being aligned with the selection system[2].

There were a number of confounding variables that could not be controlled for that may have contributed to the results of the study. Data was collected on 'time since commissioning', and 'number of deployments' to help mitigate these effects. However, it was beyond the scope of this research to be able to control for many factors relating to post commissioning experience. Examples of these factors include; the nature of operational deployment, the degree of command experienced, or the Corps that a candidate graduated to.

The nature of operational deployment and the number of operational deployments experienced varies widely within the sample, as each deployment has differing levels of command responsibility and task outputs unique to that theatre of operations. Similarly, the level of command exposure, and therefore leadership opportunity each participant has experienced, differs. This can be a result of unit size (and therefore number of command positions available), staff positions required to be filled, or capability.

It is expected that the Corps that a participant graduated into may have been a confounding variable in this study. When a participant graduates from OCS (NZ), they do so into a particular Corps (e.g. Infantry, Artillery, Engineers, Signals, Logistics). Participants are then provided a range of specialist training within their

[2] This is based on MD68 information obtained from past records pertaining to the sample. The NZDF is currently trialing a new performance appraisal system that is aligned to the NZDF Competency Framework.

Corps through various promotional and Corps specific courses. Additionally, each Corps has its own traditions, areas of expertise, and values that is adheres to. It stands to reason then, that whilst all Corps heed to certain standards of 'officer performance', they also have their own criteria against which high performance is determined. For example, an important element of an infantry officer's performance is physical fitness, whereas, for a logistics officer staff work and accounting may be more essential.

In contrast, all candidates are treated equally at OCS (NZ) (Corps are not allocated until the final weeks of the commissioning course) and their performance is graded against a single standard based on an infantry Corps. Whilst all supervisors were provided BARS to rate their subordinates, it is logical to assume that they did so based on inherent standards already present within their own Corps, introducing a confounding variable. Additionally, ratings may also have been affected by stereotypical factors. Brumback (1969) states that "it is known that certain characteristics of the ratee, even though they are extrinsic to the behaviour being rated, do influence ratings of that behaviour" (p.40).

A final factor for consideration is the degree of criterion contamination that may have been introduced into the study. Criterion contamination refers to a transfer of information between predictor assessment and criterion assessment resulting in a lack of independence between ratings (Sackett, 1987). Criterion contamination is avoided in one respect in that the results of the OSB are not made available to either training staff at OCS (NZ), or to the candidate's supervisor, however, a candidate's performance at OCS (NZ) is transferable as supervisors are privy to end of course performance reports. Whilst this may have had little effects on supervisor ratings provided on participants who have been commissioned for a longer period of time (e.g. 3-5 years), it may have biased job performance ratings for recent graduates (e.g. less than two years).

4.6.4 Methodology of the Selection System

There are two elements of the OSB methodology that may have affected the results of the study; these are the alignment of the selection system to performance appraisal

systems, and the selection principle employed. The OSB uses 16 competencies (twelve for MTO observations, and four for Syndicate Psychologist observations) against which selection decisions are made. However, these competencies are not used as assessment criteria for further performance appraisal, either in training (at OCS (NZ)) or on the job after commissioning (MD68). As the three systems are not aligned it is difficult to ensure that information pertaining to the predictor domain is in fact the same information as what is being collected in the criterion domain.

There currently exists a major project within the New Zealand Defence Force (NZDF) to ensure the alignment of military Human Resource systems with a recently developed framework called the NZDF Competency Framework. However, this framework is still under implementation and there is no current alignment of OSB competencies to OCS (NZ) performance criteria, and subsequent job performance criteria.

Additionally, as discussed previously, the OSB adheres to a principle of negative selection, whereby the objective of the OSB is to identify those who possess the minimum standard (or above) for training at OCS (NZ), rather than identifying the top candidates applying. Although it is possible to rank candidates on their MTO grades, it is worth considering that the prime focus of the MTO grading system is to determine if an applicant is "adequate" in selected criteria. Hence, the final grades provided on candidate performance at the OSB are simply classified as "Recommended" or "Not recommended".

Supervisor ratings on the other hand are not based on a principle of negative selection and are often used to help provide feedback to subordinates. Therefore, the use of a negative selection system may serve to limit the relationship between MTO ratings and supervisor ratings on performance.

Chapter 5
Summary and Conclusions

Organisational fit is an important element of organisational productivity, employing personnel with poor organisational fit can be detrimental to an organisation and be costly in terms of training and turnover (Ratcliffe, 2005). Therefore, it is essential to ensure that the measures employed by the New Zealand Army in their officer selection are both valid and reliable. At face value the OSB appears to provide adequate selection validity and utility, however, the results of this study suggest that the selections system is not able to predict future performance in either training or in subsequent employment in the officer role.

The results show no significant relationship between the B90 and future training performance and job performance and the theoretical and practical implications of these results are discussed. It is proposed that the results are likely to be resultant of both severe range restriction effects and measurement error. A number of elements within the criteria of this study that are susceptible to measurement error were identified. The most prominent of these concerns the ability of the criterion measures to capture the training and job performance construct domains, in order to ensure congruency between predictor and criterion measurement. The elements of the selection system most susceptible to measurement error are the MTO ratings on assessment competencies.

There are a number of limitations identified with the conduct of this study. The most obvious relates to sample size and these resultant effects on the range restriction of the sample, and the quality of quantitative performance information available at OCS (NZ), which limited the criterion and introduced elements of measurement error.

Despite these factors, this research has identified a number of key issues that should be investigated within the current New Zealand Army officer selection process. The first of which is to investigate the current competencies used within the OSB to determine if they are still representing the essential characteristics of the job analysis. Secondly, the study identified key discrepancies between current selection

competencies and the competencies used to measure performance whilst under training and subsequent performance post-commissioning. The research also identified a potential inability of the ACER B90 to accurately measure the domain of GMA, thus questioning its use in the selection system.

The current research also highlights a number of issues involving the use of the current performance appraisal document, the MD68, and suggests that it may be open to a number of subjective biases. This suggests that caution might be necessary for future research that uses such criterion. Additionally, where observer ratings are employed as predictor or criterion measures, thorough observation training and BARS should be employed.

Finally, readers are reminded that the results of this research should be interpreted with caution due to a number of methodological concerns that have been identified with the sample size, range restriction, and measurement error.

Chapter 6
Recommendations

On the basis of the results and methodology concerns identified in this study the following recommendations are proposed to the Army Psychology Service to improve the OSB:

1. *Job Analysis* Conduct a thorough job analysis of the Officer Cadet role and revisit the junior officer job analysis to determine the job characteristics and essential personal characteristics associated with good performance. This will help determine whether the competencies identified with good training performance are those competencies that are identified with good job performance.

2. *Alignment of Competencies* The competencies across selection, training and job performance need to be aligned to ensure consistency. In particular OCS (NZ) should determine a set of assessment criteria that allows for the quantitative measurement of these competencies, which will assist in validation of the selection process. In addition, these measures should remain consistent across time.

3. *Validation of the ACER B90* An investigation into the reliability and validity of the B90 is recommended. Given that it is used by all three services for Officer selection, this could be a tri service activity. Results do spread doubt onto the test's ability to accurately measure the domain of GMA. Further investigation into the relationship between the B90 and Raven's APM as suggested by Gracie (2004) is also recommended to determine the unique variance of each test.

4. *Assessor Training* The training provided to MTOs during the OSB should be revisited to ensure it is current. The APS could investigate developing a FOR type training package and align this with competency behaviours. In

addition, an OSB training course could be conducted similar to the RNZAF, which assessor attend prior to attending the OSB.

5. *Number of Competencies* It is recommended that the APS investigate the number of competencies employed by assessors and reduce the current competencies used. It is suggested that this occur during the NZDF Competency alignment process.

6. *Exercise Effects* An investigation into the level of exercise effect present in the OSB, to determine the MTO biases/accuracy in rating use of the competencies, is recommended. This will also provide insight into the key competencies employed by MTOs in their assessment and potentially help to limit the number of competencies used.

7. *Final Grading Score* It is recommended that a final grading score (with a degree of variability rather than a dichotomous variable) be introduced into the OSB. This could be provided by the MTO or provided by the Final Board. It is recommended that such score be derived mechanically (e.g. total of competency grades). This would allow for a more consistent predictor variable.

The following recommendations are proposed for future OSB validation research:

1. *Longitudinal Research* To help avoid some of the effects of range restriction present in this research it is recommended that a longitudinal validation approach is undertaken, whereby participant performance data is collected at time of selection, throughout training, and then during intermittent period during job performance (concentrating in particular on the first three years post-commissioning). This would allow for collection of more thorough data for analysis, and help with eliminating some of the confounding variables present in this study. It would also assist in greater sample numbers being available.

2. *Sample Size* To help avoid some of the methodological problems existent in this study it is recommended that the sample size of future validation efforts be no less than 150 participants.

3. *Ethnicity Studies* This study was not able to determine the effects of ethnicity as a factor of participant performance due to sample size constraints. It is recommended that future validation studies assess the impact of ethnic diversity on performance both in the predictor and criterion domains.

References

Aguinis, H., Henle, C. A., & Ostroff, C. (2001). Measurement in work and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of Industrial, Work and Organizational Psychology* (Vol. 1 Personnel Psychology). London: Sage Publications.

Allen, K. (2000a). *Ethnic differences in ACER B90 scores for officer applicants from June 2000 to March 2001*. Wellington: Army Psychology Service.

Allen, K. (2000b). *Gender differences in ACER B90 scores for officer applicants from 1998 to mid 1999*. Wellington: Army Psychology Service.

Allen, K. (2000c). *Officer Selection Criteria and Methods- New Zealand and other countries*. Wellington: Army Psychology Service.

Allen, K., & Mirfin, K. (1999). *The effects of age, gender and education on B90 scores for officer applicants in 1997 and 1998*. Wellington: Army Psychology Service.

Anderson, N., Born, M., & Cunningham-Snell, N. (2001). Recruitment and selection: Applicant perspectives and outcomes. In N. Anderson, D. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of Industrial, Work and Organizational Psychology* (Vol. 1). London: Sage Publications Ltd.

APS. (1995). *A revision of Officer Selection procedures*. Wellington, New Zealand: Army Psychology Service.

APS. (1996a). *Board Psychologist Folder*. Wellington: Army Psychology Service.

APS. (1996b). *OSB Folder*. Wellington, New Zealand: Army Psychology Service.

APS. (2004). *Officer Selection Board (OSB) Guidelines*. Wellington: Army Psychology Service.

ARRC. (2005). *Army Recruiting Recruiters Handbook*. Wellington: NZ Army Regional Recruiting Unit.

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A Meta-analysis of the criterion-related validity of assessment centre dimensions. *Personnel Psychology, 56*(1), 125-155.

Arthur, W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment centre dimensions: A conceptual and empirical examination of the assessment centre construct-validity paradox. *Journal of Management, 26*(4), 813-835.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27*, 519-533.

Atkins, P. W. B., & Wood, R. E. (2002). Self-versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55*, 871-904.

Atkinson, J. L. (2003). *The psychological assessment of officers and aircrew at the RNZAF selection board.* Wellington: RNZAF Psychology Service.

Barker, D. E. (1997). *A look into the distribution and the differences between applicant scores on the ACER B90.* Wellington: Army Psychology Service.

Bartram, D. (1995). The predictive validity of the EPI and 16PF for military flying training. *Journal of Occupational and Organisational Psychology, 63*(3).

Bass, B. (1990). *Bass and Stogdills handbook of leadership.* New York: Free Press.

Bennett, C. L. (1990). *Regular Officer Selection Board - Examining the validity of officer selection procedures.* Wellington: Army Psychology Service.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387-409.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*(3).

Bowden, C. (1999). *The usefulness of personality questionnaires in officer selection and training.* Massey University, Palmerston North.

Braun, P., Wiegand, D., & Aschenbrenner, H. (1991). The assessment of complex skills and of personality characteristics in military services. In R. Gal & A. D. Mangelsdorff (Eds.), *Handbook of Military Psychology.* West Sussex: England: John Wiley and Sons Ltd.

Brumback, G. (1969). A note on criterion contamination in the validation of biographical data. *Educational and Psychological Measurement, 29*(439-443).

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72*(3), 463-474.

Byham, W. C., & Thornton, G. C. (1986). Assessment Centers. In R. A. Berk (Ed.), *Performance Assessment: Methods and Applications.* Baltimore; Maryland: John Hopkins University Press Inc.

Carston, M. C. (2002). *Officer selection procedures and the professional responsibilities of New Zealand Army psychologists* (Research Report): Army Psychology Service.

Chan, D. (1996). Criterion and construct validation of an assessment center. *Journal of Occupational and Organizational Psychology, 69*(2).

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cullen, E. (1995). *An investigation into the reliability and validity of the B90.* Wellington: RNZAF Psychology Service.

Curran, P. G. (2000). *The effectiveness of the New Zealand Army officer selection process in measuring leadership potential.* Auckland: RNZAF Command and Staff College.

Daft, R. L., & Noe, R. A. (2001). *Organizational behavior*. Mason, OH: South-Western Publishing.

Damitz, M., Manzey, D., Kleinmann, M., & Severin, K. (2003). Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. *Applied Psychology: An International Review, 52*(2), 193-212.

Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment center: An efficient tool or a hammer to kill a fly? *Personnel Psychology, 55*, 827-849.

Dick, D. (1993). *A brief note on restriction of range*. Wellington: Army Psychology Service.

Dobson, P., & Williams, A. (1989). The validation of the selection of male British Army officers. *Journal of Occupational Psychology, 62*, 313-325.

Dover, S. H. (1991). A system approach to selection research and application in the military. In R. Gal & A. D. Mangelsdorff (Eds.), *Handbook of Military Psychology*. West Sussex: England: John Wiley and Sons Ltd.

Driskell, J. E., & Salas, E. (1991). Overcoming the effects of stress on military performance: Human factors, training and selection strategies. In R. Gal & A. D. Mangelsdorff (Eds.), *Handbook of Military Psychology*. West Sussex: England: John Wiley and Sons Ltd.

DSHHR. (2005). *New Zealand Defence Force Monthly Figures: December 2005*. Wellington: HQ Personnel Branch, NZDF.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and clarification systems. *Annual Review of Psychology, 30*, 477-525.

Eagar, J. P. (2004). *Peer assessments in the New Zealand Army: A selection and training perspective*. Wellington: Army Psychology Service.

Elliot, S., Lawty-Jones, M., & Jackson, C. (1996). Effect of dissimulation on self-report and objective measures of personality. *Personality and Individual Differences, 21*(3), 335-343.

Elshaw, C. C., Abram, C. M., & Weston-Lovelock, K. (1997). *The validation of the Behavioural Anchored Rating Scales and the Overall Assessments at the Regular Commissions Board.* Farnborough, Hants: UK: DERA.

Eysenck, H. J. (1970). *The structure of human personality* (3 ed.). London: Methuen.

Eysenck, S. B. G., & Eysenck, H. J. (1991). *Manual of the Eysenck personality scales.* London: Hodder and Stroughton.

Flanagan, D., Genshaft, J., & Harrison, P. (1997). *Contemporary intellectual assessment: Theories, tests and issues.* New York: The Guildford Press.

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10*(3), 391-335.

Furnham, A. (1992). *Personality at work: The role of individual differences in the workplace.* New York: Routledge.

Gatewood, R. D., & Feild, H. S. (2001). *Human Resource Selection* (5th ed.). USA: Harcourt College Publishers.

Gatewood, R. D., Thornton, G. C., & Hennessey, H. W. (1990). Reliability of exercise ratings in the leaderless group discussion. *Journal of Occupational Psychology, 63*, 331-342.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*(3), 493-511.

Gaugler, B. B., & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*(4), 611-618.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). Measurement theory for the behavioural sciences. In. New York: W. H. Freeman and Company.

Gordon, L. V. (1993). *The Gordon Personal Profile-Inventory.* San Antonio, TX: PsychCorp, A brand of Harcourt Assessment, Inc.

Gracie, E. (2004). *Comparison of the suitability of the B90 and Raven test for use in officer selection by the New Zealand Army.* Wellington: Army Psychology Service.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*, 59-61.

Guion, R. M. (1987). Changing views for personnel selection research. *Personnel Psychology, 40.*

Guion, R. M. (1998). The Gordon Personality Profile-Inventory (Revised). In J. C. Impara & B. S. Plake (Eds.), *The Thirteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Guion, R. M. (2002). Validity and reliability. In S. G. Rogelberg (Ed.), *Handbook of Research Methods in Industrial and Organizational Psychology.* Oxford, UK: Blackwell Publishers Ltd.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multiple regression analysis. In J. F. Hair, R. E. Anderson, R. L. Tatham & W. C. Black (Eds.), *Multivariate Data Analysis* (5 ed., pp. 141-214). Englewood Cliffs, NJ: Prentice Hall.

Harmer, P. (2006). New Zealand Army Officer Statistics (pp. Email). Wellington: HQ Personnel Branch, NZDF.

Harper, G., & Hayward, J. (2003). *Born to lead.* Auckland, New Zealand: Exisle Publishing Ltd.

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

Harrison, M. J. (2005). *An investigation into a working assessment centre: Validating the Royal New Zealand Navy's Final Officer Selection Board.* Auckland University, Auckland.

Horn, H. (2005). OSB Information (pp. Email). Auckland.

Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of Industrial, Work and Organizational Psychology* (Vol. 1 Personnel Psychology). London: Sage Publications.

Howard, A. (1997). A reassessment of the assessment center: Challenges in the 21st century. *Journal of Social Behavior and Personality, 12*(5), 13-52.

Hughes, P. (2000). *A qualitative study of the future conduct of leadership in the New Zealand Army.* Unpublished MBS Research Report, Massey University, Palmerston North.

Hunter, D. R., & Burke, E. F. (1995). *Handbook of pilot selection.* Aldershot, England: Ashgate Publishing Ltd.

Irvine, S. H., & Berry, J. W. (1983). *Human assessment and cultural factors.* London: Plenum Press.

Jacobs, R. R. (1986). Numerical Rating Scales. In R. A. Berk (Ed.), *Performance Assessment: Methods and Applications.* Baltimore; Maryland: John Hopkins University Press Inc.

Jennings, M. (1998). *Structured Team Officer interviewing for OASB's.* Wellington: RNZAF Psychology Service.

Jensen, A. R. (1980). *Bias in Mental Testing.* New York: Free Press.

Jones, A. (1991). The contribution of psychologists to military officer selection. In R. Gal & A. D. Mangelsdorff (Eds.), *Handbook of Military Psychology.* West Sussex; England: John Wiley and Sons Ltd.

Jones, A., Herriot, P., Long, B., & Drakeley, R. (1991). Attempting to improve the validity of a well-established assessment center. *Journal of Occupational Psychology, 64*, 1-21.

Kearney, S. (2005). OSB Information (pp. Email). Waiouru.

Keatly, Y. (1998). *Employee selection in large NZ organisations: Current practices and recent changes.* Waikato University, New Zealand.

Kelly, B. J. (1994). *An investigation of the incremental validity provided by the Royal New Zealand Army selection procedure components.* University of Auckland, Auckland.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in multidimensional forced choice performance evaluation scale. *Journal of Applied Psychology, 65*, 507-516.

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology, 78*(6), 988-993.

Klem, L. (2003). Path Analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics* (pp. 65-98). Washington D.C.: American Psychological Association.

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*, 243-260.

Klimoski, R., & Strickland, W. (1977). Assessment centers-Valid or merely prescient. *Personnel Psychology, 30*, 353-361.

Kline, P. (1993). *The handbook of Psychological testing.* London: Routledge.

Lance, C. E., Newbolt, W., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*(4), 323-353.

Lane, M. (1998). *Gender differences in the B90 test of general reasoning ability.* Wellington: RNZAF Psychology Service.

Licht, M. H. (2003). Multiple regression and correlation. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics* (pp. 19-64). Washington D.C.: American Psychological Association.

Lievens, F. (1998). Factors which improve the construct validity of the Assessment Centre: A review. *International Journal of Selection and Assessment, 6*(3).

Lievens, F., & Conway, J. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*(6).

Lievens, F., Harris, M. M., VanKeer, E., & Bisqueret, C. (2003). Predicting cross-cultural training performance: The validity of personality, cognitive ability and dimensions measured by an assessment center and a behaviour description interview. *Journal of Applied Psychology, 88*(3), 476-489.

Lowry, P. E. (1993). The assessment centre: An examination of the effects of assessor characteristics on assessor scores. *Public Personnel Management, 22*(3), 487-501.

Lowry, P. E. (1997). The assessment center process: New directions. *Journal of Social Behavior and Personality, 12*(5), 53-63.

Novick, M. R. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79*(6).

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

Pech, R. J. (1998). *Mapping cognitive architectures: An informational processing approach.* Unpublished PhD, Massey University, Palmerston North.

Preece, C. (2005). OCS General Outline (pp. Email). Papakura.

Pynes, J. E., & Bernardin, H. J. (1989). Predictive validity of an entry level Police Officer assessment. *Journal of Applied Psychology, 74*(5).

Ratcliffe, V. (2005). *Partial validation of the Royal New Zealand Navy Final Officer Selection Board.* Auckland: Royal New Zealand Navy Psychology Service.

Raven, J. C. (2000). The Raven's Progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Reber, A. S., & Reber, E. (2001). *The Penguin dictionary of psychology (3rd Ed)*. London, England: Penguin Books Ltd.

Ree, M. J., Caretta, T. R., & Steindl, J. R. (2001). Cognitive ability. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of Industrial, Work and Organizational Psychology* (Vol. 1 Personnel Psychology). London: Sage Publications.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44*.

Reid, N., & Croft, C. (1991). *ACER Advanced Test New Zealand Edition: Administrators manual.* Wellington: New Zealand Council for Educational Research.

Riggio, R. E. (1996). *Industrial/Organizational psychology*. New York: HarperCollins College Publishers.

RMAS. (1985). *Job specification: RMAS Young Officer*: Royal Military Academy Sandhurst.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology, 55*, 171-183.

Romaine, K. A. (2004). Developing Lieutenants in a transforming Army. *Military Review, Jul-Aug*, 72-80.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 73*(3), 322-327.

Russell, C. J., & Domm, D. R. (1995). Two field tests of an explanation of assessment centre validity. *Journal of Occupational and Organisational Psychology, 68*, 25-47.

Russell, D. (2005). Research Information, *Email dated 9 Oct*. Auckland.

Rust, J. (1991). Review of the ACER Advanced Test B90: New Zealand edition. In J. C. Conoley & J. C. Impara (Eds.), *The Twelfth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*, 13-25.

Sackett, P. R. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112-118.

Sackett, P. R., & Larson, J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2 ed., Vol. 1, pp. 419-489). Palo Alto, CA: Consulting Psychologists Press, Inc.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., deFruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the european community. *Journal of Applied Psychology, 88*(6), 1068-1081.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs methods and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of Industrial, Work and Organizational Psychology* (Vol. 1). London: Sage Publications Ltd.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for Frame of Reference Training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*(4), 735-746.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*, 162-173.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*(3), 432-439.

Schmitt, N., & Robertson, I. (1990). Personnel selection. *Annual Review Psychology, 41*, 289-319.

Shore, T. H., Shore, L. M., & Thornton, G. C. (1992). Construct validity of self and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77*(1), 42-54.

Siegel, A. I. (1986). Performance Tests. In R. A. Berk (Ed.), *Performance Assessment: Methods and Applications.* Baltimore; Maryland: John Hopkins University Press Inc.

Smith, M. (1994). A theory of the validity of predictors in selection. *Journal of Occupational and Organizational Psychology, 67*, 13-31.

Snider, D. M. (2003). Officership: The professional practice. *Military Review, Jan-Feb.*

Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50*, 71-90.

Steege, F. W., & Fritscher, W. (1991). Psychological assessment and military personnel management. In R. Gal & A. D. Mangelsdorff (Eds.), *Handbook of Military Psychology*. West Sussex: England: John Wiley and Sons Ltd.

Sumer, H. C., Sumer, N., Demirutku, K., & Cifci, O. S. (2001). Using a personality-oriented job analysis to identify attributes to be assessed in officer selection. *Military Psychology, 13*(3), 129-147.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics.* Massachusetts: Allyn & Bacon.

Taskforce on Assessment Centre Guidelines (2000). Guidelines for ethical considerations for assessment center operations. *Public Personnel Management, 29*3, 315-331.

Taylor, K. (2005). *NZ Army Officer Careers.* Wellington.

Turner, J. B. (1988). *Officer Selection Board assessment procedures.* Wellington: Defence Psychology Unit.

Tziner, A., Meir, E. I., Dahan, M., & Birati, A. (1994). An investigation of the predictive validity and economic utility of the assessment center for the high-management level. *Canadian Journal of Behavioural Science, 26*(2), 228-245.

Tziner, A., Ronen, S., & Hacohen, D. (1993). A four year validation study of an assessment center in a financial coorporation. *Journal of Organizational Behaviour, 14*, 225-237.

Walsh, J. P., Weinberg, R. M., & Fairfield, M. L. (1987). The effects of gender on assessment centre evaluations. *Journal of Occupational Psychology, 60*, 305-309.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validty of the employment interview. *Journal of Occupational Psychology, 61*, 275-290.

Zaal, J. N. (1998). Assessment centre methods. In P. J. D. Drenth, H. Thierry & C. J. deWolff (Eds.), Handbook of work and organizational psychology (2nd Ed) (Vol. 3: Personnel Psychology). United Kingdom: Psychology Press.

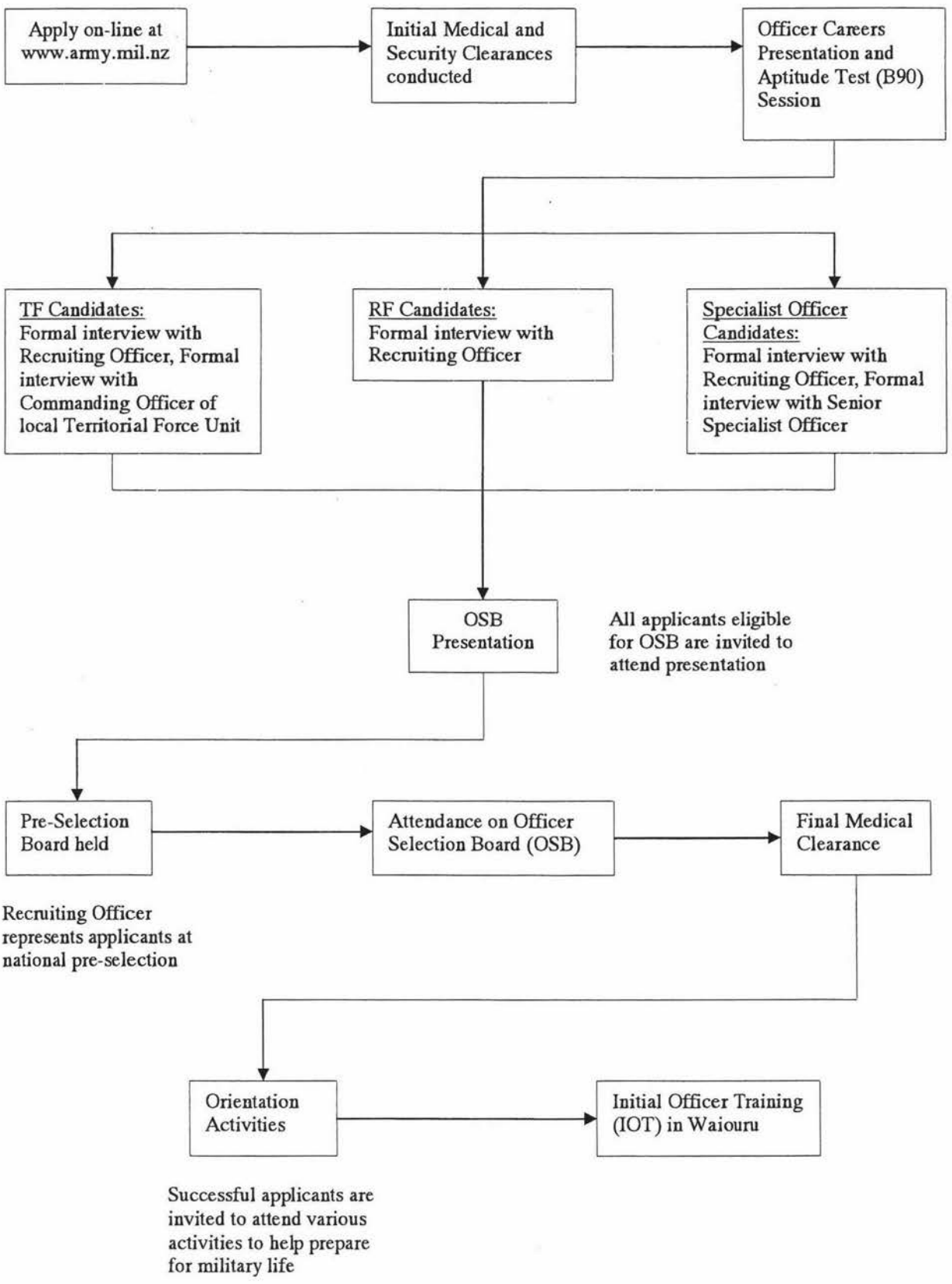
Appendix 1

This appendix has been removed. Please contact the Senior Psychologist Army, NZDF for access to this appendix.

Thank you.

Appendix 2

New Zealand Army Officer Selection Recruitment Process



Appendix 3

This appendix has been removed. Please contact the Senior Psychologist Army, NZDF for access to this appendix.

Thank you.

Appendix 4

MD68 Assessment Criteria

The following assessment criteria from is completed by the participants immediate superior each year. The grading (1-9) is based on the officers current rank. This assessment is only part of the overall appraisal document.

	N/O	* 1	* 2	3	4	5	6	7	8	* 9	Comments
Section A - Personal Qualities											
Appearance and Bearing											
Co-operation											
Loyalty to Others											
Determination and Industry											
Integrity and Fairness											
Sense of duty											
Power of Oral Expression											
Social Conduct											
Section B - Proficiency in Administration											
Understanding Instructions											
Performance of Own Work											
Knowledge of Org and Procedures											
Power of Written Expression											
Section C - Proficiency in Management of Personnel											
Delegating Authority											
Maintaining Discipline											
Developing Teamwork											
Ability to Inspire Confidence											
Section D - Proficiency in Planning and Directing											
Taking Responsibility											
Decisiveness											
Judgement											
Initiative											
Planning Ahead											
Section E - Effectiveness in Primary Appointment or Duties											
Proficiency											
Knowledge											
Reliability											
Determination											
Section F - General Ability and Overall Evaluation											
Proficiency in Primary Appointment											(Appt)
Proficiency in Secondary Appointment											(Appt)
Proficiency in Any Other Appointment											(Appt)
Attitude To Further Service with Officer											
Assessment of Officer's Future Potential											
Your Overall Assessment											

Appendix 5

This appendix has been removed. Please contact the Senior Psychologist Army, NZDF for access to this appendix.

Thank you.

Appendix 6

Approval to Conduct Research within NZDF

NEW ZEALAND ARMY

Ngati Tumatauenga

Army Psychology Service
Army General Staff, Headquarters, New Zealand Defence Force, Private Bag 39997, Wellington.
Telephone: +64 4 496 0085, Facsimile: +64 4 496 0488,
Internet: www.army.mil.nz

SPA 5/OSB

16 Apr 05

The Human Ethics Committee
Psychology Department
Massey University
Palmerston North

APPROVAL TO CONDUCT RESEARCH IN THE NZDF: A CRITERION VALIDATION OF THE NEW ZEALAND ARMY OFFICER SELECTION PROCESS

1. Lt Kathryn Benjamin is currently employed by the Army Psychology Service as a research officer. Lt Benjamin has been given approval to conduct a validation study of the New Zealand Army Officer Selection process. The conduct of this research will be supervised by the Army Psychology Service.

2. Lt Benjamin acknowledges that the research will be conducted in accordance with the Privacy Act 1993, and ethical guidelines outlines in Schedule Two of Annex B to DFO 21/2002.

H.M. HORN
Major
Senior Psychologist Army

Appendix 7

Massey University Ethics Approval

Massey University

OFFICE OF THE ASSISTANT
TO THE VICE-CHANCELLOR
(ETHICS & EQUITY)
Private Bag 11 222
Palmerston North
New Zealand
T 64 6 350 5573
F 64 6 350 5622
humanethics@massey.ac.nz
www.massey.ac.nz

25 July 2005

Kathryn Benjamin
4 Redmount Place
Papakura
AUCKLAND 1703

Dear Kathryn

Re: HEC: PN Application – 05/44
A criterion validation of the New Zealand Army Officer selection process

Thank you for your letter received 20 July 2005.

On behalf of the Massey University Human Ethics Committee: Palmerston North I am pleased to advise you that the ethics of your application are approved. Approval is for three years. If this project has not been completed within three years from the date of this letter, reapproval must be requested.

If the nature, content, location, procedures or personnel of your approved application change, please advise the Secretary of the Committee.

A reminder to include the following statement on all public documents: *"This project has been reviewed and approved by the Massey University Human Ethics Committee, Palmerston North Application 05/44. If you have any concerns about the ethics of this research, please contact Dr John G O'Neill, Chair, Massey University Campus Human Ethics Committee: PN telephone 06 350 5799 x 8635, email humanethicspn@massey.ac.nz".*

Yours sincerely

J. O'Neill

Dr John G O'Neill, Chair
Massey University Campus Human Ethics Committee: Palmerston North

cc Dr Fiona Alpass
School of Psychology
PN320

Professor Ian Evans, HoS
School of Psychology
PN320

Appendix 8

Participant Information and Consent Form

A Criterion Validation of the New Zealand Army Officer Selection Process

Candidate Information Sheet

INTRODUCTION

My name is Kate Benjamin and I am currently employed as a Research Officer with the Army Psychology Service. Presently I am working towards completion of my Master's degree in Industrial/Organisational Psychology (MSc) at Massey University. I would like to invite you to participate in this research project and read the information below.

RESEARCH PROJECT

The project is intended to validate the current selection process employed by the New Zealand Army for Officer Selection. The intent is to compare data obtained from the OSB, against candidate performance at OCS, and subsequent performance post graduation whilst working as a junior officer in the Army. This validation process will help to determine whether the criteria used for selection on the OSB are providing valid information for selection decisions. Data gained from this study will also provide benefit to future validation studies within the Army and New Zealand Defence Force.

It has been a significant period of time since the OSB was last validated. During this time the selection process has changed significantly with different measures being utilised and different dimensions being assessed. On the whole the objective remains the same with the focus on identifying leadership potential. Due to the cost to both the military and to personnel selected it is crucial to the army that only those candidates assessed as having the potential to train and work as an officer are selected.

The project discusses the tests and measures used in the OSB and looks at how well they predict performance or officer potential. As a professional institution we are obligated to ensure that our selection procedures are valid and ethical, that involves ensuring that the OSB is actually a good selection system. More importantly however, as an organisation that depends on teamwork and leadership it is important to ensure that our selection systems are selecting the best candidates for OCS and the junior officer job. Hence this project is designed to assess the validity of the OSB process.

PARTICIPATION

In order to produce results that are more reliable and have greater influence it is important to obtain a large participant pool. All Regular Force graduates of OCS (NZ) from the period 1996-2003, have been invited to participate. Due to the structure of the project a large number of participants are required to make the analysis valid. This is why all graduates from this period have been invited to participate. Those wishing to participate in the study need to fill out the attached consent form and return it in the envelope provided. Participation is completely voluntary.

If you decide to participate, data will be collected on the following criteria:

a. OSB Performance data This will be collected by Kate Benjamin from files held by the Army Psychology Service and will be coded prior to entry onto the project database. The coding process simply replaces your name with a number so that your name is not identified with your data. It is done to provide greater confidentiality to your identity.

b. OCS Performance data This will be collected by Kate Benjamin from files held at OCS and will be coded prior to entry onto the project database.

c. MD68 Grading This will be collected by the Army Psychology Service, and coded prior to entry onto the project database. Kate Benjamin will only have access to this information in a coded numeric format.

d. Immediate Supervisor Rating Questionnaire This data will be collected from a questionnaire that will be sent to your immediate supervisor, coded and then entered onto the database.

The consent form asks you to provide your consent to each of these criteria. You have the right to withhold your consent on data collection of any of criteria.

RESEARCH PROCEDURES

Data is collected for the sole purpose of validation of the OSB and New Zealand Army selection procedures. Once data has been collected, it will then be coded to remove identifying information, and held by the Army Psychology Service. You will have access to the summary findings of the research and a copy of the summary findings will be mailed to you at your unit address.

PARTICIPANT RIGHTS

You are under no obligation to accept this invitation. If you decide to participate, you have the right to:

a. Decline to answer any particular question,
b. Withdraw from the study prior to 30 July 2005,
c. Ask any questions about the study at any time during participation,
d. Provide information on the understanding that you name will not be used unless you give permission to the researcher, and
e. Be given access to a summary of the project findings when it is concluded.

CONTACT DETAILS

Should you wish to discuss your participation or any details of this research Kate Benjamin can be contacted by phone or email as shown below. Additionally, if you wish to discuss elements of the research with my Massey University supervisor, or with the Senior Psychologist Army, they can be contacted as shown below.

Full Name: Kathryn Benjamin
Employer: New Zealand Defence Force (Army)
Telephone: 021 1136 322
Email: Kathryn.Benjamin@nzdf.mil.nz

Supervisor Details
Full Name: Fiona Alpass
School/Department: Psychology Department/Massey University
Region: Palmerston North
Telephone: (06) 356 9099 Ext. 2071
Email: F.M.Alpas@massey.ac.nz

Senior Psychologist Army
Full Name: Maj Helen Horn
Telephone: (04) 496 0085
Email: Helen.Horn@nzdf.mil.nz

COMMITTEE APPROVAL STATEMENT

This project has been reviewed and approved by the Massey University Human Ethics Committee PN Application 05/44. If you have any concerns about the conduct of this research, please contact Dr John G. O'Neill as follows:

Full Name: D John G. O'Neill
Position: Chair
Massey University Campus Human Ethics Committee
Palmerston North
Telephone: (06) 350 5799 xtn 8635
Email: humanethicspn@massey.ac.nz

This project has also been reviewed and approved by the CA, Maj Gen J. Mateparae on the 02 Apr 2005.

A Criterion Validation of the New Zealand Army Officer Selection Process

Participant Consent Form

I have read the information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time. I agree to participate in this study under the conditions set out in the Information Sheet.

I provide consent for the following: (Please tick the appropriate box)

		YES	NO
1.	For data to be collected from my OSB performance records by Kate Benjamin.	☐	☐
2.	For data to be collected from my OCS performance records by Kate Benjamin.	☐	☐
3.	For data to be collected from my MD68 records by the Army Psychology Service.	☐	☐
4.	For data to be collected from my immediate supervisor on my performance in the form of a questionnaire by Kate Benjamin.	☐	☐

Please provide immediate supervisor details:

Full Name: _____

Rank: _____

Unit: _____

5.	For collected data to be held on file by NZDF for validation and selection purposes, under the understanding that access to this information is restricted to approval by the Senior Psychologist (Army).	☐	☐

6. Please provide the following demographic details:

a. Age: _____ yrs

b. Gender: Male ☐ Female ☐

c. Ethnicity: NZ Maori ☐ NZ European/Pakeha ☐

Pacific Islander ☐ Asian ☐

Other (Please Specify) _____

d. Time since commissioning: _____ yrs and _____ mths

e. Number of deployments: _____

	YES	NO
7. I would like a summary of the results on completion of the research.	☐	☐

Signature: _____

Full Name: _____

Date: _____

Appendix 9

Letter to Participants from Deputy Chief of Army

NEW ZEALAND ARMY
Army General Staff
MINUTE

6725/1

25 Jul 05

Selected Army Officers

VALIDATION OF THE OFFICER SELECTION BOARD (OSB)

1. By receiving this minute you are being asked to participate in a research project to validate the Officer Selection Board (OSB) conducted by the Army Psychology Service.

2. This research is intended to help ensure that the selection decisions made during the OSB are fair, robust and meet the needs of the Army. This research will also help to identify areas in which we can improve our selection methods and better select candidates suited to the Army's professional and dynamic culture. In addition it addresses the NZDF Human Resources Principles and Values and has at its core Army Strategic Goal 1 which is to "Recruit, Develop and Retain the right people" in order to be "a world class Army that has mana".

3. As the President of the OSB I encourage you to take the time to complete the consent form and take part in this valuable research.

J.B. VRYENHOEK
BRIG
DCA

Enclosure:
1. OSB Validation Research Consent Form and Information Sheets

Appendix 10

Supervisor Information and Consent Form

A Criterion Validation of the New Zealand Army Officer Selection Process

Supervisor Information Sheet

INTRODUCTION

My name is Kate Benjamin and I am currently employed as a Research Officer with the Army Psychology Service. Presently I am working towards completion of my Master's degree in Industrial/Organisational Psychology (MSc) at Massey University. I would like to invite you to participate in this research project and read the information below.

RESEARCH PROJECT

The project is intended to validate the current selection process employed by the New Zealand Army for Officer Selection. The intent is to compare data obtained from the OSB, against candidate performance at OCS, and subsequent performance post graduation whilst working as a junior officer in the Army. This validation process will help to determine whether the criteria used for selection on the OSB are providing valid information for selection decisions. Data gained from this study will also provide benefit to future validation studies within the Army and New Zealand Defence Force.

It has been a significant period of time since the OSB was last validated. During this time the selection process has changed significantly with different measures being utilised and different dimensions being assessed. On the whole the objective remains the same with the focus on identifying leadership potential. Due to the cost to both the military and to personnel selected it is crucial to the army that only those candidates assessed as having the potential to train and work as an officer are selected.

The project discusses the tests and measures used in the OSB and looks at how well they predict performance or officer potential. As a professional institution we are obligated to ensure that our selection procedures are valid and ethical, that involves ensuring that the OSB is actually a good selection system. More importantly however, as an organisation that depends on teamwork and leadership it is important to ensure that our selection systems are selecting the best candidates for OCS and the junior officer job. Hence this project is designed to assess the validity of the OSB process.

PARTICIPATION

There are two categories of participants in this project:

Group 1: Candidates

Candidates will include personnel who attended selection boards between 1995-2003, who were commissioned from the New Zealand Officer Cadet School (OCS).

Group 2: Immediate Supervisors

Immediate supervisors of current serving candidates will be requested to provide performance data on candidates via a questionnaire. The number of supervisors approached will reflect the number of candidates used in the research.

Your participation is requested to enable me to collect the immediate supervisor ratings on a candidate. The candidate has already provided consent for this information to be collected about their performance. However I am still required to obtain your consent to participate in this research.

You are invited to complete the attached consent form and Immediate Supervisor Rating Questionnaire on the candidate (Candidate name is indicated on the questionnaire) and return both forms in the envelope provided before the **30 July 2005.**

RESEARCH PROCEDURES

Data is collected for the sole purpose of validation of the OSB and New Zealand Army selection procedures. Once data has been collected, it will then be coded to remove identifying information, and held by the Army Psychology Service. You will have access to the summary findings of the research and a copy of the summary findings will be mailed to you at your unit address.

As an immediate supervisor participant, participation in this study will not be time consuming. You are only required to complete the attached consent form and questionnaire and return it in the envelope provided.

PARTICIPANT RIGHTS

You are under no obligation to accept this invitation. If you decide to participate, you have the right to:

a. Decline to answer any particular question,
b. Withdraw from the study prior to 30 July 2005,
c. Ask any questions about the study at any time during participation,
d. Provide information on the understanding that you name will not be used unless you give permission to the researcher, and
e. Be given access to a summary of the project findings when it is concluded.

If you decide not to participate in this research, **you must still return the documents in the envelope provided (in order to protect the integrity of the OSB).**

CONTACT DETAILS

Should you wish to discuss your participation or any details of this research Kate Benjamin can be contacted by phone or email as shown below. Additionally, if you wish to discuss elements of the research with my Massey University supervisor, or with the Senior Psychologist Army, they can be contacted as shown below.

Full Name:	Kathryn Benjamin
Employer:	New Zealand Defence Force (Army)
Telephone:	021 1136 322
Email:	Kathryn.Benjamin@nzdf.mil.nz

Supervisor Details

Full Name:	Fiona Alpass
School/Department:	Psychology Department/Massey University
Region:	Palmerston North
Telephone:	(06) 356 9099 Ext. 2071
Email:	F.M.Alpas@massey.ac.nz

Senior Psychologist Army

Full Name:	Maj Helen Horn
Telephone:	(04) 496 0085
Email:	Helen.Horn@nzdf.mil.nz

COMMITTEE APPROVAL STATEMENT

This project has been reviewed and approved by the Massey University Human Ethics Committee PN Application 05/44. If you have any concerns about the conduct of this research, please contact Dr John G. O'Neill as follows:

Full Name: D John G. O'Neill
Position: Chair
Massey University Campus Human Ethics Committee
Palmerston North
Telephone: (06) 350 5799 xtn 8635
Email: humanethicspn@massey.ac.nz

This project has also been reviewed and approved by the CA, Maj Gen J. Mateparae on the 02 Apr 2005.

A Criterion Validation of the New Zealand Army Officer Selection Process

Participant Consent Form

This form is to be returned in the envelope provided before 30 July 2005.

I have read the information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time. I agree to participate in this study under the conditions set out in the Information Sheet.

I provide consent for the following: (Please tick the appropriate box)

	YES	NO
1. For questionnaire data provided by me to be used in the project.	☐	☐
2. For collected data to be held on file by NZDF for validation and selection purposes, under the understanding that access to this information is restricted to approval by the Senior Psychologist (Army).	☐	☐
3. I would like a summary of the results on completion of the research.	☐	☐

Signature: _____

Full Name: _____

Date: _____

Statistical Appendix 1

This appendix has been removed. Please contact the Senior Psychologist Army, NZDF for access to this appendix.

Thank you.

Statistical Appendix 2

This appendix has been removed. Please contact the Senior Psychologist Army, NZDF for access to this appendix.

Thank you.