THE USEFULNESS

OF

PERSONALITY QUESTIONNAIRES

IN

OFFICER SELECTION AND TRAINING

A paper submitted in fulfilment of the Master of Science Degree

Charlotte Bowden
Department of Psychology
Massey University
1999

Acknowledgements

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

My supervisor, Associate Professor Douglas Paton for his guidance and support; the NZ Army for the use of their facilities and support; Major Kate Mirfin, Major Clare Bennett and the NZ Army psychologists for their assistance in developing the Workplace Behaviour Questionnaire and proofreading the thesis; Duncan Hedderley for his assistance with statistical analysis; and Michael Duggan for his generous assistance with helping me understand the data and putting up with all my questions.

CONTENTS

CONTENTS (cont.)

Results (cont.)

Abstract

The aim of the current research was to assess whether the Revised Eysenck Personality Questionnaire (EPQ-R) and the Gordon Personal Profile-Inventory (GPP-I) could predict future behaviour in a sample of NZ Army officers and officer cadets. Personality questionnaire data completed at the time of selection was correlated with a workplace behaviour questionnaire (WBQ) developed specifically for the purposes of the research. It was hypothesised that (1) EPQ-R and GPP-I scales should correlate significantly with their corresponding scales on the WBQ, (2) the Neuroticism/Lie and Psychoticism/Lie correlation should indicate the presence of faking, (3) officers serving longer than three years should show more similar personality profiles than officers serving less than three years, (4) immediate superiors of the same gender and ethnicity should rate participants more favourably than those of a different gender and ethnicity, and (5) scores on the WBQ measuring High Psychoticism, High Neuroticism, Low Emotional Stability, Low Ascendancy, and Low Cautiousness should not be endorsed highly if selection has been effective. Only the fifth hypothesis was supported and the results are discussed in light of methodological shortcomings and earlier research.

Introduction

Selectors are faced with the task of assessing the current abilities of the job applicant and predicting whether the person will continue to perform those abilities while employed. Selectors use many tools to aid this process, such as application forms, interviews, and questionnaires. In particular, cognitive ability test scores have been associated with success in many different occupational areas (Robertson, 1994). Personality questionnaires have often been used because they help the selector judge whether a person will "fit" in an organisation. In addition, research has found that some personality traits are good predictors of job performance criteria (see Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). Personality questionnaires have not, however, enjoyed widespread usage in selection contexts. The reasons for this form the focus of the literature review, together with how the validity and, hence, usefulness of personality questionnaires can be improved for selection purposes.

Overview of Issues

This study aimed to assess the usefulness of two of the personality questionnaires used by the NZ Army, the Eysenck Personality Questionnaire (Revised) (EPQ-R) and the Gordon Personal Profile-Inventory (GPP-I), for officer selection. A workplace behaviour questionnaire (WBQ) was developed specifically for the purposes of the research to measure behaviour associated with each of the personality questionnaire scales. As the EPQ-R and GPP-I are used for many purposes in the NZ Army, for example, officer selection and placement, it is important to ensure they measure what they claim to measure, and that they are fair, efficient, and effective. This is because employment and career paths often constitute a major part of someone's life, therefore, these decision-making tools need to be as accurate as possible. Furthermore, inaccurate decisions are costly to an organisation in terms of resources and spending extra time and money in repeating the selection process when attrition occurs. The literature review will focus on important issues relevant to the use of personality questionnaires in selection settings, in particular, the military selection setting.

First, why use personality questionnaires in the first place? For one thing, they have been shown to provide incremental validity over the more standard cognitive ability tests used in personnel selection. Consequently, large organisations, such as the NZ Army, may use an assessment centre selection process that combines cognitive ability tests with personality questionnaires. Other sources of data, such as leaderless tasks and situational tests, are also combined to increase the validity of the assessment centre data and make a more accurate decision. Second, validity research has focused on solving the lack of structure of personality traits for selection purposes. Having a structure provides a firm basis for developing questionnaires to measure the personality construct. Construct definition, then, is an important step in ensuring sound measures are developed. The Five Factor Model (FFM) was one answer to defining the personality construct and has been useful in predicting future performance. Third, for a questionnaire to be useful, it should be based on an empirically-validated theory. Furthermore, for selection purposes, this theory needs to be linked to theories concerning job performance so that criterion development can occur. Job analyses using personality-relevant criteria provide one way in which personality can be linked to job performance.

Fourth, while scale development issues are important, the usefulness of any personality questionnaire will depend on the context for which it was designed. Personality questionnaires specifically developed for selection settings have shown higher validity than standard personality questionnaires used in a selection setting. How the data is to be used, then, remains an important consideration in validating personality questionnaires for selection purposes.

Finally, a major issue concerns the ease with which personality questionnaires can be faked in selection settings. Some researchers have developed methods to combat faking, such as lie scales, response formats, and peer rating forms, all with varying degrees of success.

In this study, data from personality questionnaires completed at the Officer Selection Board (OSB) stage were correlated with data obtained from a workplace behaviour questionnaire (WBQ) developed for the purposes of the research. The WBQ contained items that described behaviours that each of the EPQ-R and GPP-I scales were thought to exhibit. These items were written to assess whether the EPQ-R and GPP-I measured what they

claimed to measure. Job analysis criteria were not used as the NZ Army's criteria were based on Australian and British Army officer job analyses. A review of the officer selection process, of which the current research forms a part, is underway and includes the development of a job analysis of the NZ Army officer role. It was thought that research into personality and job performance should wait until this job analysis had been completed. Hence, the aim of the current research was to see whether each of the scales from the personality questionnaires correlated with their corresponding behavioural scale in the WBQ. If they did, then the two personality questionnaires could be seen as useful selection tools as they would be successful in helping identify future behaviour associated with that particular personality scale. Furthermore, if selection based on these personality questionnaires had been successful, then there should be a low endorsement rate of items associated with the undesirable personality profiles.

Literature Review

Early Research

Early meta-analyses did not show promising results for the validity of personality questionnaires in selection. For example, Guion and Gottier (1965) performed a meta-analysis of validity studies conducted during a twelve-year period that focused on the relationship of personality questionnaire scores and successful behaviour in civilian employment. They found that validities reported were weak, and some were negative. Although Guion and Gottier (1965) conceded a need for personality measures to predict workplace behaviour, they did not recommend their use because the studies surveyed contained poor research designs, there was no theory relevant to workplace behaviour, and only weak validities were found (Guion & Gottier, 1965). Subsequent research has been directed at identifying how the validity of personality questionnaire data can be improved for selection purposes. In particular, later research has shown that personality questionnaires were useful for selection as some traits, e.g., Conscientiousness, were predictive of successful job performance (Barrick & Mount, 1991).

One main area of research concerned the incremental validity of personality questionnaires. Day and Silverman (1989) defined incremental validity as the ability of personality questionnaires to predict job criteria over and above that of cognitive ability tests.

Overall performance was often comprised of both task and people requirements and, while cognitive ability was related to task requirements, personality was thought to better account for the people requirements (Day & Silverman, 1989). Research has supported this theory. For example, McHenry, Hough, Toquam, Hanson, and Ashworth (1990) performed a study as part of a large-scale project to develop an officer selection procedure for the US Army. They sought to assess the ability of the Assessment of Background and Life Experiences (ABLE) to provide incremental validity over cognitive ability tests. McHenry et al. (1990) hypothesised that the ABLE scores would add significant predictive validity to the job performance criteria of Effort and Leadership, and would best predict Personal Discipline, and Physical Fitness and Military Bearing. Correlates were significant at 0.33, 0.32, and 0.37 respectively for the three criteria, which provided support for their hypothesis (McHenry et al., 1990). Furthermore, the ABLE composites were the poorest predictors of the task-related criteria, which were better predicted by the cognitive ability composites. However, when used as a composite with other predictors, the ABLE scores predicted the task-related criteria better than they predicted the people-related criteria (McHenry et al., 1990). Finally, regression analyses found that the ABLE accounted for the greatest increase in incremental validity. This research, then, provided support for the use of personality questionnaires in personnel selection as they were shown to add meaningful information over and above that provided by cognitive ability tests.

Similarly, Black (1997) sought to determine whether the Revised NEO Personality Inventory (NEO-PI-R) displayed incremental validity over cognitive ability tests. The NEO-PI-R was administered to police recruits during the first month of training. Performance scores following completion of basic training and the pre-selection cognitive ability test scores were obtained for the recruits. Black (1997) found that the cognitive test score correlated the highest with the job performance score (0.33). Regression analyses showed that the NEO-PI-R did provide incremental validity over cognitive ability test scores as a predictor of job performance. The correlation for cognitive ability tests was increased to 0.43 with the addition of the personality questionnaire global factors and further raised to 0.47 with the addition of the subfactors (Black, 1997). It is worth noting that Black (1997) referred to the "job performance" score when, in fact, the performance measure was that of training performance, an arguably different construct to job performance.

Summary and implications

Early uses of personality questionnaires were not promising for selection settings. However, later research showed that personality questionnaires provided a picture of a person's general character and could offer incremental validity over cognitive ability tests. In particular, the increased validity provided by these questionnaires gave more accurate information about a candidate for an officer role in the military and in the police force. As an officer's role can be stressful, and requires good leadership, energy, and assertiveness, using personality questionnaires with demonstrated validity of the required traits will aid the selection process, as these traits may not be tapped by cognitive ability tests. Personality questionnaires are only a part of the selection process; hiring decisions are not made on the basis of these results alone. Rather, many parts make up the process, and the overall impression given from the other parts of the selection process determine the likelihood of being hired. To further improve the validity of the selection process, some organisations, such as the NZ Army, combine personality questionnaires and cognitive ability tests with other exercises to form an assessment centre.

Assessment Centres

An assessment centre is an example of a comprehensive selection process in which applicants or potential management candidates are involved in a number of exercises designed to assess leadership potential and which resemble the environment in which they would work should they be successful. A military assessment centre, therefore, would consist of tasks that an officer could expect to carry out during a normal day, but also tasks that may be required during an operational posting such as strategising and problem-solving tasks. Although costly, assessment centres are seen as highly valid and efficient selection procedures. They let psychologists and other managers gain valuable information about individuals by using a variety of assessment techniques in a residential programme administered over a period of days (Goodstein & Lanyon, 1999). Measures completed include psychological tests such as personality questionnaires and intelligence tests, interviews, and biodata forms (Goodstein & Lanyon, 1999). Empirical evidence has shown that using personality questionnaires as part of this process can predict managerial success (Goodstein & Lanyon, 1999). The personality questionnaires are often completed, scored and interpreted at the beginning of the programme before any other information is obtained

about the applicant. This information is generally not discussed with the other raters until the end of the programme (Goodstein & Lanyon, 1999).

Assessment centres are also comprised of leaderless discussion groups and situational tests (Goodstein & Lanyon, 1999). During a leaderless task, raters observe the dynamics as a group discusses an assigned topic or performs a practical task. The applicants are often rated on dimensions such as energy, initiative, planning, communication skills, interpersonal skills, decision-making, and persuasiveness (Goodstein & Lanyon, 1999). Goodstein and Lanyon (1999) did not state how these dimensions were developed, for example, whether they were developed from job analyses and the like. Peer ratings may be taken from participants during the process as well (Goodstein & Lanyon, 1999). It is not difficult to see how these dimensions relate to success in the armed forces. Leaders and, therefore, officers, require energy as well as the ability to take initiative. Officers may be placed in situations where they must think and decide on a course of action quickly. Tasks need to be planned much as managers would plan tasks for their employees. Officers may need to come up with new ideas and, importantly, they need to communicate their decisions to their soldiers to ensure the soldiers understand their tasks. Furthermore, officers need to be able to relate well to their soldiers as teamwork is an important part of being a member of the armed forces. Persuasiveness, as it relates to the armed forces, concerns the ability to lead a team, and to be respected and accepted by that team so that subordinates will perform tasks they may not wish to do.

Situational tests involve placing the applicant in a situation where a senior manager is away for the day and the applicant must take over the manager's work. The applicant is required to sort through the contents of an in-tray and prioritise the tasks (Goodstein & Lanyon, 1999). This task is essentially a problem-solving task, and gives assessors opportunities to see how candidates come up with solutions. Again, problem-solving ability in the armed forces is very important, particularly in a wartime situation where dangerous situations may require quick-thinking. In the case of a job applicant, the resulting information gathered from the assessment centre tests and tasks is used to decide whether or not to hire the person. In the case of management potential, the information is used to decide whether the individual fits in with the organisational culture so that a management development plan can be written (Goodstein & Lanyon, 1999).

How, then, do assessment centres measure up as selection tools? Goodstein and Lanyon (1999) reported that assessment centres have shown validities of 0.37, and therefore, were valid means of selecting people. Eatwell (1998) reported a validity of 0.41 for an assessment centre, and claimed that, depending on the tools and procedures used, validity could range from –0.04 to 0.74. Similarly, Robertson (1993) reported validities obtained through meta-analyses of 0.41, 0.43, and 0.25 against performance and supervisor ratings. Borman (1982) conducted a study that sought to develop, run, and evaluate an assessment centre designed to measure potential for success as a US Army recruiter (Borman, 1982). The subjects were assessed on first impression, physical attractiveness, and likeability ratings, structured interviews, cold calls; interviews; interview with a concerned parent; five minute speech about the army; an in-basket exercise; and assessment of human relations, selling, organising, and overall performance (Borman, 1982). The ratings were correlated with criteria of scores on three tests measuring mastery of prospecting and selling techniques, and ratings of telephoning and interviewing techniques (Borman, 1982). The validity for the exercises was significant at 0.32, but the first impression, likeability, and physical attractiveness ratings were not significant (Borman, 1982). When the assessment ratings were unit weighted on each dimension and pooled across the six exercises, the validity for the exercises was higher (0.48). However, a range restriction occurred in that seven people dropped out who had either very high or very low ratings (Borman, 1982). When corrected for range restriction, the validity rose to 0.53 (Borman, 1982). Borman (1982) concluded that the assessment centre was reasonably successful in predicting training performance in a military sample.

Perkins (1998) investigated whether High Extraversion and Low Neuroticism could predict leadership ability in an assessment centre used for selecting British Army officers. Perkins (1998) hypothesised that High Extraversion and Low Neuroticism would show significant positive correlates with passing officer selection and with scores on the Regular Commissions Board (RCB) performance dimensions. Also hypothesised was that significant correlates would exist between individual item scores on a personality questionnaires and passing officer selection and scores on the RCB dimensions (Perkins, 1998). The OCEAN, a personality questionnaire based on the FFM, was administered to officer candidates during a pre-RCB selection phase. The candidates were told the results would not affect officer selection (Perkins, 1998). Those who passed this initial phase went on to complete the RCB, an assessment centre of three days duration in which performance on written tests, group

discussions, analysis and planning, lecturettes, leaderless group tasks, command tasks, obstacle course, interviews, and a race were assessed (Perkins, 1998).

Perkins (1998) did not find any significant correlates between High Extraversion, Low Neuroticism, and passing officer selection. However, OCEAN facets showed a negative correlation between worrying, shyness and passing officer selection (Perkins, 1998). Furthermore, no significant correlates existed between High Extraversion and Low Neuroticism and RCB dimensions, although some of the subfactors (Worrying, Shy and Bashful, and Socially Active) did significantly correlate with some of the dimensions. Scores on individual items of the OCEAN Extraversion and Neuroticism scales did, however, display significant correlates with passing officer selection and with RCB performance dimensions (Perkins, 1998). Perkins (1998) also found that the OCEAN did not detect faking and impression management which he gave as one explanation for why the OCEAN only correlated with passing officer selection and RCB dimensions at the item level. Perkins (1998) concluded the Big Five was not adequate for selection purposes as some of the factors were confused and the global factors did not predict passing officer selection or RCB dimensions as well as the sub-factors. Assessment centres, then have been shown to demonstrate good validity depending on the exercises and measures used. The NZ Army use an assessment centre for their officer selection process.

The New Zealand Army Officer Selection Board (OSB)

The NZ Army OSB is used to select candidates for officer training. Assessors are comprised of Military Testing Officers (MTOs) and Army Psychologists. The selection process consists of a pre-selection phase in which candidates complete a cognitive ability test, an essay-writing test, a medical screening form, and an interview with a recruiting officer. If successful at a "paper" selection board, where the Senior Psychologist (Army) and the Military Secretary decide on the basis of the pre-selection exercises who is able to continue to the next stage, the candidate then attends the 4½ day OSB.

At the OSB, the candidates are divided into groups called syndicates. The first 1½ days consists of group and individual indoor and outdoor activities designed to allow assessors to assess applicants on criteria found, through British Army and Australian Army

officer job analyses, to be important in the officer role. Peer ratings are gathered, personality questionnaires are administered and interviews are conducted during the OSB. Successful applicants later attend a medical board.

The final day features the final board meeting where the MTOs present their gradings on the selection criteria and the buddy ratings and then rank orders them. The MTOs then give their gradings of In or Out for each candidate. The psychologists, deputy president, and president give their gradings of In or Out and these are all discussed to decide finally who has been successful. The president then informs each candidate of the decision. The MTOs are trained in debriefing unsuccessful candidates and feedback is provided on strengths and weaknesses to both successful and unsuccessful candidates. This latter part is important as the feedback given to successful candidates can assist them to work on particular areas prior to commencing officer training.

The NZ Army officer selection process, then, does not rely solely on the basis of personality questionnaire results. The OSB process utilises information across the range of exercises that make up the OSB; information across all components is used to assess candidates against the criteria identified from job analyses as relevant to future job performance as a junior officer.

Summary and implications

These findings suggest that, in general, assessment centres are a valid method of selection. The validity is affected by the nature of the criteria used, the type of exercises, and whether peer ratings are used. Furthermore, assessment centres are quite expensive to run, therefore, only larger organisations, in particular, the military, tend to put in the time and resources. The NZ Army officer selection procedure involves the use of an assessment centre, at which time the EPQ-R and GPP-I personality questionnaires are administered. However, personality questionnaire use is still controversial. Later research has attempted to remedy the problems associated with using personality questionnaires in selection. Guion and Gottier (1965) highlighted the need for an adequate organising structure of personality dimensions to increase their validity in selection. Research has shown the Five Factor Model (FFM) to be a promising taxonomy of personality traits for selection purposes.

The Five Factor Model (FFM)

Why were taxonomies of personality characteristics considered important for selection? For one thing, they organised personality traits so as to provide a greater understanding of the behaviours that made up those traits. Further research could then determine the relationship of these traits to job-relevant criteria. Consequently, people-requirements of jobs could be defined. For officer selection, the selectors could target those individuals who possessed the personality characteristics required to succeed in the officer role. The FFM is a widely-used taxonomy useful for describing normal personality and, in particular, has been shown to predict job performance criteria. The FFM was empirically-derived through a series of factor analyses of adjectives commonly used to describe individuals' characteristics. Following factor analyses of these trait adjectives by Cattell in 1945, Fiske in 1949, and other researchers, five factors have consistently emerged (Goodstein & Lanyon, 1999). Measures of the Big Five, such as the NEO-PI-R used in Black's (1997) research have divided the five factors into subfactors, or facets. The five factors were commonly labelled Neuroticism or Emotional Stability, Agreeableness, Conscientiousness, Extraversion, and Openness to Experience or Intellect (Goodstein & Lanyon, 1999). Typically, Neuroticism measured how anxious, or stable a person was, Agreeableness reflected co-operation, trust, goodnaturedness; Conscientiousness reflected traits such as responsibility, dependability, hard-working, and organised; Extraversion referred to how sociable, assertive, talkative, ambitious a person was; and Openness referred to how imaginative, creative, artistic and intellectual a person was (Goodstein & Lanyon, 1999).

In particular, Neuroticism and Extraversion are two factors measured by the EPQ-R. The GPP-I measures Ascendancy and Sociability which many researchers have argued should be combined as they have been shown to measure Extraversion (Gordon, 1993). The GPP-I also measures Emotional Stability, and Original Thinking, which possesses similarities to the FFM Original Thinking scale. GPP-I Responsibility may be likened to Conscientiousness as it measures traits such as reliability, dependability, and perseverance. Therefore, research into the strength of the relationship between these scales and future behaviour was particularly pertinent to the current research, as the current study sought to determine whether the EPQ-R and GPP-I scales used in officer selection measured what they claimed to measure. For officer selection purposes, they needed to be predictive of behaviour

associated with the NZ Army officer role. Thus, FFM research can provide evidence to support the use of Extraversion, Neuroticism, Responsibility, and Original Thinking in selection settings.

Since Guion and Gottier's (1965) discouraging conclusions, there have been at least two comprehensive meta-analyses which have addressed the validity of personality questionnaires based on the FFM for selection purposes. These studies have shown that the FFM factors were useful for predicting future performance and behaviour at work. Barrick and Mount (1991) performed a meta-analysis of studies that assessed the Big Five dimensions' ability to predict three job performance criteria: job proficiency, training proficiency, and personnel data for five occupational groups: professionals, which included doctors, teachers, engineers and so forth; police; managers; sales; and skilled/semi-skilled people. In their study, Barrick and Mount (1991) hypothesised that Conscientiousness and Emotional Stability would be the most valid predictors for the three criteria in all five occupational groups. In addition, Barrick and Mount (1991) hypothesised that Extraversion and Agreeableness would be predictive of successful performance in jobs that required high people contact such as sales or management. Finally, Barrick and Mount (1991) hypothesised that Openness to Experience would predict training proficiency as it measured traits such as curiousity, broad-mindedness, and intelligence - traits linked with positive attitudes towards learning.

Criteria used were job performance ratings, training performance ratings, productivity data, and personnel data such as salary level and turnover (Barrick & Mount, 1991). They found that Conscientiousness was, indeed, a valid predictor for all occupational types (p ranged from 0.20 to 0.23), but Emotional Stability was not (p ranged from –0.13 to 0.12). Extraversion was found to be predictive in sales ($p = 0.15$) and management ($p = 0.18$) groups but Agreeableness was not ($p = 0.00$ and 0.10 respectively). Openness to Experience did predict the training proficiency criteria ($p = 0.25$) and Extraversion was also a significant predictor ($p = 0.26$). Barrick and Mount (1991) cautioned that the validities were, more than likely, underestimated as only studies reporting zero-order correlates and average correlates were included. Despite these caveats, the study showed that personality dimensions, in particular Conscientiousness, were predictive of job performance criteria. As the NZ Army officer role can be likened to a managerial role, the correlation between EPQ-R Extraversion and the WBQ Extraversion should demonstrate whether Barrick and Mount's (1991) finding

were supported in the current research. Furthermore, if EPQ-R Neuroticism and GPP-I Emotional Stability were not found to predict workplace behaviour, this would support Barrick and Mount's (1991) finding that FFM Emotional Stability was not a valid predictor in occupation settings.

Similar research has been conducted using NZ police recruits. These studies provide an interesting comparison with the present research as the police officer role shows similarities to the army officer role. In particular, physical training and rank structure in the police force are similar to the army. The NZ Police and the NZ Army do, in fact, work together closely on some aspects such as bomb disposal, search and rescue, crime-scene searching, crowd control, and counter-terrorist activities. Consistent with Barrick and Mount's (1991) work, Black's (1997) study using NZ police recruits found that Conscientiousness correlated the highest next to cognitive ability with job performance (0.27), and Extraversion correlated significantly and positively with job performance (0.16). Contrary to Barrick and Mount's (1991) research, Black (1997) found that Neuroticism correlated significantly and negatively with job performance (-0.16). Black (1998) assessed the differences in personality profiles using the NEO-PI-R between unsuccessful applicants and successful applicants for entry into the NZ Police. Like Barrick and Mount (1991), Black (1998) hypothesised that unsuccessful applicants would score higher on Neuroticism and lower on Extraversion than successful applicants. As a result of independent t-tests, Black (1998) found that unsuccessful applicants did score higher than successful applicants on Neuroticism, but there was no difference between the two groups on Extraversion scores. However, successful applicants were found to score lower on Agreeableness than unsuccessful applicants (Black, 1998). In the current study, low levels of WBQ Neuroticism and WBQ Emotional Stability in the officer sample would provide further support for Black's (1998) conclusions.

Summary and implications

The Big Five was one answer to the problem of an inadequate taxonomy of personality traits and dimensions. When questionnaires were based on the Big Five, they were shown to predict job performance and training performance criteria with reasonable accuracy. Therefore, these personality questionnaires could be useful tools for predicting

how well a job applicant will perform on the job. The EPQ-R and GPP-I used by the NZ Army include some scales that are similar to the FFM scales. For example, Neuroticism is represented by EPQ-R Neuroticism and GPP-I Emotional Stability (reverse scored). Therefore, FFM research findings that link personality traits with job performance can be further supported if the similar scales in the current research also demonstrate links with workplace behaviour. It is not enough, however, simply to have an organised structure of personality dimensions or traits. Each trait within the taxonomy must be carefully defined so that questionnaires developed can accurately measure those traits. Construct definition is an important step in the process of improving the validity of personality questionnaires for selection purposes.

Personality Questionnaires and Construct Validity

In a selection context, decisions are made concerning an individual's career based, in part, on the inferences drawn from the questionnaire used. If a questionnaire does not measure the underlying construct that it should, considering the theory, then the data obtained from the measure is useless. Demonstrating construct validity, then, is the essence to ensuring a useful questionnaire. The current research is but one study that tests the ability of the EPQ-R and GPP-I to measure what they should measure. The WBQ items were developed to measure behaviour associated with the EPQ-R and GPP-I personality scales. The WBQ assessed whether the EPQ-R and GPP-I measured the constructs they were supposed to measure so that the questionnaires could be seen as useful tools for officer selection in the NZ Army. In a selection setting, one could not have confidence in personality questionnaire results if the questionnaire was not an accurate measure of the personality construct. Similarly, officer selection would not be accurate if the personality questionnaires used did not measure the desired traits. This problem is minimised if the trait can be clearly defined through well-designed research. Messick (1988) argued that, although other types of validity exist, construct validity is the most important. The *Standards for Educational and Psychological Testing* (American Psychological Association (APA), 1985) defined validity as "...the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores." (p. 9). That is, validity refers to actual interpretations made from the test scores gained, as opposed to the validity of the test itself.

Hershberger (1999) argued that constructs interpret and categorise response consistencies on test items, hence, defining constructs was critical in order to understand test responses. This point was particularly important when the measure was used for personnel selection because these decisions had implications for selecting capable employees and ensuring their use was consistent with equal employment opportunity (EEO) policies. EEO issues concern fair employment practices, in particular, ensuring that a selection procedure did not discriminate against certain groups in society, such as people with disabilities, people with families, pregnant women, and racial groups.

In order to assess construct validity, the construct should be defined clearly based on existing theory and research. Parker (1993) claimed that if constructs were not defined properly, the resulting measures and methodology used would not be defined properly and, therefore, would not show a good relationship to the construct under investigation. The lack of a sound relationship meant that a good interpretation of the construct could not be adequately inferred from the results (Parker, 1993). Concerning the present study, if the EPQ-R and GPP-I scales were not based on well-defined constructs that related to personality, then they could not be deemed an accurate measure of personality. Hence, they would not be useful as selection tools for the NZ Army. Clark and Watson (1995) argued that the most precise and efficient measures were based on well-articulated theories that were supported by good empirical evidence. As such, one must first conceptualise the construct precisely and in sufficient detail in an appropriate context (Clark & Watson, 1995). The WBQ measure, then, was developed to assess whether the EPQ-R and GPP-I measured the construct they purported to measure.

Furthermore, to eliminate desirability and irrelevant scale variance, convergent and discriminant validity evidence should be provided at the earliest stage of item development. For example, Ozer and Reise (1994) described an approach where a researcher began with a vague idea of a particular construct, generated a large number of items to measure the construct, collected data on the construct, and revised the original theory. This process continued until the construct had been defined accurately, displaying acceptable levels of convergent and discriminant validity (Ozer & Reise, 1994). Clark and Watson (1995) provided an overview of the scale construction process incorporating both theoretical and empirical approaches. They argued that the construct should first be defined using nomological nets following a thorough literature review to determine the construct and other

constructs that might be closely related. Nomological nets referred to a collection of behavioural statements that defined a theory (Orton, 1987). Once construct definition occurred, the construct should be validated by empirical testing methods. This involved creation of an item pool thought to measure the construct and related constructs, followed by empirical testing using well-designed studies to refine the item pool (Clark & Watson, 1995). The EPQ-R was developed in a similar manner to that suggested by Clark and Watson (1995). It was based on psychophysiological theories that linked Extraversion to arousal, Neuroticism to lability of the nervous system, and Psychoticism to the level of circulating androgens. Items were developed through factor analyses and other empirical methods to measure these constructs. On the other hand, the GPP-I was not based on a particular theory, but developed through empirical and factor analytic means. The WBQ was developed based on behaviour linked to the personality construct behind each scale, with items refined in conjunction with army psychologists.

Controversy has surrounded which of the two methods behind the EPQ-R and GPP-I was better. Scale development processes were particularly important to ensure accurate definition of the construct to be measured and, hence, identification of an accurate item pool. These factors would then enhance the construct validity of the questionnaire scores. Early researchers have argued for and against both methods of scale construction. Jackson (1971) was a strong proponent for theoretically-derived scales. He argued that theory was highly important in scale construction as construct validity was demonstrated through the construct's link to theory. Other arguments were centred on whether measures should be developed purely based on theory, purely through empirical-criterion keying methods, or a combination of the two. It was generally agreed that blind empiricism was not useful and did not result in highly valid measures. The empirical scales may only be better if they were first based on construct definitions, such as seen in the EPQ-R.

Rationally-derived scales, or theoretically-derived scales, began by defining the construct based on a particular theory, followed by generating item content reflecting that theory (Ozer & Reise, 1994). Empirically-keyed scales served to differentiate responses given by two different groups on one criterion (Ozer & Reise, 1994). Scales using the empirical approach were developed by generating items based on factor analytic and correlational evidence from previous studies. These scales were then tested to see which items differentiated between two groups. As such, empirically-derived scales could show

higher validity coefficients as they were more statistically sound. Other researchers, however, argued that theoretically-derived scales could show just as good, if not better, validity as empirically-derived scales. Jackson and Paunonen (1980) argued that rational scales constructed with little effort proved to be as good as elaborately-constructed empirical scales. Research suggests that a combination of the two approaches may be the most accurate way of developing a scale.

Jackson (1971) gave an example of the Personality Research Form (PRF), a theoretically-derived instrument, where very few of the irrelevant items (those written for a different scale) correlated more highly than those directly written for the particular scale. He believed that items that bore no relation to theory but were included in empirically-derived scales were, more than likely, admitted due to errors in initial item selection (Jackson, 1971). For example, in the MMPI, an empirically-derived instrument, some items on the same scale showed negative correlates with each other (Jackson, 1971). Jackson (1971) did, however, concede that some empirical or "subtle" items were needed. The endorsement rate for an item such as "I enjoy torturing animals" as a measure of sadistic impulses would, quite likely, be fairly low (Jackson, 1971).

A study conducted by Knapp and Fitzgerald (1973) provided empirical evidence for theoretical scales. The aim of their study was to examine whether the Personal Orientation Inventory (POI) could predict personality change following an encounter experience. They administered the POI, a theoretically-derived inventory, to a group of Navy personnel volunteers before, and between one and eight months after, an encounter experience. In addition, empirically-derived POI scales based on earlier criterion and factor analytic studies were administered. Five of the twelve rationally-derived scales showed significantly higher means at post-test compared with three of the nine empirically-derived scales. Contrary to Jackson's (1971) findings that theoretically-derived scales did not show item content overlap, Knapp and Fitzgerald (1973) found that the empirically-derived items showed a mean intercorrelation of 0.26 compared with 0.47 for the rationally-derived scales. Although Knapp and Fitzgerald (1973) did not make conclusions regarding the two types of scales, their data showed that the rationally-derived scales were more effective in predicting post-encounter group change than the empirically-derived scales.

Summary and implications

In order to ensure good construct validity, a measure must be based on sound theory that has been empirically tested. Only then could a test user be confident that inferences drawn from test scores were accurate, leading to the right hiring or placement decision. Centring efforts on construct definition led to clear, well-articulated, and robust traits that could then be used accurately to assess personality. As a result, a selector would have a more clear idea of the characteristics of the job applicants and could rely on the data to provide an accurate picture. Arguments for and against theoretically-derived versus empirically-derived scales have centred around which of the two showed greater validity. Generally, a combination of the two where specific hypotheses about the construct were developed based on theory and then empirically-validated has been found to be the most acceptable method of scale development. The EPQ-R was developed in this way, whereas the GPP-I was developed through empirical means only. The EPQ-R, then, may be a more robust measure of personality than the GPP-I. However, before firm conclusions can be drawn regarding which questionnaire is better, further factors exist that affect the robustness of a questionnaire.

Personality-Relevant Criteria

One such factor concerns the relationship between the personality traits and the criteria to be predicted. In particular, validity of scores from personality questionnaires were improved when the criteria were conceptually-linked to the personality questionnaires. Demonstrating this relationship provided further evidence that a personality questionnaire was useful for predicting people-related behaviours required for a certain position. For officer selection, the personality questionnaires used should be conceptually-linked to the people requirements of the officer role, such as the Personal Discipline, Effort and Leadership, and Physical Fitness and Military Bearing traits identified by McHenry et al. (1990). Schneider, Hough, and Dunnette (1996) claimed that the best way to enhance criterion-related validity which, in turn, provided evidence for construct validity, was to determine relevant performance dimensions for a given job and then to link specific personality traits to those specific performance dimensions. In essence, then, to enhance the validity of personality questionnaire data in selection contexts, the constructs underlying the questionnaire should be linked theoretically and empirically to the construct underlying the

criteria. Standard 10.8 in the *Standards* (APA 1985) stated clearly that two links were required if construct validation was the major support of validity in personnel selection:

> "...First, there should be evidence for the validity of the test as a measure of the construct, and second, there should be evidence for the validity of the construct as a determinant of major factors of job performance. There should be a clear conceptual rationale for this linkage. Both the construct and the job factors to which it is linked should be defined carefully. A consistent pattern of results should point toward the hypothesized relationship..." (p. 61).

Construct definition, then, should occur for both predictor and criterion constructs based on theory and followed up by empirical testing. In particular, specific hypotheses should be developed to provide evidence for the link between the two constructs. For example, Hough (1992) described an earlier study (see Hough, Eaton, Dunnette, Kamp, & McCloy, 1990) assessing the relationship between nine personality factors and five criteria of job proficiency, training success, educational success, commendable behaviour, and law-abiding behaviour. Hough (1992) described the development process of the ABLE, including an extensive literature search and review to gather information about the relationship between the personality scales and external criteria. The validities found were summarised within each predictor-criterion construct combination and the ABLE scales were correlated with the five job performance criteria (Hough, 1992). Uncorrected correlates of the personality constructs with the job proficiency critierion ranged from -0.03 to 0.18 and, when corrected, were not much higher (Hough, 1992). However, Hough (1992) found that the criterion-related validities were increased due to the conceptual linkage of the predictor-criterion construct and that all nine personality constructs showed different relationships with the five performance criteria. In particular, Achievement was shown to be the best predictor, with correlates ranging from -0.24 to 0.27 for each of the job performance criteria (Hough, 1992). It is worth noting at this stage that, although Hough (1992) saw the importance of explicitly conceptualising the link between each personality construct and the criterion constructs, this was not actually performed in the study described. Robertson (1993) argued that this resulted in a mixture of large and small validity coefficients, leading to low overall validity. In a similar vein, Hogan and Roberts (1996) argued that, when any scale is used to predict any criteria, very few links occur. However, if single construct measures are used to predict relevant criteria, validity increases (Hogan & Roberts, 1996). In the current research, then,

linkages between each scale of the EPQ-R and GPP-I should be demonstrated for each relevant scale of the WBQ.

Binning and Barrett (1989) conceptualised the links between predictor and criterion, using job performance as an example of a criterion, in the following diagram:

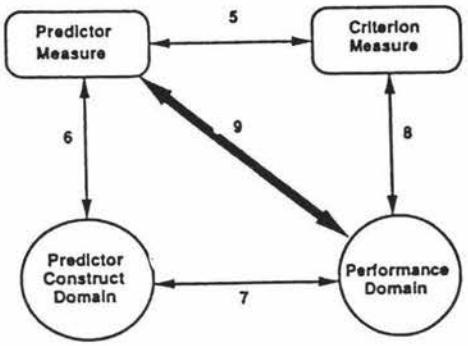

Figure 1. How inferences are commonly viewed in personnel selection (Binning & Barrett, 1989, p. 480).

The diagram shows the predictor measure to be related to the criterion measure, and is a sample from a psychological construct domain; the predictor construct domain overlaps with the performance domain; the criterion measure is a sample from the performance domain; and the predictor measure is related to the performance domain (Binning & Barrett, 1989). The bold arrow between the predictor measure and the performance domain emphasises the greater importance of this link. Using the EPQ-R as an example of a predictor measure that samples the predictor domain of personality, and the criterion of job performance, this link says that the EPQ-R should predict job performance, because its underlying construct of personality is linked to job performance.

One problem with providing conceptual linkages was that the resulting scales designed to measure each construct could overlap in item content. Any correlates between the two may, therefore, only reflect similar item content rather than any relationship between the two constructs (Nicholls, Licht, and Pearl, 1982). However, Friedman (1983) argued that item similarity did not affect construct validation too severely, providing that additional evidence of construct validity associated with the face valid items could be obtained. That is,

face valid items also need to display convergent and discriminant validity, and if they do, there is no reason why they should be omitted simply because they contain similar content to other items.

Job analyses

How, then, can item selection and scale development for a job performance criterion take place? The *Standards* (APA, 1985), under Standard 10.4, stated that, in job selection, characterisation of the content domain should be based on a thorough job analysis. Other research supports this view. Robertson (1993) pointed out that several personality variables were often measured at the selection stage, and they were better linked to specific job competences than to overall performance. Again, selection of traits should be linked through empirical and theoretical evidence to personality requirements identified through job analyses. An understanding of the nature of the job is an important consideration in determining which personality characteristics are required. Similarly, in a military setting, a delineation of the officer role together with personality characteristics required is invaluable in determining the conceptual linkages between personality constructs. Undertaking job analyses to define the role will further improve the usefulness of personality questionnaires in personnel selection. For example, Binning and Barrett (1989) described the selection process as involving a job analysis to determine the appropriate behaviours that make up the relevant performance domain which then guides the selection of appropriate measures that will accurately predict how well an applicant will perform. Criterion measures should be developed with the values, vision, mission, and goals of the organisation in mind as the selection process aims to select those candidates who display the right behaviours deemed necessary for success in the organisation (Binning & Barrett, 1989).

Tett, Jackson, and Rothstein's (1991) study involved conducting a meta-analysis of studies that assessed the validity of personality questionnaires in selection. They analysed studies according to the type of analysis involved: whether a confirmatory or exploratory factor analysis was used, and whether the measure had been developed from a job analysis. Textbook definitions of confirmatory factor analysis stated that this type of analysis referred to factors derived following testing of specific hypotheses (Murphy & Davidshofer, 1998). This definition could be likened to empirically validating theoretically-derived scales.

Exploratory factor analysis served only to find the best statistical fit to the data obtained (Murphy & Davidshofer, 1998) and could be likened to empirical-criterion keying methods. Tett et al. (1991) found that correlates obtained from confirmatory as opposed to exploratory factor analyses were higher, and where a job analysis was used, correlates were higher still (ρ = 0.38, corrected for unreliability) (Tett et al., 1991). Furthermore, they argued that personality traits were more widespread than cognitive ability and were also less intercorrelated. This meant that personality measures need to be developed for each job type, from a job analysis that outlined the specific personality requirements for success on the job. Again, the personality traits required should be based on conceptual linkages with performance criteria. Tett et al. (1991) argued that:

> "...the full potential of personality traits in personnel selection will be realised only when confirmatory research strategies employing personality-oriented job analysis become the standard practice for determining which traits are relevant to predicting performance on a given job and when greater attention is directed to the selection of psychometrically-sound construct valid personality measures." (p. 732, Tett et al., 1991).

Other research has also demonstrated the worth of using job analyses for criterion development, with meaningful results despite a small sample size. Day and Silverman (1989) sought to determine important personality characteristics for effective job performance in accountants. They argued that the reluctance to use personality questionnaires in selection may have stemmed from poor methodological designs. More importantly, however, they believed that differences in validities may occur in different occupational groups (Day & Silverman, 1989). They argued that, when personality characteristics were matched with the occupation and organisation, validities were likely to be higher (Day & Silverman, 1989).

Their study aimed to demonstrate links between the personality variables of work orientation, degree of ascendancy, and interpersonal orientation, and performance components based on theories that linked the two (Day & Silverman, 1989). They hypothesised that impulse expression and control, orientation toward direction from others, and intellectual and aesthetic orientation should not be related to job performance components (Day & Silverman, 1989). They did not specify, however, which personality variables should be related to which performance components as recommended by Hogan and Roberts (1996). A job analysis comprised of interviews and the critical incidents method

were used to identify important job performance criteria of potential for success, technical ability, timeliness of work, client relations, cooperation, and work ethic (Day & Silverman, 1989). An overall performance score was also derived. A sample of accountants were administered the Personnel Classification Test (a cognitive ability measure) and scales derived from the Personality Research Form, prior to any employment decision being made. Their grade point averages were also collected as a cognitive ability measure. Data was collected until an adequate sample size (n = 43) was attained to account for effect size. No selection decisions were made on the basis of these scores, however, Day and Silverman (1989) did not state whether the accountants were aware of this. The three scales hypothesised not to predict any job performance criteria were, in fact, found not to predict the criteria. Furthermore, the cognitive ability measure accounted for significant amounts of variance for the technical ability criteria only; the personality dimensions accounted for significant increases in explained variance (Day & Silverman, 1989). Finally, they found that work orientation, degree of ascendancy, and degree and quality of interpersonal orientation were significantly related to supervisor ratings of at least three performance dimensions (Day & Silverman, 1989).

The present study employed a slightly larger sample size than Day and Silverman's (1989) sample. Theoretically, then, it should be possible to obtain similar results. However, the current study focused on looking at behaviour associated with personality scales to address whether the EPQ-R and GPP-I measured what they claimed to measure. Currently, the NZ Army use a British Army officer job analysis together with a list of criteria important for success in the officer role developed from this job analysis and an Australian Army job analysis. It was decided not to validate the personality scales against these criteria as a multi-method job analysis of the New Zealand army officer role is currently being developed as part of a large-scale review of the officer selection process. Part of this review includes distribution of the WBQ questionnaire used for the current research, together with additional items developed by the NZ Army that addressed behaviour relevant to OSB criteria, cognitive ability test scores, and further items concerning the officer's temperament, leadership ability, motivation, and academic ability. It was felt that research directed at validating the personality questionnaires against performance criteria should wait until the NZ Army had completed this review. Therefore, the links between the EPQ-R and GPP-I scales and criteria relevant to the officer role have not yet been conceptually-defined.

Research has shown (Tett et al., 1991; Day & Silverman, 1989), then, that personality measures based on a well-defined theory and linked to specific job competences through the use of job analyses showed significant validities. These findings were a far cry from those reported by Guion and Gottier (1965) and provided support for the use of personality questionnaires as effective predictors of workplace behaviour. However, one main issue concerning performance measures and criterion development is that organisations are often unable to undertake a full, rigorous development procedure that adheres to research principles. For one thing, such an undertaking is time-consuming and only the larger organisations may deem an exercise worthwhile. Furthermore, such a project can be costly in terms of personnel required and resources needed. Whether managers have the knowledge and skill to undertake criterion development as a part of job analysis is an important issue and, unless money is available for outside consultants, organisations must often make do with the resources they have. These resources may be limited and mean that a rigorous job analysis followed by carefully-developed and tested performance criteria may simply not be feasible.

Summary and implications

In summary, then, the validity of personality questionnaire data used for selection purposes will be enhanced if the questionnaire is based on a well-defined theory. Furthermore, the conceptual linkages between the personality questionnaire data and the criterion the data is hypothesised to predict needs to be supported by theoretical and empirical research. In particular, validity is enhanced if the criteria are personality-relevant. One way of developing personality-relevant criteria is through job analyses, which only large organisations may undertake. As a large-scale review of the officer role is currently underway, of which this research forms a part, the current research focuses on developing personality-relevant items that reflect workplace behaviour thought to represent each personality profile in the EPQ-R and GPP-I.

Context

Despite using carefully-developed measures based on sound theory and linked to personality-relevant criteria, other factors affected the validity of personality questionnaires in selection. One of these factors concerned the context for which the personality questionnaire was designed. Researchers may get caught in the validity generalisation trap where a questionnaire designed for one purpose was used in a previously-unvalidated setting. The resulting validities, not surprisingly, were often weak (see Guion & Gottier, 1965). For example, when personality questionnaires designed for use in clinical contexts were used in employment situations; the resulting validities were low (Hogan, Hogan, & Roberts, 1996). More to the point, the questionnaires commonly in use have not been standardised with applicant populations (Goodstein & Lanyon, 1999). While individuals with severe psychological disorders would clearly not be desirable in an organisation setting, employers were more concerned with normal personality characteristics and how they affected the individual's behaviour on the job (Goodstein & Lanyon, 1999). Goeters, Timmerman, and Maschke (1993), argued that most applicants did not suffer psychological disorders, therefore, personality questionnaires used in selection should not be used to diagnose these disorders, but rather, to assess how the applicant reacted to stress, their work orientation and their sociability. The NZ Army use the EPQ-R for this purpose, in particular, to determine how well officer applicants cope with stress, whereas the GPP-I is used to determine the work-oriented aspects of personality. Hogan (1991) gave an example of the MMPI, an inventory measuring a large number of traits and initially designed for studying mental illness, or psychopathology. Hogan (1991) believed that many people thought the MMPI to be the prototypical inventory, however, it performed poorly as a selection device. The reason being? It was not designed for use in selection contexts, and in fact, only one of its scales, Adjustment, bore any resemblance to a predictor in selection (Hogan, 1991).

Qualls and Moss (1996) argued that if standardised personality questionnaires were to be used in a number of situations, the usefulness of any information gained would be dependent on the context in which it was used and the abilities of the test user. In particular, if decision-making was to occur as a result of test scores, the presence of relevant empirical evidence supporting the test's use in that particular situation must be documented (Qualls & Moss, 1996). Goeters et al.(1993) argued that the Personality Research Form (PRF)'s

reliability declined in an applicant situation. They administered the PRF and the Temperament Structure Scales, an inventory designed specifically for the purposes of selecting aviation personnel, as part of an assessment centre for selecting people for pilot training. They found that, although the PRF reliability significantly decreased ($\alpha = 0.60$ is the cut-off) under selection conditions, the TSS reliability did not. This finding could be explained partly because the TSS was specifically designed for selection, and partly because the PRF may have been more susceptible to faking than the TSS.

Binning and Barrett (1989) argued that, in personnel selection contexts, the types of inferences involved in selection process decisions and the nature of the evidence to support those inferences must be defined. Validity will be higher, then, if a measure is developed specifically for the situation in which it is to be used. In the current research, validity will be further improved if evidence can be found to support the use of the EPQ-R and the GPP-I in military settings, as well as selection settings. In a letter from the New Zealand Council for Educational Research (NZCER) (dated 30 October 1998) to the NZDF concerning the use of the EPQ-R, the NZCER stated "However good a test is in psychometric terms, if it does not measure the domain of focus, then its use is invalid. The EPQ-R is a personality inventory. There is an abundance of empirical and theoretical support for the notion that information about personality can be usefully related to vocational decisions." The EPQ-R, then, needs to be validated to ensure that it measures behaviour associated with personality in a selection setting. The EPQ-R manual (Eysenck & Eysenck, 1991) did not report whether the EPQ-R had been validated for military contexts, however, the GPP-I manual (Gordon, 1993) contained several studies supporting the use of the GPP-I in military contexts. The NZ Army developed their own norms for the EPQ-R and GPP-I based on an applicant population of approximately $n = 1500$ which are updated every six months, are specific to each group being assessed, and are used to assess any differences between groups. However, at the time the current research was conducted, these norms were not available and, therefore, could not be used. The EPQ-R was not specifically designed for use in selection or military contexts, rather, it was designed to measure what Eysenck and Eysenck (1991) believed to be the major dimensions of personality. The EPQ-R was intended, therefore, for a wide variety of situations to assess personality traits. It has not, however, been validated as a measure for use in personnel selection against job performance criteria. Rather, it is a questionnaire that seeks to provide information concerning the make-up of an individual's personality, in particular,

highlighting predispositions towards malfunctioning (Eysenck & Eysenck, 1991). The WBQ was also not developed using job performance criteria as it was intended to measure behaviour associated with personality in a selection context. The current study, then, is one study that seeks to validate the WBQ for selection purposes.

Summary and implications

In summary, then, the validity of test scores from personality questionnaires is increased if the questionnaire is specifically designed and/or validated for the purpose in which it is to be used. A personality questionnaire that was specifically designed to be used in selection settings, was likely to show more valid inferences than one designed for clinical use but used in selection. The EPQ-R was designed to pinpoint individuals who might react under stress with emotional difficulty, whereas the GPP-I contains many work-related items and has been well-validated in industrial and military settings. Therefore, the EPQ-R should be useful in screening out individuals who do not cope well with stress and the GPP-I should provide a general picture of an individual's personal style on the job. Validation in selection contexts, however, is not always an easy task. A major argument against the use of personality questionnaires in personnel selection concerns the ease with which personality questionnaires can be faked in these contexts.

Faking

Faking represented what was commonly termed socially-desirable responding where, rather than responding to item content, people responded to items according to whether they were socially acceptable or not (Hogan, 1991). For example, most people would regard the item "I often have strange and unusual thoughts" as socially undesirable and would respond accordingly (Hogan, 1991). Paulhus (1986) distinguished between impression management and self-deception. He claimed that impression management referred to a conscious decision to distort responses, to lie, in the hope of gaining something desired, such as a job. Self-deception, on the other hand was unconscious, and was aimed at protecting one's self-esteem (Paulhus, 1986). Similarly, Leary and Kowalski (1990) distinguished between impression motivation, where an individual was motivated to control how others see them, and

impression construction, where, once motivated, an individual would alter their behaviour to present a different image to others. They stated that, job applicants were likely to make sure they presented themselves in a more positive light, that showed they could handle the job easily, and that they had the ideal employee attributes (Leary & Kowalski, 1990). Likewise, Hogan et al. (1996) claimed that an individual would use the way they answer a questionnaire to tell an unknown interviewer who they were and how they would like to be seen. In this way, responses to items were self-presentations, not self-reports.

Item content can affect the ease with which a personality questionnaire can be faked. This issue reflected the nature of scale development; whether a scale was rationally-derived or empirically-derived. Hough et al. (1990) contended that many studies showed the face-valid items demonstrated greater validity than the subtle items, and subtle items introduced as a scale may actually reduce the validity of the inferences drawn. This line of thinking was similar to Jackson (1971) who believed that items based on theory were more valid than those based on empirical methods of item selection. However, other studies have shown that transparent items, where test-takers guessed the meaning behind the item and responded accordingly, were one problem with rationally-derived scales, and led to faking of questionnaires. For example, Perkins (1998) pointed out that items as part of a lie scale often contained different content to personality items. The more intelligent candidates could pick up on this and answer these questions truthfully but not the personality items. One item on the Lie Scale of the EPQ-R was "Do you always wash before a meal," an item that is clearly not personality-related. Another transparent EPQ-R item, this time measuring the Neuroticism scale, was "Are you are worrier?" Subjects in a selection setting may be less likely to endorse these items.

Hence, in a situation such as applying for a job, applicants may be highly motivated to distort their responses to present the best possible image. What problems has faking posed? Faking may lead to incorrect hiring decisions, which may subsequently lead to increased attrition if the person does not perceive they fit in (Schneider, Smith, Taylor, & Fleenor, 1998). This may be costly to an organisation as selectors may then have to repeat the whole selection process. In the NZ Army, many resources, including time, money, and people are devoted to ensuring accurate selection. However, if part of the selection process is faulty leading to inaccurate decisions, then these resources may be wasted.

How prevalent, then, is faking? Much evidence exists to support the argument that response distortion could occur under high motivation conditions. For example, Mahar, Cologon, and Duck (1995) assessed whether applicants faked a questionnaire according to the stereotype of the job role. They administered the Myers-Briggs Type Indicator (MBTI) to a sample of undergraduate students and then a week later asked them to complete it again. In the first administration, they answered it in order to maximise their chances of getting a job as a psychiatric nurse; the second, they answered it to give the best possible impression of themselves (Mahar et al., 1995). Following another two weeks they again completed the MBTI responding as they thought a typical psychiatric nurse would (Mahar et al., 1995). The MBTI was also administered to a sample of psychiatric nurses. They found that subjects were able to distort their responses according to instructions, but were not able to fake the MBTI according to the perceived stereotype of a psychiatric nurse. However, the strategy used in attempting to fake a profile was that of a stereotypical view of the psychiatric nurse occupation – the fake-job profiles were very similar to the stereotype profile than to any other profile (Mahar et al., 1995). Therefore, it is likely that participants in the current study may have faked the personality questionnaires according to their perception of those personality characteristics desirable for the officer role.

In contrast, Hogan et al. (1996) argued that while people may be able to intentionally distort their responses when asked to do so, the actual level of faking in an applicant setting was low. However, Hogan et al. (1996) did not provide any empirical evidence to support their claim. Hogan (1991) reported a study which assessed personality inventories in personnel selection where subjects were divided into four groups: fake good-honest, honest-fake good, fake bad-honest, and honest-fake bad, and were asked to complete the inventories twice. Faking good was defined as trying to be selected, whereas fake bad was defined as trying not to be selected. The data was compared with applicants who were waiting to hear whether they had been selected. The study showed that, while responses could be distorted and detected with a validity key, applicants did not actually distort their responses (Hogan, 1991).

Although it has been shown that personality questionnaires can easily be faked in selection contexts, the implications of faking are not clear. Some researchers argued that faking would reduce the validity of the personality questionnaire data and, consequently, have added special scales to their questionnaires to detect socially desirable responding. The

EPQ-R is an example of one such questionnaire, as it contains a Lie scale designed to detect both impression management and self-deceptive enhancement. However, Costa and McCrae (1997) argued that adding such scales may actually reduce the validity. People endorsed such items because they genuinely believed they possessed these qualities; the endorsements, then, were an accurate reflection of their personality (Costa & McCrae, 1997). They believed the only real way to verify socially-desirable responding was through peer-reports or other outside sources (Costa & McCrae, 1997). Likewise, Paulhus (1986) argued that if socially desirable responding was not considered a nuisance variable, then trying to control it with social desirability scales served to reduce the instrument's validity (Paulhus, 1986). This, however, was the case with self-deception; impression management, on the other hand, led to skewed distributions particularly in clinical and personnel settings (Paulhus, 1986). Eysenck and Eysenck (1991) recommended examining the EPQ-R Neuroticism/Lie correlation and the Psychoticism/Lie correlation as a determinant of faking for the EPQ-R. They stated that if the correlation approached or exceeded -0.50, then it was likely faking had occurred. Lower, nonsignificant Neuroticism/Lie and Psychoticism/Lie correlates measured the stable conformity factor corresponding to self-deceptive enhancement. Other researchers have designed their questionnaires using a forced-choice format, such as the GPP-I, where statements of equal social desirability were paired together in one item. However, Hough et al. (1990) argued that this type of format has not reduced a subject's ability to fake responses. In fact, they discovered that people were still able to distort their responses in a forced-choice format when instructed to do so (Hough et al., 1990). Therefore, the use of scales and specific response formats to combat faking is controversial, and may even serve to reduce the validity of the questionnaire.

On the other hand, some researchers claimed that faking did not affect the validity of the personality questionnaire data at all. Tett et al. (1991) found, in their meta-analysis, that validity of the data gained from job incumbents was not significantly higher than the validity of data obtained from job applicants. Tokar, Fischer, and Subich (1998) reported a meta-analysis of the cumulated correlates between the Big Five and social desirability scores, and between job performance and social desirability scores. The meta-analysis also calculated a correlation between personality and job performance controlling for social desirability and found no effect on the predictive validity of the Big Five (Tokar et al., 1998). Barrick and Mount (1996) assessed whether self-deception and impression management affected the predictive validity of Conscientiousness and Emotional Stability in two samples of truck

driver applicants. They also sought to address whether the effect of response distortion was influenced by the type of response distortion. They administered the Personal Characteristics Inventory (PCI) and the Balanced Inventory of Desirable Responding (BIDR) as part of the selection process. The success of the hiring decision was measured by voluntary turnover and supervisor ratings (Barrick & Mount, 1996). They found that Conscientiousness was able to validly predict both voluntary turnover and supervisor ratings (p = -0.26 and 0.41 respectively in the first sample and –0.26 and 0.39 respectively in the second sample) as was Emotional Stability (p= -0.23 and 0.23 respectively in the first sample and –0.21 and 0.27 respectively in the second sample) (Barrick & Mount, 1996). Structural equation modelling showed that, although subjects did distort their responses through both self-deception and impression management, this did not affect the validity of Conscientiousness or Emotional Stability.

Hough et al.'s (1990) study also aimed to determine the validity of personality for predicting job performance, the effectiveness of validity scales, whether response distortion affected the validity, and the degree to which responses could be distorted (Hough et al., 1990). They designed a six-factor model of personality using a construct-oriented approach and developed the ABLE to measure it. As well as scales that measured the six personality constructs, the ABLE had four validity scales labelled social desirability, poor impression, self-knowledge, and nonrandom response (Hough et al., 1990). The ABLE was administered to enlisted army personnel and to personnel who had just been sworn in but not yet placed. They were told that the ABLE would be used to make placement decisions, so Hough et al. (1990) argued this was similar to an applicant setting. The criterion measures used were supervisory ratings; tests of school, job and soldiering knowledge; hands-on tests; letters of commendation; and other awards. They found that the ABLE predicted job performance well (validities were in the 0.20s, but were not corrected for range restriction or unreliability), lending further support to the use of personality questionnaires in selection. Hough et al. (1990) also found that soldiers did distort their responses and this was detected through the validity scales. However, the mean scores in the applicant sample were very similar to the scores obtained from the incumbents. Therefore, Hough et al. (1990) concluded that criterion-related validity was not affected by response distortion.

These studies contained certain methodological flaws that may render the conclusions dubious. First, the samples used may not be true applicant samples. Barrick and Mount's

(1996) study was based on a sample of long-haul truck driver applicants who were told at the time of administration of the questionnaires that the questionnaires would not be used for selection purposes. However, Barrick and Mount (1996) noted that some applicants, through comments made to the test administrator, believed the questionnaires would be used to hire them. This confusion may have lead to inaccurate results. Rosse, Stecher, Miller, and Levin (1998) criticised Hough et al.'s (1990) findings by saying their applicant sample did not reflect a true applicant population as the soldiers had already been sworn in and were only awaiting placement. Therefore, this could not be deemed a selection setting and consequently, faking could still affect the validity of personality questionnaires in selection.

Finally, although not a methodological flaw, Hough (1990) et al developed their own inventory to measure personality constructs through what appears to be purely empirical means. Using inventories that are not well-validated may lead to inaccurate results. Therefore, before any firm conclusions can be drawn regarding Hough et al's (1990) study, their ABLE inventory needs to be further validated. In addition, they did not provide a rationale for the linkage between each personality construct and each criterion construct. The criteria used were based on criterion validity coefficients from earlier criterion validity studies. Tett et al. (1991) argued it was not enough to use validity coefficients, rather, the traits needed to have a conceptual link to the criteria proposed.

In a study designed to assess the effect of response distortion in an actual applicant setting, Rosse et al. (1998) argued that studies showing no effect on validity of response distortion did not distinguish between self-deception and impression management. In a similar vein to Paulhus's (1986) argument, Rosse et al. (1998) argued that studies showed self-deception to be a personality variable as opposed to a nuisance variable, whereas impression management reflected a distinct intent to distort responses dependent on the situation (Rosse et al., 1998). Their study aimed to assess the effect of response distortion in personality questionnaires completed by actual applicants and whether this had any effect on hiring decisions. The job applicants completed the NEO-PI-R and the impression management scale of the BIDR during a selection process, while the job incumbents completed the NEO-PI-R. They found that applicant response distortion scores were higher than incumbents, the applicant distribution was negatively skewed, Neuroticism and Conscientiousness were more highly correlated with response distortion, Agreeableness and Extraversion were less so, and Openness to Experience was not at all correlated (Rosse et al.

1998). Furthermore, job applicants showed higher means on all positively-worded facets except one and lower means on all Neuroticism scale facets. This finding indicated that applicants were trying to present a positive image of themselves. They also found a greater chance of applicants with high response-distortion scores selected, supporting the claim that response distortion does affect hiring decisions (Rosse et al., 1998). Rosse et al. (1998) concluded that the NEO-PI-R was perhaps not appropriate for personnel selection as it did not sufficiently control for response distortion bias.

Summary and implications

Faking, then, added to the reluctance of organisations to use personality questionnaires in their selection processes. There are, however, strategies for reducing faking, from scales that measure socially-desirable responding to types of response formats. However, other research has shown these strategies did not enhance the validity of the data, but, in fact, could reduce the validity of the data. Some research suggested faking did not affect the validity of the data at all and, therefore, such validity scales were not really necessary. However, these studies contained methodological flaws that may have affected the interpretation of the results. The type of scale development used may also affect faking. For example, one flaw with rationally-derived scales was that, because item content based on theory was relatively transparent, these scales were more susceptible to faking. The EPQ-R's Lie scale items are transparent, as such, respondents can guess the meaning and answer accordingly. The GPP-I was designed to combat faking by using a forced-choice format. This response format is still, however, susceptible to faking.

Summary

Personality questionnaires have enjoyed increasing usage since the discouraging conclusions of early research. Later research showed that personality questionnaires provided incremental validity over cognitive ability tests. Organisations such as the NZ Army, combine the EPQ-R and GPP-I personality questionnaires with other exercises to form an assessment centre used to select officers for training. However, careful attention needs to be paid when using personality questionnaires in selection as a number of factors affect their

validity. In order to demonstrate construct validity, the personality questionnaire should be based on a sound model that is, in turn, based on well-defined constructs. The FFM is one such model on which many questionnaires have been based, and strong evidence supports the FFM traits' ability to predict job performance and other workplace behaviour. The EPQ-R was based on psychophysiological theories that identified a three-factor model whereas the GPP-I contained an eight-factor structure derived through empirical means. Validity is further enhanced when the instruments designed to measure the constructs are developed through sound scale development processes. Research supports the use of theoretically-derived instruments with items that have been refined through empirical means. The EPQ-R scales were developed in this way, however, the GPP-I scales were developed through empirical means only. Finally, for selection purposes, the personality constructs should be conceptually-linked to the criteria they are trying to predict. Personality questionnaires can be linked to criteria by conducting thorough job analyses that identify personality-relevant criteria based on the people-requirements of the job. In the current research, the EPQ-R and GPP-I scales were linked to workplace behaviour in the NZ Army officer role. Taken together, these factors improve the validity of personality questionnaire data when used in selection settings.

However, some factors decrease the validity of the data. Research has shown that standardised personality questionnaires are less valid in selection contexts. If a personality questionnaire was not specifically designed for selection use, then it needs to be validated for selection before it is used. The EPQ-R was not specifically designed for selection, however, it is used to identify individuals who may not cope with the demands of being an officer. The GPP-I has been well-validated in selection and military settings. Another factor concerns the ease of faking personality questionnaires in selection contexts. Some researchers claimed that faking reduces the validity of personality questionnaire data in selection. Although some personality questionnaires, such as the EPQ-R, contain special scales or special response formats, such as in the GPP-I, to detect faking, these strategies may further reduce the validity of the questionnaire data.

The Present Research

The aim of the current research was to assess the relationship between the EPQ-R and GPP-I, and subsequent behaviour as measured by the WBQ for NZ Army officer applicants. The WBQ was completed by the immediate superiors of participating junior officers and officer cadets. The participants' personality questionnaire data obtained at the OSB were correlated with the corresponding scales of the WBQ to determine whether the EPQ-R and GPP-I measured what they claimed to measure.

Variables

One variable that may affect the data was the corps the officer belonged to. Like Day and Silverman (1989) who argued that personality questionnaires showed increased validity when personality characteristics were matched with the organisation and occupation, Schneider et al. (1998) showed that different organisations could be characterised by shared personality characteristics that served to differentiate them from other types of organisations. Similarly, Hogan et al. (1996) argued that meta-analyses did not separate out studies according to occupational type and this served to reduce the validity coefficients. They claimed that each Big Five dimension predicted performance in these jobs differently, therefore, aggregating the results from jobs with different psychological requirements would reduce the validity (Hogan et al., 1996). Therefore, personality questionnaires may differentiate the corps.

Length of time served was another important variable. Officers, such as Captains, have served more time than Lieutenants who, in turn, have served more time than Second Lieutenants. Because of this, Captains may be more similar in personality to each other than the others because of organisation socialisation processes. Length of time served has also correlated positively with ratings (Borman, 1991) presumably because the rater would have known the subject for longer. In fact, Borman (1982) suggested that assessors' ratings following several sessions with candidates at the assessment centre may be more valid than earlier ratings. Similarly, Binning and Barrett (1989) claimed that personality traits affected job performance more when incumbents had spent more time on the job.

Research has shown that gender differences can occur on personality questionnaires, and gender and ethnicity of the rater and ratee can affect performance ratings. For example, Borman (1991) found that males were rated more highly on traditionally male jobs than females and females were rated more highly on traditionally female jobs than males. In addition, ratees of the same ethnicity as their raters were rated more highly than when the two were of different ethnicity (Borman, 1991). Martin and Kirkcaldy (1998) performed a study where the short form of the EPQ-R and seven scales measuring traits associated with work attitudes was administered to 100 male and female university students. They found that females scored significantly higher on Neuroticism than males, and males scored significantly higher on Psychoticism than females (Martin & Kirkcaldy, 1998). No sex difference was found for scores on Extraversion. Furthermore, they found a significant correlation between Neuroticism and the Lie scale for females, indicating faking was present. It would be interesting to see whether these findings could be replicated in a military sample. In light of this research, data on demographic variables of gender and ethnicity of immediate superiors were gathered, as well as length of time served, age, and rank data. Education level, although not shown to affect ratings, was also included in both the officers/officer cadets and immediate superiors' demographic questions for completeness.

Hypotheses

Scales on the WBQ were developed by a theoretical approach, under the assumption that personality characteristics determine behaviour. Between one and three items were written that reflected behaviour associated with the personality profile of each scale and interrelationships between scales in the EPQ-R and GPP-I. Each WBQ scale, then, was assumed to be conceptually-linked to its corresponding EPQ-R or GPP-I scale and should, therefore, correlate significantly with its corresponding scale.

Hypothesis 1: Scores on the EPQ-R and GPP-I should correlate significantly with corresponding scales on the WBQ

As applicants for the NZ Army know their personality questionnaire results will be used for selection purposes, it is highly likely their responses will be subject to faking.

Therefore, the correlation between EPQ-R Neuroticism and the Lie scale, and between EPQ-R Psychoticism and the Lie scale should be examined, as recommended by Eysenck and Eysenck (1991), to determine whether faking has occurred

Hypothesis 2: The Neuroticism/Lie correlation and the Psychoticsm/Lie correlation will be significant and negative

Research has also shown that employees who have spent a longer time in an organisation may show more similar personality profiles than those who have only been in the organisation a short time.

Hypothesis 3: Officers who have spent three or more years in the NZ Army will show more similar personality profiles than those who have spent less than three years in the NZ Army

Gender differences have been found on the EPQ-R scales, and the gender and ethnicity of the rater and ratee can affect ratings of performance. It is unknown whether these variables could affect ratings of behaviour, therefore, it was decided to assess whether gender and ethnicity affected the behavioural ratings of the current study.

Hypothesis 4: Immediate superiors of the same gender and ethnicity as the officer or officer cadet will give more favourable ratings than those of different gender and ethnicity

Finally, if the EPQ-R and GPP-I are effective selection tools, then there should be no officers and officer cadets with undesirable personality characteristics as measured by the WBQ.

Hypothesis 5: WBQ scores reflecting High Psychoticism, High Neuroticism, Low Emotional Stability, Low Ascendancy and Low Cautiousness should not be endorsed highly.

Design

Sample

The sample was selected from the general population of officers at all camps, and officer cadets in the NZ Army. Participants were not randomly selected, rather, participants needed to meet criteria for inclusion. All officer cadets and junior officers of Second Lieutenant, Lieutenant, and Captain ranks were included in the sample. Senior officers, which are those of Major and above, were not included as potential participants as they were more likely to be immediate superiors of the participants and, if included, may have received the same material twice. In addition, many candidates prior to 1994 were assessed on activities performed at a selection board. The OSB now focuses on the assessment of criteria. Therefore, majors were more than likely selected under the old scheme. Some captains had served in the army prior to 1994 as non-commissioned officers and other ranks, but completed an OSB, and were "commissioned from the ranks" as officers. Therefore, these captains had spent a longer time in the NZ Army than other officers. Officers from overseas armies were excluded as they had not completed an OSB. Finally, the immediate superior had to be someone who had been in charge of the participant for at least four weeks to accurately fill in the questionnaire.

These criteria narrowed the potential sample from 500 to 250 officers and officer cadets. Of these, 59 were officer cadets and 191 were officers. There were 197 males and 53 females. The rank structure was as follows: 59 officer cadets, 45 second lieutenants, 103 lieutenants, and 43 captains. Of this initial sample, 40 officers and 16 officer cadets, giving a total sample of 56, chose to participate. Two questionnaires, one from an officer cadet and one from an officer, were returned without being completed by their immediate superior. This reduced the sample to 15 for the officer cadets and 39 for the officers, yielding a return rate of 21.6%.

One problem encountered in obtaining a representative and adequate sample was the occurrence of APEC, which was held six weeks after the questionnaires were sent out. Of

greater significance was the East Timor crisis that coincided with sending out the questionnaires. Both these events meant that officer workloads were high. The APEC conference required much planning and preparation prior to the actual event. It was believed that the first send-out did not effect a significant return rate due to APEC preparations. Unfortunately, the escalation of the East Timor crisis coincided with the reminder notice sent out to officers and officer cadets asking them to consider participating. This significantly reduced the potential return rate, as even those officers not actually deployed were still heavily involved in preparations for the deployment of others.

Measures

The Revised Eysenck Personality Questionnaire (EPQ-R)

Theory

The EPQ-R consists of five scales measuring Psychoticism, Extraversion, Neuroticism, Addiction, and Criminality, in addition to a Lie scale designed to detect subjects who distort responses to present themselves in a positive light. The scales within the questionnaire were developed through studies using self-ratings, peer-ratings, observational studies, and analyses of psychophysiological and biochemical experiments over a forty-year period (Eysenck & Eysenck, 1991). Essentially, then, the EPQ-R was based on the relationship between personality and underlying physiological processes in the human body, in particular, heredity. Most of the norming research concerning the EPQ was conducted on normal subjects and also psychotics, neurotics, depressive patients, and prisoners (Friedman, 1984).

The EPQ-R originated as the Maudsley Medical Questionnaire in 1952, which measured Neuroticism (Eysenck & Eysenck, 1991). Subsequent revisions saw the Maudsley Personality Inventory and the Eysenck Personality Inventory that measured Extraversion and Neuroticism. The EPI also saw the addition of the Lie scale and two alternate forms to be used on the same population (Eysenck & Eysenck, 1991). This inventory was psychometrically advanced to the MPI, reporting higher reliabilities and validities. Eysenck and Eysenck (1991) reported that Neuroticism and Extraversion could explain much of the

variance seen in the earlier Phlemagtic-Choleric-Sanguine-Melancholic personality dimensions, and later work by Cattell and Guilford showed that their personality dimensions could be explained by a higher-order two-factor structure. However, a third dimension independent of Neuroticism explained the tendency towards psychotic behaviour (Eysenck & Eysenck, 1991). Subsequent factor analyses and experimental studies have supported the existence of this dimension (Eysenck & Eysenck, 1991). Consequently, the Psychoticism scale was added in 1975 and the inventory became the EPQ (Friedman, 1984). The Psychoticism scale came under much criticism for its psychometric properties and these were improved in an EPQ revision to become what is now the EPQ-R. Much of the research is on the EPQ rather than the EPQ-R and changes made in the revised version largely concerned the Psychoticism scale, plus the addition of the Addiction and Criminality scales. The Addiction and Criminality scales consisted of six appended items omitted from the Psychoticism scale of the questionnaire. Two new items have been added to the Extraversion scale, one new item has been added to the Neuroticism scale, and the Lie scale remained as it was. Item scores were summed to give the scale score. A description of these scales, together with reliability and validity evidence is discussed below.

The EPQ-R Scales

Psychoticism

It is important to define Psychoticism properly (Eysenck & Eysenck, 1991). The trait does not refer to what is considered psychopathic behaviour but, rather, the *tendency* towards this behaviour. It is abnormal behaviour associated with a normal personality. Eysenck and Eysenck (1991) defined Psychoticism as "an underlying dispositional personality trait which is present in varying degrees in all persons; if present in marked degree it predisposes a person to the development of psychiatric abnormalities." (p. 1). In other words, a high score on the Psychoticism scale does not mean an individual is psychotic, rather, that the individual is more likely to develop psychoses than someone who scored low on the scale. Eysenck and Eysenck (1991) suggested it may be better to label the Psychoticism scale "Tough-mindedness," a more emotionally neutral term that referred to a set of attitudes opposite to tender-minded attitudes. They found high correlates (0.4) between Psychoticism and Tough-Mindedness as measured on attitude scales (Eysenck & Eysenck, 1991).

As the scale was developed from psychophysiological theories and validated through the empirical-criterion keying method, which showed how it discriminated between two different groups, its actual nature and content was hard to define. Even Eysenck and Eysenck (1991) described the Psychoticism scale as "peculiar" (p. 5), and defined a high scorer as someone who was a loner, did not care about others, was troublesome, cruel, insensitive, and did not fit in with others. In addition, the High Psychoticism profile reflected someone who did not care about danger, liked to upset people and make fools out of them (Eysenck & Eysenck, 1991).

Psychometric properties of the Psychoticism scale

The original Psychoticism scale in the EPQ contained 25 items. In the EPQ-R, the Psychoticism scale was revised by deleting some items and adding new items followed by factor analyses based on the original theories behind Psychoticism (Eysenck & Eysenck, 1991). This was because the Psychoticism scale showed low reliability, low scoring range and a highly skewed distribution (Pearson, 1989). Six items were deleted and later appended for use with the Addiction scale. Thirteen new items were generated through factor analyses, giving a total of 32 items for the Psychoticism scale on the EPQ-R.

The alpha coefficient of the EPQ Psychoticism scale was reported at 0.74 for males and 0.68 for females (Eysenck & Eysenck, 1991). In remedying these reliabilities, Eysenck and Eysenck (1991) performed a study using the new items with 408 male and 494 female students, teachers, and other subjects. The four factors were extracted, and Eysenck and Eysenck (1991) reported higher alpha coefficients of 0.78 for males and 0.76 for females. The new test-retest reliabilities using a one month interval were 0.77 and 0.81 respectively for males and females (Eysenck & Eysenck, 1991). Although these reliabilities may not seem much higher, Eysenck and Eysenck (1991) argued that Psychoticism is a heterogeneous trait that measures a variety of different facets, therefore, its reliability will never be much higher.

The second criticism of the Psychoticism scale referred to the low range of possible scores. Kline (1993) argued that the Psychoticism scale lacked discrimination amongst

normal subjects. The mean for females was 2.63 with a standard error of 2.36, resulting in a large number of zero scores on the Psychoticism scale (Kline, 1993). The mean for males was 3.78 (Corulla, 1987), with a standard error of 3.09 (Eysenck & Eysenck, 1991). This has since been corrected in the EPQ-R. Using the same study that improved reliability, Eysenck and Eysenck (1991) reported a new mean of 7.19 and a standard error of 4.60 for males and a mean of 5.73 with a standard error of 3.85 for females.

Pearson (1989) assessed the differences between the EPQ Psychoticism scale and the EPQ-R Psychoticism scale on occupational therapy students. The students were administered the EPQ during an intial interview and then completed the EPQ-R during their first-year psychology course. Pearson (1989) found that the EPQ Psychoticism scale only correlated 0.06 with the Psychoticism scale of the EPQ-R. He concluded this was not due to distortion, as the Lie scale did not differ greatly in relationship to the Psychoticism scale of either questionnaire (Pearson, 1989). This finding suggested that the Psychoticism scale measured different phenomena in the two versions of the questionnaire. This finding was, however, surprising as Eysenck and Eysenck (1991) reported correlates of 0.88 for males and 0.81 for females between the EPQ Psychoticism and the EPQ-R Psychoticism scales.

Finally, the skewness of the Psychoticism scale was addressed. Eysenck and Eysenck (1991) computed the moments between the original EPQ Psychoticism scale items and the new Psychoticism scale items, and plotted these on histograms. They found that the distribution was improved, although still not normal. However, Eysenck and Eysenck (1991) argued that the Psychoticism scale could never be normal, since by its nature, the scale measures traits not considered to be associated with normal personality.

Extraversion

Extraverts were generally described as outgoing and sociable, and sought out and enjoyed others' company. Eysenck and Eysenck (1991) described an Extrovert as someone who liked parties, had lots of friends, craved company and excitement, was impulsive, and was generally easy-going, optimistic and carefree. The Introvert was the exact opposite, being shy and retiring, cautious, serious, and well-ordered (Eysenck & Eysenck, 1991). Friedman (1984) believed that the EPQ Extraversion scale was more a measure of sociability

than the impulsivity measured by the EPI Extraversion scale. Low Extraversion profiles were more introverted and tended to work more slowly and accurately than High Extraversion scorers (Friedman, 1984). In the revision of the EPQ, two new items were added to the Extraversion scale. The new alpha coefficient for the EPQ-R Extraversion scale was 0.90 for males and 0.85 for females, based on the sample of 408 male and 494 female students, teachers, and others (Eysenck & Eysenck, 1991). Test-retest reliabilities with a one month interval were 0.83 and 0.89 for males and females respectively (Eysenck & Eysenck, 1991).

Neuroticism

Neuroticism was a measure of the emotional stability of an individual. Eysenck and Eysenck (1991) described an individual scoring high on the Neuroticism scale as over-emotional, anxious, and often depressed, who suffered insomnia and psychosomatic disorders, and took a long time to return to normal after a stressful experience. In particular, their reactiveness predisposed them to neurotic disroders when placed in a stressful situation (Friedman, 1984). In the revised version of the EPQ, one extra item was added to the Neuroticism scale. Alpha coefficient reliabilities for the EPQ-R Neuroticism scale were reported at 0.88 and 0.85 respectively for males and females, and test-retest reliabilities with a one month interval were reported at 0.76 for males and 0.81 for females (Eysenck & Eysenck, 1991).

Lie scale

The Lie scale was designed to measure distortion of responses. No changes were made between the EPQ and EPQ-R versions of the Lie scale. The alpha coefficient reliabilities were 0.82 and 0.79 for males and females respectively (Eysenck & Eysenck, 1991). The test-retest reliabilities, using a one month interval, were 0.76 and 0.80 for males and females respectively (Eysenck & Eysenck, 1991). Eysenck and Eysenck (1991) claimed that the Lie scale was designed to detect those who "faked good," such as in employment situations. Elliott, Lawty-Jones, and Jackson (1996) reported evidence that subjects showed elevated Lie scores when instructed to "fake good" on the EPQ-R, and that police service applicants showed elevated Lie scores compared with controls, recruits, and incumbents. In

order to detect faking, the correlates between the Lie scale and Neuroticism, and Psychoticism should be evaluated. If the Lie/Neuroticism correlation was significantly negative and approached or exceeded -0.5, then faking was likely to have occurred (Eysenck & Eysenck, 1991). Similarly, if the Lie/Psychoticism correlation was significant and negative, faking was likely to have occurred. Eysenck & Eysenck (1991) suggested performing regressional analyses where the top 5% of the High Lie scale scores were removed and correlates recalculated until the Lie scale influence could be determined.

Not only did the Lie scale measure distortion of responses; it was also found to measure a stable personality dimension. Eysenck and Eysenck (1991) noted that, while the Lie scale scores could be manipulated in experimental situations, showing it could adequately measure distortion, the reliability of the scale had not been found to differ under high and low faking conditions. In conditions where motivation to distort was not apparent, Eysenck and Eysenck (1991) claimed that the Lie/Neuroticism correlation should be low and, therefore, high Lie scale scores reflected a stable personality dimension. Similarly, Davies, French, and Keogh (1998) argued that, although significant associations may occur between socially desirable responding and test scores, these associations may, in fact, reflect a need for approval, rather than an intention to distort responses. Similar to Paulhus (1986), they argued that, when self-deceptive enhancement was an important part of a measure, controlling for socially desirable responding reduced the predictive validity of the measure. However, if impression management was involved and was independent of the measure, then controlling for socially desirable responding was justified. Elliott et al. (1996) argued that the Lie scale should be scored in two parts; one to detect faking and the other to detect the conformity factor. Eysenck and Eysenck (1991), however, argued that if the motivation to distort was low, reflected by the Lie/Neuroticism correlation, then the Lie scale should not be used as a correction for distortion but as a measure of the conformity/social naivety factor. Corulla (1987) found that the Lie scale loaded highly onto the Extraversion scale, suggesting that high Lie scorers were also sociable and, as Eysenck and Eysenck (1975) suggested, measured a degree of social naivety.

Addiction

The EPQ-R manual (Eysenck & Eysenck, 1991) reported only brief information concerning the Addiction scale. It was developed through item analysis and empirical criterion-keying methods, whereby items that differentiated drug addicts and normals at a significance level of 0.001 or above were included in the scale. In the EPQ-R, the six items from the Psychoticism scale deleted were required for the Addiction scale, hence, they were appended (Eysenck & Eysenck, 1991). The manual provided norms for drug addicts on all scales. Test-retest reliabilities of 0.75 for males and 0.80 for females were reported, based on a sample of 109 males and 120 females using a one month interval (Eysenck & Eysenck, 1991). The alpha coefficient reliabilities were 0.78 and 0.84 respectively for males and females (Eysenck & Eysenck, 1991). Studies reported in the manual showed that bulimics scored highly on the Addiction scale, and that Addiction scores varied with age (Eysenck & Eysenck, 1991).

Criminality

Once again, the manual reported only brief information on the Criminality scale, with no studies reported that empirically tested it. It was developed to better account for the differences observed on the Psychoticism, Neuroticism, and Extraversion scales between criminals and non-criminals, using an empirical criterion-keying approach (Eysenck & Eysenck, 1991). Using a sample of 189 non-criminal and 934 criminal subjects, the reliabilities were 0.75 for non-criminals and 0.75 for criminals (Eysenck & Eysenck, 1991). The manual did not state what type of reliability this was. Test-retest reliabilities appeared to have been performed on a different sample of 109 males and 120 females. These reliabilities were 0.77 for males and 0.76 for females (Eysenck & Eysenck, 1991).

Because the EPQ-R manual was sparse on normative information relating to the Addiction and Criminality scales, the NZ Army have not previously collected information on these scales. However, in order to ensure responsible use of these scales and to supplement the available information, the NZ Army have begun to develop their own norms for these

scales, therefore, have recently commenced collecting data for Addiction and Criminality in officer selection.

Validity

The EPQ-R has demonstrated high reliability. Validity information in the manual, however, was sparse. Intercorrelates among the scales were reported in the manual and, at first glance, suggested evidence of discriminant validity. However, Eysenck and Eysenck (1991) did not state whether these intercorrelates were significant or not. In addition, the manual outlined only a few studies using the EPQ-R, none of which were conducted in a selection setting. Eysenck and Eysenck (1991) described the outcome of these studies, but failed to report statistical evidence for most of the studies. Few studies have been carried out on the Addiction and Criminality scales as these were recent additions. The research cited primarily focused on the Psychoticism revisions, as the other scales of the EPQ did not undergo major changes. Therefore, EPQ research for these other scales should adequately describe the EPQ-R use.

The EPQ was not developed for specific use in selection contexts. While Eysenck and Eysenck (1991) claimed that personality differences were important for selection contexts, no studies or information were cited that provided evidence for the EPQ's use in these contexts. Furthermore, normative studies were not adequate for occupational contexts. The manual provided age norms for males and females, but these norms were based on a sample of 902 students and schoolteachers and "other willing and varied subjects." (p. 15, Eysenck & Eysenck, 1991). Kline (1993) reported norms for the EPQ carried out on 21 accountants, 29 actors, 19 butchers, bakers, and cooks, 9 shop assistants, and 9 physiotherapists. Kline (1993) argued that these norm samples were far too small to be meaningful. Since the NZ Army use their own local norms to combat this problem, information derived from EPQ-R scales in the NZ Army officer selection process is more useful.

Kline (1993) argued that the factors were too broad to be used for selection purposes as they do not possess the amount of discrimination required. Goeters et al. (1993) also argued that the EPI was too simply-structured and could not provide an adequate level of

differentiation for selection purposes. Despite this shortcoming, Kline (1993) noted that factor analysis showed the EPQ to possess excellent validity for all three factors. Kline (1993) claimed that the Extraversion scale was related to low arousability (meaning that High Extraverts will tend to seek stimulation to increase their arousability), Neuroticism was related to lability of the autonomic nervous system, and Psychoticism was related to the level of circulating androgens. These findings suggested that the EPQ (and therefore, the EPQ-R) had a sound basis in psychophysiology as postulated by Eysenck and Eysenck (1991).

The EPQ-R also has a published short form and a form for use in Africa. Wilson and Doolabh (1992) performed a study using the EPI, the EPQ, the EPQ-R, the short-version EPQ-R and the African EPQ (EPQ-A) to compare the validity and reliability of the instruments. In particular, Wilson and Doolabh (1992) argued that additional reliability and validity evidence was not provided in the revisions as evidence of similarity of the scales. Wilson and Doolabh (1992) administered the questionnaires to secondary school and teachers' college students in Zimbabwe. They found that the internal consistency of the EPI Extraversion and Lie scale were poor (0.41 and 0.46 respectively for males and 0.54 and 0.48 respectively for females). The EPQ-R Psychoticism scale only showed a marginal improvement over the EPQ Psychoticism scale and EPQ-A Psychoticism scale (0.67, 0.61, and 0.67 respectively for males, and 0.63, 0.49 and 0.52 respectively for females) (Wilson & Doolabh, 1992). The reliability of the Psychoticism scale was also found to be inadequate in the short form of the EPQ-R (0.50 for males and 0.45 for females) (Wilson & Doolabh, 1992). Wilson and Doolabh (1992) found the factorial validity to be moderate for the Psychoticism scale, but the other dimensions of all the questionnaires showed good validity. The Extraversion, Neuroticism, and Lie scales all correlated well between all the questionnaires, however, the Psychoticism scale showed lower correlates. In addition, the Psychoticism scale showed restriction of range but were less skewed than in western samples. They concluded that all versions essentially measured the same constructs, but that the Psychoticism scale was the least stable for use in Zimbabwe, and needed refining (Wilson & Doolabh, 1992).

Corulla (1987) assessed the relationship between the EPQ-R scales and the I7 Impulsiveness questionnaire, that measured Impulsiveness, Venturesomeness, and Empathy. The EPQ-R and the I7 were administered a week apart to 307 male and female university students. Corulla (1987) found that the Psychoticism scale was less skewed, but closer to

zero for males than for females. Corulla (1987) concluded that the Psychoticism distribution was closer to normal for males than for females. The reliabilities of all the scales for both questionnaires ranged from 0.69 to 0.90. He found significant correlates between the Psychoticism scale and Impulsiveness, and between Neuroticism and Empathy, but although Extraversion and Venturesomeness correlated positively, the correlation was not significant (Corulla, 1987). Corulla (1987) concluded that the EPQ-R Psychoticism scale showed improved internal consistency, improved range of scoring, and a more normal distribution primarily for males.

Faking

Empirical evidence suggested that the EPQ-R was subject to faking. In Elliott et al.'s (1996) study, they administered the EPQ-R, the IVE questionnaire, and a self-rating questionnaire comprised of selected items from the IVE framed positively and negatively to measure impulsive/neutral/non-impulsive behaviour. Two objective measures, the response time for completing both the EPQ-R and the IVE, and the time taken to trace a circle template when told to do so as slowly as possible, were also included. The response time was deemed a measure of impulsivity, where the faster a subject took to complete the questionnaire, the more impulsive they were (Elliott et al., 1996). Elliott et al. (1996) hypothesised that neurotic, extraverted subjects would be more impulsive and more unable to regulate their behaviour. The circle template test was designed to assess whether this group of subjects would complete the task faster than the more "stable" subjects. Subjects were assigned to Stockbroker, Librarian or Control conditions. The subjects in the experimental conditions were asked to imagine they were in a job interview for their respective condition and to complete the EPQ-R and the IVE accordingly. The subjects in the control condition were asked to complete the measures honestly. They were all secretly timed. Subjects then performed the objective tasks under the same set of instructions and were timed. Subjects then completed the self-rating questionnaire designed for the purposes of the study. Neuroticism scores were lower in both the Librarian and Stockbroker conditions compared with the control. The objective tests both measured impulsiveness, demonstrating concurrent validity. Elliott et al. (1996) found that the Lie scale did not effectively discriminate honest subjects from those that were faking, particularly in the Stockbroker condition, where subjects were motivated to fake bad in order to present a tough image. Furthermore, the

claim that the Neuroticism score would decrease in conjunction with an increased Lie score did not occur. The Stockbroker condition showed a positive correlation whereas the Librarian condition showed a negative correlation in the opposite direction (Elliott et al., 1996). Elliott et al. (1996) concluded that the EPQ-R and the IVE were subject to faking but the more objective measures were not, and though face validity was compromised, the objective tests were a better measure of impulsive personality.

Some researchers have claimed that faking the EPQ-R was scale-specific. For example, Davies et al. (1998) reported the EPQ-R Neuroticism scale was negatively correlated with social desirability and the EPQ-R Lie scale was positively correlated with social desirability, whereas the EPQ-R Extraversion scale was not associated at all with social desirability. They administered the EPQ-R and the BIDR to college students and correlated scores. They found that Extraversion correlated positively and significantly with self-deceptive enhancement. Neuroticism correlated negatively and significantly with self-deceptive enhancement, Psychoticism correlated negatively and significantly with impression management. Finally, Lie scale scores correlated significantly and positively with both self-deceptive enhancement and impression management, with the impression management correlation the strongest (Davies et al., 1998). Davies et al. (1998) concluded that socially desirable responding should be controlled for the Psychoticism scale but not for the Extraversion or Neuroticism scale.

Summary

The EPQ-R, then, was developed to assess normal personality, despite the emotion-laden descriptors for Neuroticism and Psychoticism. The EPQ-R was designed to measure the *tendency* towards developing these sorts of behaviour. The manual provided good reliability evidence, however, was sparse on normative and validity information for groups other than students and normal adults. Other studies have shown the EPQ-R scales to demonstrate good validity, however, the Psychoticism scale showed variable validity in different studies. The validity of the Addiction and Criminality scales have yet to be adequately demonstrated.

Although Eysenck and Eysenck (1991) claimed the EPQ-R could be used in a variety of contexts, no empirical evidence, except for Psychoticism scale studies, was provided to support this claim. Furthermore, Eysenck and Eysenck (1991) cautioned against using the

EPQ-R in selection contexts as subjects were likely to show high Lie scale scores. Another problem with using the EPQ-R in selection contexts was the lack of appropriate occupational norms, although the NZ Army have developed their own local norms. However, provided the correlates between the Neuroticism and Psychoticism scales with the Lie scale are examined and found to be acceptable, the EPQ-R should be interpretable in contexts where the motivation to fake is high.

The Gordon Personal Profile-Inventory (GPP-I)

Theory

The GPP-I was comprised of two personality questionnaires, the Gordon Personal Profile (GPP) and the Gordon Personal Inventory (GPI) combined to form the GPP-I for ease of administration. The GPP-I was not based on any particular theory, but rather, the constructs were defined through factor analyses (Guion, 1998). The GPP measured four personality traits: Ascendancy, Responsibility, Emotional Stability and Sociability. The GPI measured four additional traits: Cautiousness, Original Thinking, Personal Relations, and Vigour. The GPP and GPI were originally developed from six factors identified in Cattell's, Guilford's and Thurstone's work, corresponding to Ascendancy, Responsibility, Emotional Stability, and Sociability, and Original Thinking and Personal Relations. Notably, Ascendancy and Sociability have often been combined to form the Extraversion that is part of the Big Five, and factor analyses have confirmed the GPP-I structure in this respect (Gordon, 1993). However, Gordon (1993) argued that the two dimensions are functionally distinct and should not be merged.

The GPP was developed into a forced-choice format where pairs of items were matched for equal social desirability, one negative pair and one positive pair. The resulting tetrad had one statement that measured each factor (Guion, 1998). The GPI traits were also identified through factor analyses of Cattell's work, Original Thinking and Personal Relations had already been identified in the GPP development but could not be included as their item content did not match any other items for socially desirable pairing. Cautiousness and Vigor were identified through factor analyses of Cattell's work. The GPI was also developed into a forced-choice format in the same way as the GPP, with high and low preference items paired

together to form a tetrad. In answering GPP-I items, subjects had to choose the phrase within each tetrad which was most like them and which was least like them. A global edition was developed in 1993 where item wording was changed slightly so that the questionnaire could be understood in different cultures (Gordon, 1993). To avoid a purely ipsative arrangement where high scores on two scales necessarily meant low scores on the other two scales, two of the statements in each tetrad were favourably worded and two were unfavourably worded, so that a range of possible points within each set could occur (Gordon, 1993). Item scores were summed to give the total scale score. A further refinement was to the composite total score of the four GPP dimensions that served as a measure of self-esteem. In the GPP-I the composite was officially labelled Self-Esteem and was a scale in its own right. The number of desirable or undesirable items endorsed across all statement sets indicated a measure of self-worth, and showed a perfect correlation with the total score of the four dimensions (Gordon, 1993). This total score, then, was the score for Self-Esteem. The resulting GPP-I was arguably broad enough for a wide variety of uses, and has been used extensively in organisation contexts as most of its items show work-related content. Gordon (1993) claimed that it was not easily faked, had good test-taker support, and items were relevant to job-success, thereby showing its use for selection contexts.

The GPP-I manual (Gordon, 1993) provided reliabilities for high school and college students and for male managers. The reliability of the global edition, which was the edition used by the NZ Army, was based on a sample of college students. Therefore, reliabilities presented for the following scales were taken from the GPP-I manual (Gordon, 1993) and reported for both college students, based on a sample of 134 students; and male managers, based on a sample of 218 as the male manager role more closely resembled an officer role. Test-retest reliabilities were reported for an unknown sample size of US Naval recruits using a 27 week interval. Separate reliabilities for males and females were not provided. Norms were provided for college students, adults, and male and female managers.

The GPP-I Scales

Ascendancy

The GPP-I manual (Gordon, 1993) described individuals who scored high on the Ascendancy scale as self-assured, who took an active role in groups, and were independent. Those who scored low were more passive and preferred to support others than take the lead. Alpha reliability coefficients for the Ascendancy scale were 0.85 (college students) and 0.82 (male managers). The test-retest reliability was 0.70.

Responsibility

A high scorer on the Responsibility scale was characterised by perseverance and reliability (Gordon, 1993). Low scorers were deemed irresponsible and "flighty" (p. 2, Gordon, 1993). Alpha coefficients for this scale were 0.87 and 0.84 for college students and male managers respectively. The test-retest reliability was 0.66.

Emotional Stability

A stable individual who did not suffer much from anxiety, worry or nervous tension would score highly on this scale. Low scorers did not tolerate frustration well, and were anxious, nervous, and hypersensitive (Gordon, 1993). Alpha coefficients were 0.88 for college students and 0.82 for male managers. The test-retest reliability was reported at 0.50.

Sociability

High scores on this scale reflected an individual who enjoyed others' company, was sociable and gregarious. Low scores reflected an individual who was the opposite of these traits and who may even avoid social contact. Alpha coefficients were reported at 0.86 for college students and 0.85 for male managers. The test-retest reliability was 0.65.

Self-Esteem

Norms for the Self-Esteem score were based on a sample of college students (n = 1709) and reliability information stated that split-half, alpha coefficients, and test-retest reliabilities clustered around 0.94 value (Gordon, 1993). A score was not developed from the Inventory section as these scales were not conceptually linked to self-esteem, and the items were not as clearly favourable and unfavourable (Gordon, 1993). As the distribution was often negatively-skewed, scores were more interpretable for low scorers on the scale (Gordon, 1993). Those who obtain low scores endorsed more unfavourable items compared with favourable items, and Gordon (1993) argued this served as an operational definition of low self-esteem. Low scorers were characterised as anxious, unreliable, withdrawn, and not self-confident (Gordon, 1993). Gordon (1993) did note, however, that individuals with a very high Self-Esteem score could be trying to present themselves favourably. In essence, then, this finding could be similar to results gleaned from the EPQ-R Lie scale.

Cautiousness

High scores on the Cautiousness scale reflected an individual who did not like to take risks, planned carefully, and were cautious. Those who obtained low scores were impulsive, made hasty decisions, took risks and sought excitement. The alpha coefficients for this scale were 0.86 for college students, and 0.83 for male managers. The test-retest reliability was 0.70.

Original Thinking

Individuals who were intellectual, enjoyed creative problem-solving, and discussions scored highly on this scale. In contrast, low scorers were those who did not take a great interest in discussions or acquiring knowledge. Alpha coefficients for this scale were 0.87 for college students and 0.81 for male managers. The test-reliability for this scale was 0.79.

Personal Relations

High scores on this scale were characterised by individuals who were trusting, patient, tolerant, and empathic. Low scorers were not trusting, were easily annoyed and irritated, and may criticise. The alpha coefficients for this scale were 0.83 for both college students and male managers. The test-retest reliability was 0.67.

Vigor

Vigorous individuals were energetic and enthusiastic, and worked quickly. Low scores on the Vigor scales represented lethargic, slow-paced, and easily-tired individuals. The alpha coefficients for this scale were 0.85 for the college students and 0.82 for the male managers. The test-retest reliability was 0.65.

Validity

Guion (1998) reported that the reliability and validity of the GPP-I was adequate, with alpha coefficients at least 0.79 and stood the test of time through test-retest reliabilities. However, Guion (1998) criticised the GPP-I for not meeting current standards, rather, little has changed since its development forty years ago. He claimed that little technical data was reported in the manual, and the latest research performed on faking the GPP-I was in 1973. Furthermore, Guion (1998) argued that factor analysis evidence provided in the manual was superficial and did not allow for criterion unreliability or range restriction. Guion (1998) also criticised the norms provided, which were mainly for high school students. Some norm samples contained managers, however, Gordon (1993) strongly advised development of local norms. Guion (1998) concluded the GPP-I does still seem to be relevant for personality measurement today, that no evidence exists that the FFM is actually any better than the eight-factor structure behind the GPP-I.

Although the GPP-I was designed with selection and organisational settings in mind, Hess (1998) noted the GPP-I did not correlate well with Holland's Occupational Themes, but

did not explain why. Correlates reported in the GPP-I manual with Holland's Occupational Themes were also low (Gordon, 1993). Furthermore, Hess (1998) noted that, while there was good evidence for the GPP-I's correlates with assessment centres, Peace Corps deselection, and managerial advancement, there was little evidence to suggest that the GPP-I was more effective than other personality questionnaires. In contrast, Hogan (1991) claimed that the GPP-I was an effective predictor of workplace behaviour. He reported significant correlates of 0.28, 0.27, and 0.31 between Ascendancy and peer ratings of management potential in three groups of manufacturing managers (Hogan, 1991). Hogan (1991) also reported significant correlates between Ascendancy and a progress measure in management trainees and sales trainees.

Some validity evidence for the GPP-I was provided in the manual, where further factor analyses confirmed the eight factor structure behind the GPP-I scales (Gordon, 1993). One study reported in the manual compared the correlation of scores on the GPP-I and the EPI, particularly the EPI Neuroticism scale and the GPP-I Emotional Stability Scale, and EPI Extraversion scale and the GPP-I Sociability scale. The EPI and the GPP-I were administered to undergraduates and correlates calculated. Extraversion was found to significantly correlate with Sociability (0.59), and Neuroticism correlated significantly with Emotional Stability (-0.55) (Gordon, 1993).

Another study by Gillis and Lee (1979) compared scores on the 16PF, GPP, and GPI. Subjects were high school students who were administered the questionnaires on three occasions separated by one-week intervals. Multiple regression analyses showed that half of all the correlates were significant, and the 16PF was a better predictor of GPI and GPP scores. They concluded that the questionnaires essentially measured the same personality dimensions but that the 16PF factor structure, due to its greater comprehensiveness, was a better predictor (Gillis & Lee, 1979).

Schippmann and Prien (1989) sought to distinguish general mental ability and personality characteristics of successful managers from unsuccessful managers. Subjects were managers participating in an assessment programme and were classified into top management, middle management, low management, and non-management. Using a criterion of management success developed specifically for the purposes of the research, Schippman and Prien (1989) correlated scores on the Edwards Personal Preference Schedule

(EPPS), Self-Descriptive Inventory (SDI) and the GPP-I with this criterion. They found that Cautiousness and Responsibility correlated significantly and negatively with the success criterion (-0.21 and -0.20 respectively), whereas Original Thinking (misnamed Order in the study), Vigor, and Ascendancy correlated significantly and positively with the success criterion (0.27, 0.29, and 0.28 respectively). The authors concluded that personality characteristics could be useful in describing managerial behaviour, however, despite the purposes of their study, no results were reported as to whether the personality tests distinguished between the different management levels.

Faking

Gordon (1993) argued that the GPP-I was not as susceptible to faking as questionnaires that employed a yes-no format (such as the EPQ-R). Studies reported in the GPP-I manual (Gordon, 1993) showed only the Responsibility scale had a significantly higher mean difference in a simulated employment condition than in a simulated vocational guidance condition. In a different study, the Original Thinking scale showed a higher mean for the employment condition. Another study showed higher means for Responsibility and Emotional Stability in an actual selection condition for the GPP and a higher mean for Original Thinking of the GPI in the same condition. Gordon (1993) concluded that, while in some cases responses on some scales could be distorted, the magnitude of the distortion was small. Hess (1998), however, argued that the method used to equate each pair within a tetrad was not specified in the manual, therefore, it was not possible to determine the strength of the item development process.

Braun (1963) assessed the fakability of the GPI using undergraduate students. The subjects were divided into three groups and were administered the GPI twice. In the first administration, the subjects completed the GPI honestly. In the second administration, Group 1 was instructed to fill in the GPI as if they were the worst candidates for a top-level job (fake-bad), Group 2 were instructed to fill in the GPI as if they were the best candidates for the job (fake-good), and Group 3 were instructed to fill in the GPI as if they were the best candidates for the job but to try and disguise the fact they were faking the questionnaire (Braun, 1963). Braun (1963) found that the means were significantly higher in the fake-good sets for all scales except the Personal Relations scale. The means were significantly lower in

the fake-bad set for all scales. Braun (1963) also found that faking could still be detected in Group 3 subjects. He concluded that the GPI could be faked when subjects were motivated to do so.

Schwab (1971) performed two studies addressing response distortion using the GPI. Subjects were divided into two groups, the first were instructed to answer the GPI honestly and the second to answer as if they were candidates for a top management position. Three weeks later, subjects were again divided into two groups, one comprised of those who had faked and half of those who had answered honestly, and the other group comprised of the remaining subjects. The first group was dismissed, and the latter group was further subdivided. Half were asked to respond to the GPI honestly and the other half was asked to fake the inventory. The result was four groups of subjects counterbalanced for honest versus fake responding. In Study 2, the same procedure was repeated on another sample. Schwab (1971) found a significant distortion effect for the Original Thinking scale in both studies where subjects faked. Furthermore, when control groups were not analysed, faking scores increased. When the control groups were added to the analysis, six of the eight tests showed a decrease in the honest-fake difference (Schwab, 1971). Therefore, when compared with control groups, the GPP-I was not as susceptible to faking.

Summary

The GPP-I is an example of an empirically-derived questionnaire that was designed to be used in industrial settings. It consists of two questionnaires joined together and has shown good reliability and validity in a variety of organisational and military settings. Despite claims in the manual that the GPP-I is relatively resistant to faking, earlier research has shown that the GPP-I can actually be faked in selection settings, like any other personality questionnaire.

Development of the Workplace Behaviour Questionnaire (WBQ)

The present study involved developing a workplace behaviour questionnaire to assess personality-relevant behaviours on the job. This questionnaire formed part of a larger questionnaire designed by the NZ Army that assessed behaviour related to OSB criteria and cognitive ability tests. In keeping with research on developing personality-relevant criteria, questionnaire items were developed to be consistent with the scale interpretations of the EPQ-R and GPP-I. It was decided to assess the construct validity of the EPQ-R and GPP-I by writing behavioural items that reflected each personality scale as the NZ Army's job analysis criteria were based on British and Australian Army job analyses; a New Zealand Army job analysis is currently under development. The questionnaire was developed in conjunction with the Army Psychologists, in particular, the Senior Psychologist and the Assistant Director of Psychological Research (ADPR).

Initially, the Senior Psychologist (Army) provided detailed information on the scale interpretations of the EPQ-R and the GPP-I. From these interpretations, a list of behaviourally-anchored items was generated that described workplace behaviour relevant to the personality scale profiles. Workplace behaviour in the military context is different to workplace behaviour encountered in civilian organisations. In the military, personnel do not work a typical 9am to 5pm job. Rather, they are "on duty" 7 days a week, 24 hours a day. As such, workplace behaviour also includes behaviour in the field and on operations, behaviour at social and organised regimental functions, behaviour in the officers' mess, and behaviour at the OCS mess in addition to behaviour displayed during the day at their place of work. Therefore, items were not written specifically to contain work-related behaviours because officers and officer cadets were considered to be "at work" all the time.

Between one and three items were written for each of the EPQ-R and GPP-I scales. Items were positively and negatively bipolar, and measured with a 5-point Likert scale. The EPQ-R Addiction and Criminality scales had little data collected in the NZ Army. It was decided to include an item measuring Addiction but not Criminality as officers are generally not selected if they have a criminal record, therefore, this item would show a low or zero endorsement rate.

Some scales, when interpreted together, produced a different profile to the profiles of the scales separated (Mirfin, personal communication, May 1999). For example, High EPQ-R Psychoticism and Low GPP-I Cautiousness interacted to produce a profile of a person who was impulsive, would act without thinking, and was generally considered a danger. Therefore, nine additional items were written to measure the nine additional profiles found from interactions of the scales. The items measuring the scale interactions were unipolar and were measured by a 5-point Likert scale with the range of possible responses from 1 Strongly Disagree to 5 Strongly Agree. The questionnaire also contained items measuring personality profiles of a personality questionnaire currently being trialled by the NZ Army, the Trait Self-Descriptive inventory (T-SD). It was intended to evaluate this questionnaire in the same way as the EPQ-R and the GPP-I, however, only eleven participants had completed the T-SD, therefore, it was excluded from any further analysis.

A meeting was held with all the army psychologists to refine the items ensuring they measured one behaviour per item that accurately reflected the personality trait it was designed to measure. Another meeting was held with the Assistant Director of Psychological Research (ADPR) and the Senior Psychologist (Army) to further refine the items and to discuss the format. It was decided to randomly order the items so that some began with the less positively-worded statements first while other items began with the positively-worded statements first. This decision was to ensure that the immediate superiors did not fall into a response pattern. The researcher assigned a "1" to denote "negative-wording first" and a "2" to denote "positive-wording first". Then the researcher asked the Army Psychology Service Research Officer to call out either 1 or 2 43 times. The researcher wrote these numbers down and ordered the 43 bipolar items accordingly. The research officer had called out "1" 21 times and "2" 22 times.

The final questionnaire was given to three officers to read to ensure that items and instructions were easily understood and readable. They were also asked to comment on item content.

Method

Each officer and officer cadet completed the EPQ-R and GPP-I during the OSB they sat at some stage since 1994. Candidates were aware they would be selected, in part, on the basis of these personality questionnaire scores, therefore, their responses were subject to faking. The present study was advertised in the internal Army News newspaper and a letter was sent out internally to all Army units from Army General Staff advertising the study and encouraging participation. Each officer and officer cadet in the sample was sent a package through the NZ Army internal mail system. This package contained an information sheet outlining the purposes of the research and participants' rights under the Privacy Act (1993), a consent form, and the WBQ developed specifically for the purposes of this research. A copy can be found at Appendix A, with a description of the items measuring each scale at Appendix B. A copy of the information sheets and consent form can be found at Appendices C and D. This send-out took place six weeks prior to the APEC conference described earlier. A reminder notice was sent to potential participants three weeks after the initial send-out to increase the return rate. It was reasonable to assume that the East Timor crisis affected the return rate, as only two more questionnaires were received.

The officers and officer cadets were asked to read the information sheet and, if they chose to participate, sign the consent form that asked permission to use their personality questionnaire data completed at the time of selection. They then filled in the demographic section of the questionnaire and gave the questionnaire to their immediate superior to complete. Once the immediate superior completed the questionnaire, they returned it to the Army Psychology Service where staff removed the consent form from the front page and assigned a code number to the questionnaire. This was to protect the confidentiality of the participants. Once the codes had been assigned, the questionnaires were returned to the researcher, together with a database containing the participants' EPQ-R and GPP-I scale scores. The data obtained from the WBQ was correlated with these scores to assess whether these questionnaires were measuring what they were supposed to measure.

Because of the nature of the sample, caution needs to be exercised in generalising the results of the study. First, the low return rate of 21.6% meant that the sample was not representative of the officer population from which it was drawn. Second, the criteria for

inclusion in the study resulted in a select sample which, again, may not be representative of the officer population. Finally, due to the small sample size, some variables could not be analysed. It was not possible to collapse data across corps or rank, and there were too few participants from ethnic groups other than NZ Pakeha to make meaningful analyses possible. Therefore, this study could be viewed as a pilot study from which further replication with a larger sample may yield more generalisable results.

Limitations

The current study was limited by the following factors. As mentioned before, the sample size was small, which would lead to a reduced effect size. For example, Tett et al (1991) noted that nonsignificant and negative validities could occur in studies with small sample sizes. As an example, Tett et al. (1991) reported the average sample size in an earlier meta-analysis was 102, which they believed to be too small when considering the effect of sampling error on effect size. Schmidt and Hunter (1978) argued that validity generalisation may be possible if sample sizes are large. They reported well-designed research using large samples of army subjects that showed selection tests had similar validities and regression weights over a variety of jobs. Therefore, the interpretations of the data in the present study should be made with caution as any information gleaned from the data may be an overestimate.

Although local norms are collected by the NZ Army, these were not available at the time of the research and, therefore, could not be used. Comparisons were made with data from the personality questionnaire manuals instead. Furthermore, some reliability estimates for the EPQ-R and GPP-I scales could not be calculated directly from the data as individual item scores were not entered on the database used for the current research. These item scores, although recorded elsewhere, were not available for the current research. The lack of item scores and consequent lack of reliability estimates means that validity information gained from the correlates obtained may not be meaningful.

Results

Demographic Characteristics

Tables 1 and 2 show the demographic characteristics of the officer cadet sample.

Table 1

Gender, Ethnicity, and Highest Education Qualification of the officer cadet sample

Characteristic	Number	Percentage
Gender		
Male		
Female	12	80.0
Ethnicity		
NZ Maori	1	6.7
NZ Pakeha	14	93.3
Asian	0	
Pacific Island	0	
Other	0	
Education		
None	0	
School Certificate	0	
Sixth Form Certificate	0	
University Entrance	1	6.7
Bursary/Scholarship	11	73.3
Trade/Prof Cert/Dip	0	
Part-Degree/Diploma	3	20.0
Bachelor Degree	0	
Postgraduate	0	

Table 2

Age at Time of WBQ Administration , Length of Time in RF and in TF in the Officer Cadet Sample

Characteristic	Years
Age	
Minimum	18.0
Maximum	21.0
Mean	19.1
Standard Deviation	1.0
Length of Time (RF)	
Minimum	1.0
Maximum	3.0
Mean	1.1
Standard Deviation	0.6
Length of Time (TF)	
Minimum	0
Maximum	2.0
Mean	0.3
Standard Deviation	0.7

Table 1 shows that the officer cadet sample was primarily male, NZ Pakeha, and held at least a university entrance or equivalent qualification. Table 2 shows that the officer cadet mean age was 19.1 years and all had spent less than three years in the NZ Army, as they were still in training. Few officer cadets had spent time in the Territorial Force (TF) (part-time soldiers); this was not surprising as a minimum age of 18 is required for the TF. Most officer cadets would have opted for the RF at the age of 18 as opposed to TF.

Tables 3 and 4 show the demographic characteristics of the officer sample. In Table 3, the minimum age of the sample was 21 years, with the maximum age 41 years. Since personality questionnaire data was obtained from applicants from 1994 onwards, the older officers were, more than likely, those officers who had been "commissioned from the ranks." With a standard deviation of 5.02 years, the ages were well spread-out over these years. Only a small amount of time had been spent in the TF and, again, this was not considered unusual as currently serving RF officers were targeted for the study. Table 4 shows that most of the

sample were lieutenants, and the largest number came from the Logistics regiment. Only ten of the sixteen corps were represented in this sample. No participants came from the SAS, Intelligence, Legal, Military Police, Physical Training, or Chaplains corps. A larger sample size would be needed to ensure that the corps were represented accurately. Most of the officer sample were NZ Pakeha, and had a minimum education level of school certificate on joining. The minimum education qualification was lower in the officer sample than in the officer cadet sample. This was because some officers were "commissioned from the ranks."

Table 3

Age at Time of WBQ Administration, Length of Time in RF, and in TF of the Officer Sample

Characteristic	Years
Age	
Minimum	21.0
Maximum	41.0
Mean	27.1
Standard Deviation	5.0
Length of Time (RF)	
Minimum	1.0
Maximum	25
Mean	6.7
Standard Deviation	5.8
Length of Time (TF)	
Minimum	0
Maximum	5.0
Mean	1.0
Standard Deviation	1.6

Table 4

Gender, Ethnicity, Highest Education Qualification, Rank, and Corps of the Officer Sample

Characteristic	Number	Percentage
Gender		
Male	29	74.4
Female	10	25.6
Ethnicity		
NZ Maori	2	5.1
NZ Pakeha	35	89.7
Pacific Island	1	2.6
Asian	0	
Other (Australian)	1	2.6
Education		
None	0	
School Certificate	5	12.8
Sixth Form Certificate	2	5.1
University Entrance	7	17.9
Bursary/Scholarship	10	25.6
Trade/Prof Cert/Dip	1	2.6
Part-Degree/Diploma	6	15.4
Bachelor Degree	7	17.9
Postgraduate	0	
Missing	1	2.6
Rank		
Second Lieutenant	7	17.9
Lieutenant	24	61.5
Captain	8	20.5
Corps		
Artillery	6	15.4
Armoured	2	5.1
Engineers	3	7.7
Signals	6	15.4
Infantry	7	17.9
Medical	2	5.1
Dental	1	2.6
Education	1	2.6
Nursing	1	2.6
Logistics	10	25.6

Tables 5 and 6 show the demographic characteristics of the immediate superiors.

Table 5

Age at Time of WBQ Administration, Length of Time in RF, and in TF of the Immediate Superior Sample

Characteristic	Years
Age	
Minimum	20.0
Maximum	62.0
Mean	32.5
Standard Deviation	10.5
Length of Time (RF)	
Minimum	2.0
Maximum	27.0
Mean	11.0
Standard Deviation	6.7
Length of Time (TF)	
Minimum	0
Maximum	20.0
Mean	3.8
Standard Deviation	5.1

Table 6

Gender, Ethnicity and Rank of the Immediate Superior Sample

Characteristic	Number	Percentage
Gender		
Male	49	90.7
Female	4	7.4
Missing	1	1.9
Rank		
Officer Cadet	13	24.1
Captain	7	13.0
Major	23	42.6
Lieutenant Colonel	4	7.4
Missing	7	13.0
Ethnicity		
NZ Maori	4	7.4
NZ Pakeha	46	85.2
Pacific Island	2	3.7
Asian	0	
Other	0	
Missing	2	3.7

Table 5 shows that the minimum age in the sample was 20 years. This age was low for an immediate superior. Within the officer cadets, there was an internal rank structure. This meant that those officer cadets who had served longer could have a higher rank to other officer cadets who had just started at OCS. It seemed that the officer cadet sample took the instruction to give the questionnaire to their immediate superior too literally and passed it to a senior officer cadet, instead of to an OCS instructor. Because the senior officer cadets were not yet officers, strictly speaking, they could not be classified as immediate superiors. Therefore, the age range and length of time served in the RF and TF was lower than it might normally have been. Furthermore, the senior officer cadets may have had more chances of observing the participants' behaviour than immediate superiors of officers as the officer cadets work closely together. As such, this group was analysed separately from the officer group. Table 6 shows that most immediate superiors were male, NZ Pakeha with the rank of Major.

Initial Analyses

Data obtained from the WBQ were entered onto a database using SPSS version 9.0. If more than one answer had been given for a particular item, the first of the two answers was entered. Item scores were then summed to give a scale score. It was hoped to carry out a factor analysis on the WBQ scales to see whether the items did load onto the factors that they were designed to measure. However, because of the small sample size, this factor analysis was not possible. Instead, the data was subjected to correlational analyses and nonparametric tests. The present results, therefore, should be interpreted with caution. Prior factor analyses of personality questionnaires including the EPQ-R and GPP-I were carried out in 1989 by the Army Psychology Service and revealed four factors could be extracted. Three of these related to the EPQ-R and GPP-I scales. A varimax rotation revealed that Factor 1 was comprised of GPP-I Responsibility (0.74), Emotional Stability (0.67), Vigor (0.66), Original Thinking (0.60), Cautiousness (0.59), Personal Relations (0.57) and EPQ Psychoticism (-0.40). This factor was labelled Adjustment/Stability. Factor 2 was comprised of GPP-I Original Thinking (0.37), Sociability (0.80), Ascendancy (0.78), and EPQ Extraversion (0.79). This factor was labelled Gregariousness. The final factor was comprised of EPQ Neuroticism (0.76) and EPQ Lie (-0.61) and reflected a level of adjustment factor.

Correction for Attenuation

The correlates were not corrected for attenuation or range restriction as these corrections would not add any meaningful information. Using EPQ-R Extraversion as an example, the uncorrected correlation between this scale and the WBQ Extraversion scale in the entire sample, $n = 54$, was 0.13 with a 95% confidence interval of $-0.13 \leq r \leq 0.39$. Correcting for attenuation using a computer program based on Gulliken's (1987) correction formulae yielded a correlation of 0.25 with a 95% confidence interval of $-0.25 \leq r \leq 0.74$. While the correlation had increased, so had the confidence interval. In essence, the meaning of the confidence interval was that one could be 95% certain that the corrected correlation lay between -0.25 and 0.74. This is meaningless. Hunter, Schmidt, and Jackson (1982) argued that it was important to correct for attenuation as the correction reduced the measurement error inherent in the observed correlation. However, measurement error was also affected by

unsystematic error or sampling error. When the observed correlation was corrected for attenuation, the sampling error was also increased. This resulted in the widening of confidence intervals to account for the increased sampling error (Hunter et al., 1982). The larger the correlation in the first place, the smaller the sampling error (ie the lower the standard error). If the correlation was large to begin with, then the corrected correlation would result in a larger correlation with smaller confidence intervals. Hunter et al. (1982) argued that this "substantive measurement" (p. 59) was superior to statistical corrections, and only occurred when better measurement procedures were used in the first place.

Range Restriction

A similar story is presented for restriction of range. In essence, range restriction referred to the phenomenon that occurs when those with very high or very low scores were not selected or selected themselves out (Hunter et al., 1982). Using the same example, the corrected correlation between the EPQ-R Extraversion and the WBQ Extraversion was 0.25 with a 95% confidence interval of $-0.25 \leq r \leq 0.74$. In selection studies where the sample used were usually job incumbents, the sample was comprised of the top scorers on the cognitive ability tests. In the current study, applicants with undesirable personality profiles would, more than likely, have been selected out. Therefore, the range of scores in the sample is restricted, and the estimated population correlation was reduced (Hunter, et al., 1982). The aim of correction formulae was to estimate the effect of changing the population standard deviation to a different value (Hunter et al., 1982). Using the computer program based on Gulliksen's (1987) correction formulae, the correlation corrected for attenuation was further corrected for range restriction. The resulting correlation was 0.46 with a 95% confidence interval of $-0.46 \leq r \leq 0.91$. Again, this result is essentially meaningless. The corrected correlation contains different sampling error to the sample error, and this is accounted for by the increased confidence interval. Thus, a large amount of sampling error is associated with these correlates and the best way to increase these correlates is to use better measurement procedures.

The results of the study are presented below, based first on the whole sample and then followed by the officer cadet sample and then the officer sample.

Hypothesis 1: Scores on the EPQ-R and GPP-I scales should correlate highly with corresponding scores on the WBQ.

Combined Sample

Descriptive Statistics

Table 7 presents the descriptive statistics of the EPQ-R and the GPP-I for the whole sample, n = 54.

Table 7

Means, Standard Deviations and Range of Scores on the EPQ-R and GPP-I

Variable	Mean	Standard Deviation	Minimum	Maximum	Range
EPQ-R P	2.63	1.94	0	9	9
EPQ-R E	17.80	2.75	10	22	12
EPQ-R N	5.22	4.29	0	18	18
EPQ-R L	8.35	3.67	2	16	14
EPQ-R A	4.67	2.54	0	9	9
GPP-I A	25.39	3.61	13	32	19
GPP-I R	28.81	3.22	19	35	16
GPP-I E	26.69	3.49	20	33	13
GPP-I S	23.94	3.58	12	31	19
GPP-I SE	104.74	4.78	84	109	25
GPP-I C	23.81	4.70	12	35	13
GPP-I O	31.24	4.40	22	39	17
GPP-I P	26.69	4.02	15	33	18
GPP-I V	29.26	4.23	21	38	17

Ideally, it would have been best to compare the above data with a military sample, especially the NZ Army population. However, descriptive statistics for the NZ Army officer population were not available at the time of the research. Therefore, the results in the study were compared with statistics provided in the EPQ-R manual (Eysenck & Eysenck, 1991). In addition, tests of statistical significance of the differences could not be carried out, as the individual data was, obviously, not recorded in the manual. The sample used in the manual was comprised of a variety of subjects, no data for military samples were reported. As this

was not an ideal comparison group, any differences reported in the current study should be interpreted with caution.

EPQ-R

The EPQ-R manual (Eysenck & Eysenck, 1991) gave descriptive statistics according to age groups. For the combined sample, then, the average of the means and of the standard deviations of the 16-20yr, 21-30yr, and 31-40yr age groups were used as a comparison. The 16-20yr age group was deemed to represent the officer cadets, as their ages fell into this category. Most of the officers fell into the categories 21-30yrs and 31-40yrs.

Table 8 shows the averaged means and standard deviations from the EPQ-R manual based on a sample of 902 subjects (Eysenck & Eysenck, 1991).

Table 8

Averaged Means and Standard Deviations for the EPQ-R Scales

Scale	Mean	Standard Deviation
Psychoticism	8.30	4.47
Extraversion	14.13	5.52
Neuroticism	11.37	5.58
Lie	5.85	3.72
Addiction (Addicts)	20.04	5.35
Addiction (Normals)	12.11	4.57

For the Psychoticism scale, the average mean for the three age groups was 8.30 and the average standard deviation was 4.47. These figures were much higher than those found in the current study. In particular, the difference between the two standard deviations indicated the presence of range restriction on the Psychoticism scale. The Extraversion and Lie scales also showed possible range restriction effects as their standard deviations were different. The mean for the Neuroticism scale in the current study was lower than in the general population. This finding supported Barrick and Mount's (1996) study and Black's (1997, 1998) research. It appears that the sample scored higher on the Lie scale than the general population. This may be indicative of faking.

Finally, the Addiction scale mean and standard deviation were based on a separate study of 155 male and female addicts and 155 normal male and female addicts. Ages were not given for this sample, rather, statistics were given according to gender. The mean and standard deviation for the current study are much lower, at 4.67 and 2.54 respectively. This could mean that officers are less likely to be addicts than the general population, which is likely, as people with a history of addiction are generally not selected. However, the current study was based on a sample size of only 21 for the Addiction scale, as the NZ Army have only recently begun collecting data on this scale.

GPP-I

The GPP-I manual (Gordon, 1993) presented means and standard deviations for college students and a group of male managers, amongst other groups. As officer cadets were of a similar age to the college students, the data in the current study were compared with the college students. Officers undertake a role similar to a manager (Gordon, 1993), therefore, their data was compared with the male manager group. The college students were administered the global edition of the GPP-I, but the male managers were administered the standard version. Corrected correlates between the standard GPP-I scales and the global edition scales ranged from 0.97 to 0.99 (Gordon, 1993). Therefore, it was reasonable to assume that both versions were the same for the purposes of this study and the means and standard deviations could be combined for comparison in the combined sample. Table 9 shows these averaged means and standard deviations.

Table 9

Averaged means and standard deviations for the GPP-I Scales

Scale	Mean	Standard Deviation
Ascendancy	23.45	5.6
Responsibility	26.85	5.1
Emotional Stability	25.2	5.75
Sociability	22.1	5.6
Self-Esteem	N/A	N/A
Cautiousness	25.4	5.55
Original Thinking	26.85	5.85
Personal Relations	26.6	5.35
Vigor	26.4	5.85

The GPP-I manual (Gordon, 1993) did not provide means and standard deviations for the Self-Esteem scale. Hence this scale could not be compared with the current data. The Ascendancy, Responsibility, Emotional Stability, and Sociability scales were at least two standard deviations higher than those reported in the current study, which could indicate range restriction effects. Participants in the current study had a slightly higher average on the Ascendancy, Responsibility, Original Thinking, and Vigor scales. In particular, the mean Original Thinking score was 31.24 compared with 26.85. This scale had the highest difference.

WBQ

For completeness, Table 10 shows the descriptive statistics of the WBQ scores for the combined sample. Data were missing for two participants for the Emotional Stability scale, therefore, the descriptive statistics for this scale were based on a sample of 52.

Table 10

Means, Standard Deviations, and Range of the WBQ scales

Variable	Mean	Standard Deviation	Minimum	Maximum	Range
Psychoticism	6.30	2.24	3	12	9
Extraversion	6.94	1.45	3	9	6
Neuroticism	5.00	1.58	2	8	6
Lie	4.20	1.55	2	8	6
Addiction	1.54	0.93	1	4	3
Ascendancy	7.54	1.34	4	10	6
Responsibility	4.00	0.87	2	5	3
Emotional Stability	6.81	1.12	5	9	4
Sociability	10.91	1.12	8	13	5
Self-Esteem	3.83	0.93	1	5	4
Cautiousness	3.00	0.73	1	5	4
Original Thinking	7.33	1.43	3	10	7
Personal Relations	11.50	2.04	8	15	7
Vigor	4.02	0.86	2	5	3

Correlates

For interest's sake, Table 11 shows the uncorrected intercorrelates of the EPQ-R scales and the GPP-I scales from the current sample. Those marked with a * are significant to at least the $p = 0.05$ level.

Table 11

Intercorrelates of the EPQ-R and the GPP-I scales

Var	P	E	N	L	Ad	As	R	E	S	SE	C	O	P	V
P	1	0.34*	-0.02	-0.19	-0.02	0.28*	-0.25	0.17	-0.04	0.14	-0.26	0.00	-0.09	0.04
E		1	-0.40*	-0.06	-0.24	0.57*	-0.42*	-0.05	0.50*	0.51*	-0.42*	0.25	0.09	0.12
N			1	-0.20	0.59*	-0.46*	0.27*	-0.30*	-0.39*	-0.72*	0.07	-0.42*	-0.46*	0.05
L				1	-0.70*	-0.10	0.27	0.04	0.02	0.15	0.13	-0.20	0.09	-0.06
Ad					1	-0.002	-0.37	-0.03	-0.06	-0.47*	-0.14	0.14	-0.33	-0.06
As						1	-0.62*	-0.17	0.51*	0.59*	-0.50*	0.30*	-0.002	0.11
R							1	0.09	-0.46*	-0.11	0.51*	-0.14	0.01	0.13
E								1	-0.48	0.32	0.01	0.06	0.11	-0.27*
S									1	0.44*	-0.23	0.27*	0.22	-0.03
SE										1	-0.17	0.39*	0.29*	-0.07
C											1	-0.19	0.36*	0.06
O												1	0.19	-0.05
P													1	-0.17
V														1

Many of the intercorrelates of scales in the GPP-I were significant, which indicated that the GPP-I scales did not measure separate factors in this sample. A larger sample size would confirm this. In particular, the significant correlation between Ascendancy and Sociability suggested that, contrary to Gordon's (1993) claims, these could be combined to form Extraversion. Of further note was the correlation between EPQ-R Neuroticism and GPP-I Emotional Stability, which was significant at -0.30, suggesting that these two scales measure the same trait (when Emotional Stability is reverse-scored).

EPQ-R

The intercorrelates provided in the EPQ-R manual were for males and females. Since the current sample was largely comprised of males, comparisons will be made using the intercorrelates based on the male sample in the EPQ-R manual. The GPP-I manual reported intercorrelates between the EPI, Extraversion and Neuroticism scales for a sample of male

managers. Hence, these correlates will be reported under the officer sample. No other studies were found which correlated the EPQ-R with the GPP-I, therefore, comparisons of correlates between these questionnaires could not be made. Table 12 shows the intercorrelates reported in the EPQ-R manual (Eysenck & Eysenck, 1991).

Table 12

Intercorrelates of the EPQ-R scales

Scales	Correlation
Psychoticism/Extraversion	0.23
Psychoticism/Neuroticism	0.19
Psychoticism/Lie	-0.34
Extraversion/Neuroticism	0.02
Extraversion/Lie	-0.32
Neuroticism/Lie	-0.25

Again, the sample used in the manual was that comprised of students, teachers, and other volunteers, therefore, any comparisons made should be interpreted with caution. Furthermore, Eysenck and Eysenck (1991) did not state whether their intercorrelates were statistically significant or not. Of note is the difference between the Psychoticism/Neuroticism scales where the current study reported a correlation of -0.02 and the EPQ-R manual (Eysenck & Eysenck, 1991) reported a correlation of 0.19. Similarly, the (statistically significant) correlation between Extraversion and Neuroticism was -0.40 compared with 0.02 in the EPQ-R manual (Eysenck & Eysenck, 1991). These differences may, of course, be attributed to differences in the two samples used, as well as the small sample size of the current study, therefore, no firm conclusions regarding these comparisons can be made.

GPP-I

As no sample was provided in the manual that could be used as a comparison group for the combined sample, intercorrelates of the scales are not presented here. Instead, intercorrelates for the college students will be compared with the officer cadet sample and intercorrelates for the managerial sample will be compared with the officer sample.

Table 13 shows the uncorrected correlates of the EPQ-R and GPP-I scales with the WBQ scales. Those marked with an asterix correlated at least to the $p = 0.05$ level of significance. The bolded diagonal represents the correlates of interest, namely, the EPQ-R and GPP-I scales with their corresponding WBQ scales. As data were missing for two participants on the Emotional Stability scale, this scale was based on a sample of 52.

Table 13

Uncorrected Correlates of the EPQ-R and GPP-I with the WBQ Scales

Var	QP	QE	QN	QL	QAd	QAs	QR	QE	QS	QSE	QC	QO	QP	QV
P	**-0.07**	-0.01	0.11	0.04	0.21	-0.002	-0.01	0.03	0.11	-0.26	0.03	0.005	0.05	-0.02
E	-0.06	**0.13**	0.10	0.01	0.27*	-0.18	0.02	-0.03	-0.04	0.02	0.03	-0.10	0.10	-0.05
N	-0.19	0.12	**-0.05**	-0.11	-0.08	0.18	-0.04	0.28	0.12	-0.01	-0.09	0.01	0.13	0.01
L	-0.09	-0.14	-0.24	**-0.04**	-0.12	-0.13	0.07	0.26	-0.05	0.09	0.19	-0.001	0.18	0.11
Ad	0.16	-0.25	0.35	-0.04	**0.19**	0.07	-0.48*	-0.06	-0.31	-0.09	-0.25	-0.08	-0.43*	-0.10
As	0.06	-0.09	0.18	-0.07	0.18	**-0.11**	0.11	-0.23	0.08	0.03	0.00	-0.08	-0.06	0.06
R	-0.25	0.22	-0.36*	0.02	-0.23	0.05	**0.12**	0.27	0.17	0.15	0.19	0.06	0.31*	0.21
E	0.005	-0.05	-0.01	0.10	-0.09	0.13	-0.06	**0.06**	-0.04	-0.10	0.32*	-0.11	0.04	-0.21
S	0.20	-0.11	0.10	0.08	0.16	-0.21	-0.07	-0.24	**-0.18**	-0.12	-0.27*	0.01	-0.24	0.02
SE	-0.07	-0.04	-0.06	0.13	0.05	-0.19	0.07	-0.13	0.05	**-0.04**	0.22	-0.10	-0.08	0.05
C	-0.05	0.15	-0.26	-0.003	-0.32*	-0.07	0.15	0.11	0.07	0.12	**0.02**	0.08	0.04	0.07
O	0.12	-0.13	0.07	0.01	0.09	-0.003	-0.09	0.13	-0.06	-0.10	0.02	**-0.06**	-0.20	0.06
P	-0.09	-0.02	-0.09	0.14	0.08	-0.15	0.08	-0.10	-0.05	-0.05	-0.11	-0.02	**0.04**	-0.04
V	-0.03	-0.09	0.003	0.08	-0.01	-0.22	0.32*	-0.01	0.07	0.20	-0.09	-0.06	0.20	**0.25**

None of the correlates expected to reach statistical significance reached significance. That is, the EPQ-R and GPP-I scales did not correlate with their corresponding scales in the WBQ. Only the Vigor scale neared significance, with a correlation of 0.25 at the $p = 0.07$

level. These findings could mean one of three things. First, the presence of possible range restriction on some scales of the data could have influenced the strength of the correlates. Second, the EPQ-R and GPP-I questionnaires may not have been accurate predictors of future behaviour. On the other hand, the WBQ may not have been a good measure of the personality profiles measured by the EPQ-R and GPP-I. To resolve this issue, the reliabilities of the WBQ scales were calculated. Although not shown in the table, to see whether the WBQ Neuroticism/WBQ Emotional Stability measured the same trait as the EPQ-R Neuroticism/GPP-I Emotional Stability did, the correlate between the two WBQ scales was calculated and found to be significant ($r = -0.28$, $p = 0.05$). This suggested that the WBQ Neuroticism/WBQ Emotional Stability scales did measure the same trait in the combined sample.

Table 14 shows the alpha reliability coefficient for each scale of the WBQ in the combined sample.

Table 14

Alpha Coefficients for the WBQ Scales

Var	QP	QE	QN	QL	QAd	QAs	QR	QE	QS	QSE	QC	QO	QP	QV
α	0.71	0.32	0.66	0.69	N/A	0.38	N/A	-0.47	-0.17	N/A	N/A	0.59	0.58	N/A

The alpha coefficients could not be calculated for the WBQ Addiction, Responsibility, Self-Esteem, Cautiousness, or Vigor scales as there was only one item measuring each scale. The reliability analysis on the SPSS version 9.0 software package requires at least two items. The above table shows the WBQ to be a very unreliable measure, consequently, the validity of the measure is likely to be poor also. The Psychoticism scale, however, does show a moderate reliability.

Officer Cadets

Descriptive Statistics

Table 15 presents the descriptive statistics of the EPQ-R and the GPP-I for the officer cadet sample, n = 15. All the following results should be interpreted with caution due to the small sample size. The Addiction scale was based on a sample of 14 as data was missing for one participant for this scale.

Table 15

Means, Standard Deviations and Range of Scores on the EPQ-R and GPP-I in the Officer Cadet Sample

Variable	Mean	Standard Deviation	Minimum	Maximum	Range
EPQ-R P	2.73	1.71	0	6	6
EPQ-R E	17.8	2.04	13	20	7
EPQ-R N	4.93	3.31	2	14	12
EPQ-R L	6.87	2.36	3	10	7
EPQ-R A	5.64	1.91	3	9	6
GPP-I A	26.67	3.02	20	32	12
GPP-I R	27.53	3.80	19	34	15
GPP-I E	25.73	3.17	20	31	11
GPP-I S	24.67	3.77	19	31	12
GPP-I SE	105	3.98	92	109	17
GPP-I C	23.73	5.44	12	33	21
GPP-I O	31.4	4.05	24	38	14
GPP-I P	26.6	4.17	15	33	18
GPP-I V	30.13	3.98	25	38	13

EPQ-R

Table 16 shows the means and standard deviations for the 16-20yrs age group taken from the EPQ-R manual (Eysenck & Eysenck, 1991).

Table 16

Means and Standard Deviations of the EPQ-R Scales for 16 - 20 year olds

Scale	Mean	Standard Deviation
Psychoticism	9.57	5.26
Extraversion	15.97	5.26
Neuroticism	11.12	5.68
Lie	5.37	4.18

The current sample Psychoticism scale mean and standard deviation were much lower than those reported in the EPQ-R manual. This indicated a level of range restriction. Similarly, Extraversion and Neuroticism showed different standard deviations that may be indicative of range restriction. Addiction scale means and standard deviations were not presented as these were not available for this age group in the manual.

GPP-I

Table 17 shows the mean and standard deviations reported in the GPP-I manual (Gordon, 1993) for the sample of college students. The manual did not provide a mean or standard deviation for the Self-Esteem score for this sample.

Table 17

Means and Standard Deviations for the GPP-I scales in a College Student Sample

Scale	Mean	Standard Deviation
Ascendancy	23.8	5.8
Responsibility	24.9	6.0
Emotional Stability	23.6	6.5
Sociability	22.7	5.9
Self-Esteem	N/A	N/A
Cautiousness	22.3	6.6
Original Thinking	26.6	6.5
Personal Relations	24.2	6.2
Vigor	25.9	6.8

Similar to the combined sample, the Ascendancy, Responsibility, Emotional Stability, Sociability and Original Thinking scales all showed higher means than those reported in the GPP-I manual. In addition, the Vigor scale showed a higher mean in the current study. Furthermore, these scales showed lower standard deviations than those reported in the manual, suggesting the presence of range restriction. The greatest differences in means were between the Original Thinking and Vigor scales. It was logical to think that officer cadets possessed higher Ascendancy, Responsibility, Emotional Stability, Sociability, and Original Thinking scores than the general population as these were deemed good characteristics for successful performance in an officer role. Original Thinking contained similar behaviours to the FFM Openness to Experience, which has been linked with successful training performance (see Barrick & Mount, 1991). However, the small sample size prevented any firm conclusions regarding the officer cadet sample.

For completeness, Table 18 shows the descriptive statistics of the WBQ scores in the officer cadet sample.

Table 18

Means, Standard Deviations, and Range of the WBQ Scales

Variable	Mean	Standard Deviation	Minimum	Maximum	Range
Psychoticism	6.33	2.23	3	10	7
Extraversion	6.13	1.41	4	9	5
Neuroticism	5.67	1.88	2	8	6
Lie	4.20	1.93	2	8	6
Addiction	1.53	1.06	1	4	3
Ascendancy	7.40	1.45	4	9	5
Responsibility	4.13	0.92	3	5	2
Emotional Stability	6.53	1.19	5	9	4
Sociability	11.27	1.22	9	13	4
Self-Esteem	3.73	0.88	2	5	3
Cautiousness	3.07	0.70	2	4	2
Original Thinking	7.40	0.91	6	9	3
Personal Relations	11.40	2.32	8	15	7
Vigor	4.0	0.65	3	5	2

Correlates

Table 19 shows the uncorrected intercorrelates of the EPQ-R scales and the GPP-I scales in the officer cadet sample. Those marked with a * are significant to at least the $p = 0.05$ level. Once more, many of the correlates were significant. In particular, the Ascendancy and Sociability correlate was significant suggesting that these scales measured the Extraversion trait. The EPQ-R Neuroticism/GPP-I Emotional Stability correlate was not significant, which was more than likely due to the small sample size.

Table 19

Intercorrelates of the EPQ-R and the GPP-I Scales

Var	P	E	N	L	Ad	As	R	E	S	SE	C	O	P	V
P	1	0.43	-0.37	-0.01	-0.16	0.15	-0.05	0.46	0.14	0.45	0.26	0.05	0.59*	-0.04
E		1	-0.70*	-0.38	-0.25	0.30	-0.26	0.04	0.63*	0.61*	0.12	0.44	0.45	0.07
N			1	0.26	0.52	-0.36	0.07	-0.26	-0.38	-0.81*	-0.10	-0.73*	-0.35	-0.40
L				1	-0.19	-0.07	0.30	0.19	-0.15	0.11	0.02	-0.39	0.15	0.08
Ad					1	-0.15	-0.30	-0.04	-0.03	-0.50	-0.10	-0.61	-0.41	-0.51
As						1	-0.78*	-0.24	0.62*	0.39	-0.52	0.43	-0.07	0.03
R							1	0.32	-0.57*	0.09	0.63*	-0.19	0.25	0.43
E								1	-0.49	0.33	0.12	-0.11	0.13	0.20
S									1	0.40	-0.08	0.44	0.29	-0.17
SE										1	0.31	0.50	0.59*	0.48
C											1	-0.13	0.71*	0.13
O												1	0.15	0.19
P													1	-0.08
V														1

GPP-I

The EPQ-R manual did not provide intercorrelates of the scales for the 16-20yr age group, however, the GPP-I manual provided intercorrelates for the sample of college students. Only the male data is presented below.

Table 20

Intercorrelates of the GPP-I Scales for a College Student Sample

Scale	A	R	E	S	C	O	P	V
A		0.04	0.09	0.64	-0.05	0.36	0.22	0.23
R			0.56	-0.01	0.44	0.26	0.40	0.35
E				-0.13	0.27	0.17	0.40	0.16
S					-0.14	0.16	0.19	0.18
C						0.23	0.42	0.15
O							0.33	0.32
P								0.18
V								

Some of the correlates in the current study were quite different to those reported in the manual. For example, the Ascendancy scale was significantly and negatively correlated with the Responsibility scale at -0.78, compared with the nonsignificant correlation of 0.04 reported in the GPP-I manual (Gordon, 1993). The Emotional Stability scale in the current study correlated -0.24 with Ascendancy in the current study, compared with 0.09 in the manual (Gordon, 1993). Furthermore, Responsibility correlated -0.57 with Sociability in the current study, compared with -0.01 reported in the manual (Gordon, 1993). These findings may be because undesirable profiles were selected out, leading to range restriction. However, the small sample size of the officer cadet sample renders any meaningful comparisons difficult.

Table 21 shows the uncorrected correlates of the EPQ-R and GPP-I scales with the corresponding WBQ scales in the officer cadet sample. Those marked with a * were significant to at least the $p = 0.05$ level.

Table 21

Uncorrected Correlates of the EPQ-R and GPP-I Scales with the WBQ Scales

Var	QP	QE	QN	QL	QAd	QAs	QR	QE	QS	QSE	QC	QO	QP	QV
P	**0.21**	-0.01	-0.14	0.28	-0.07	0.22	0.12	0.08	0.14	-0.15	0.08	0.12	-0.13	0.38
E	0.09	**0.31**	-0.32	0.01	0.48	0.005	0.17	0.28	0.05	-0.03	-0.44	0.39	-0.16	0.32
N	-0.17	-0.18	**0.18**	-0.24	-0.27	0.08	-0.26	-0.21	-0.21	-0.06	0.002	-0.28	-0.03	-0.43
L	-0.52*	0.03	-0.30	**-0.06**	0.17	-0.28	0.41	0.03	0.31	0.02	0.52*	-0.01	0.55*	0.51
Ad	0.36	-0.18	0.13	-0.03	**0.03**	0.27	-0.60*	-0.27	-0.46	0.03	-0.48	-0.10	-0.61*	-0.59*
As	0.34	-0.06	0.13	0.09	0.22	**0.18**	0.17	-0.13	0.01	0.07	-0.22	0.21	-0.36	0.36
R	-0.33	0.20	-0.14	0.30	-0.15	-0.27	**0.02**	0.14	0.31	0.003	0.47	-0.11	0.54*	-0.06
E	0.15	0.09	-0.24	0.30	-0.02	0.35	0.06	**0.14**	0.28	0.33	0.33	0.14	0.14	0.17
S	0.06	-0.13	-0.10	-0.24	0.33	-0.17	0.22	0.04	**-0.18**	-0.29	-0.50	0.25	-0.33	0.23
SE	0.00	0.32	-0.36	0.34	0.27	-0.11	0.39	0.32	0.50	**0.08**	0.05	0.47	0.16	0.58*
C	-0.36	0.37	-0.34	0.11	-0.32	-0.41	0.08	0.34	0.48	-0.22	**0.17**	0.22	0.32	-0.08
O	0.32	-0.06	0.16	0.21	0.03	-0.05	0.14	-0.03	-0.02	-0.05	-0.16	**0.05**	-0.20	0.13
P	-0.28	0.39	-0.41	0.06	-0.14	-0.28	0.45	0.55*	0.60*	-0.19	0.03	0.48	**0.21**	0.47
V	-0.20	0.03	-0.16	0.32	0.19	-0.12	0.01	-0.12	0.14	0.07	0.35	-0.11	0.53*	**0.16**

As with the combined sample, none of the EPQ-R and GPP-I scales correlated significantly with their corresponding WBQ scales as represented by the bold diagonal. Some correlates were even negative; taken together, this suggests that the WBQ may not have been an accurate measure of the behaviours associated with the EPQ-R and GPP-I personality profiles in this sample. Furthermore, range restriction may have influenced the size of the correlates.

To confirm this suggestion, Table 22 shows the alpha reliability coefficient for each scale of the WBQ in the officer cadet sample.

Table 22

Alpha Coefficients for the WBQ Scales in the Officer Cadet Sample

Var	QP	QE	QN	QL	QAd	QAs	QR	QE	QS	QSE	QC	QO	QP	QV
α	0.76	-0.06	0.92	0.80	N/A	0.50	N/A	-0.49*	0.01	N/A	N/A	-0.18	0.70	N/A

The alpha coefficients could not be calculated for the WBQ Addiction, Responsibility, Self-Esteem, Cautiousness, or Vigor scales as there was only one item measuring each scale. In addition, for some reason, the SPSS software package calculated the alpha coefficient of Emotional Stability at greater than -1, therefore, the average inter-item correlation was used instead. In the officer cadet sample, the WBQ actually showed good reliability, particularly for the Neuroticism scale and the Lie scale. The remaining scales showed low to moderate reliability in this sample. It is possible that reliability was improved because the immediate superiors were also officer cadets and, therefore, knew the participants better as they worked closely together. As the WBQ Neuroticism scale demonstrated excellent reliability with an alpha coefficient of 0.92, the non-significant Neuroticism correlation probably meant the EPQ-R Neuroticism scale did not measure what it was supposed to measure in this sample. The Lie scale correlation, at -0.06 suggested quite strongly that the EPQ-R Lie scale may not have measured what it was supposed to measure in this sample.

Officers

Descriptive Statistics

Table 23 shows the descriptive statistics of the EPQ-R and GPP-I scales in the officer sample. The Addiction scale was based on a sample of 7.

Table 23

Means, Standard Deviations and Range of Scores on the EPQ-R and GPP-I in the Officer Sample

Variable	Mean	Standard Deviation	Minimum	Maximum	Range
EPQ-R P	2.59	2.04	0	9	9
EPQ-R E	17.79	3.00	10	22	11
EPQ-R N	5.33	4.65	0	18	18
EPQ-R L	8.92	3.94	2	16	14
EPQ-R A	2.71	2.63	0	7	7
GPP-I A	24.90	3.73	13	31	18
GPP-I R	29.31	2.88	23	35	12
GPP-I E	27.05	3.58	20	33	13
GPP-I S	23.67	3.52	12	30	18
GPP-I SE	104.64	5.10	84	109	25
GPP-I C	23.85	4.46	14	35	21
GPP-I O	31.18	4.58	22	39	17
GPP-I P	26.72	4.02	16	33	17
GPP-I V	28.92	4.33	21	38	17

EPQ-R

The means and standard deviations reported in the EPQ-R manual (Eysenck & Eysenck, 1991) were averaged across the two age groups of 21-30yrs and 31-40yrs since only one officer fell outside this age group. These averaged means and standard deviations are presented in Table 24 below.

Table 24

Averaged Means and Standard Deviations for the 21-30 year and 31 - 40year Age Groups

Scale	Mean	Standard Deviation
Psychoticism	7.67	4.07
Extraversion	13.21	5.66
Neuroticism	11.5	5.54
Lie	6.10	3.49

Once again, the Psychoticism scale showed a much lower mean and standard deviation in the current sample, at 2.59 and 2.04 respectively. This finding indicated that officers were lower in the trait of Psychoticism than the general population, but that range restriction effects may also have been present. In addition, the Extraversion mean was higher and the Neuroticism mean was lower in the current sample than in the general population. Range restriction effects may possibly have been present for the Extraversion scale but not the Neuroticism scale, as the standard deviations were similar for this scale. Lie scale scores were similar in both studies, although the current study showed a slightly higher mean. Again, the EPQ-R manual (Eysenck & Eysenck, 1991) did not report separate scores for the Addiction scale, therefore, statistics for this scale could not be compared.

GPP-I

The GPP-I manual presented means and standard deviations for a group of male managers. These statistics are presented in Table 25 below.

Table 25

Means and Standard Deviations for Male Managers

Scale	Mean	Standard Deviation
Ascendancy	23.1	5.4
Responsibility	28.8	4.2
Emotional Stability	26.8	5.0
Sociability	21.5	5.3
Self-Esteem	N/A	N/A
Cautiousness	28.5	4.5
Original Thinking	27.1	5.2
Personal Relations	29.0	4.5
Vigor	26.9	4.9

In addition, the GPP-I manual presented means for US and UK Army officers. The data for the US sample, as it had the larger sample size of the two, is presented in Table 26 below.

Table 26

Means for the GPP-I Scales of a US Sample of Army Officers taken from Gordon (1993)

Scale	Mean
Ascendancy	24.3
Responsibility	28.9
Emotional Stability	28.0
Sociability	21.2
Self-Esteem	N/A
Cautiousness	25.6
Original Thinking	27.8
Personal Relations	23.7
Vigor	28.7

The first four scales all showed lower standard deviations in the present sample than in the sample reported in the manual (Gordon, 1993). This indicated range restriction effects. No mean was presented for the Self-Esteem scale in either the male manager or military samples. Noteworthy is that Original Thinking scores in the current study were similar to those reported in the manual, which further supports Barrick & Mount's (1991) findings that Openness to Experience was predictive of training performance but not job performance. Only the Cautiousness scale mean in the current sample appeared lower to the mean reported in the manual (Gordon, 1993). This finding suggested that officers were lower in Cautiousness than other types of managers. Considering the nature of the military job, where calculated risks are often taken, this finding was not surprising.

Finally, only the Original Thinking and Personal Relations scales showed differences between the US Army officer means and the current sample army officer means. In the current sample, the Original Thinking and Personal Relations means were higher than the US Army sample. Perhaps this finding meant that NZ Army officers were more curious and creative and, simply, got on with other people better than the Americans. The other scale means were fairly similar in comparison.

For completeness, Table 27 shows the descriptive statistics of the WBQ scores in the officer sample. The Emotional Stability scale was based on a sample of 37 as data from two participants were missing for this scale.

Table 27

Means, Standard Deviations, and Range of the WBQ Scales in the Officer Sample

Variable	Mean	Standard Deviation	Minimum	Maximum	Range
Psychoticism	6.28	2.27	3	12	9
Extraversion	7.26	1.35	3	9	6
Neuroticism	4.74	1.39	2	8	6
Lie	4.21	1.40	2	7	5
Addiction	1.54	0.88	1	4	3
Ascendancy	7.59	1.31	5	10	5
Responsibility	3.95	0.86	2	5	3
Emotional Stability	6.92	1.09	5	9	4
Sociability	10.77	1.13	8	13	5
Self-Esteem	3.87	0.95	1	5	4
Cautiousness	2.97	0.74	1	5	4
Original Thinking	7.31	1.59	3	10	7
Personal Relations	11.54	1.94	8	15	7
Vigor	4.03	0.93	2	5	3

Correlates

Table 28 shows the uncorrected intercorrelates of the EPQ-R scales and the GPP-I scales in the officer sample. Again, the Ascendancy and Sociability correlate was significant ($r = 0.46$, $p = 0.05$) as were many of the GPP-I correlates. In particular, the EPQ-R Neuroticism/GPP-I Emotional Stability (reverse-scored) correlate was significant $r = -0.33$, $p = 0.05$), suggesting these two scales measure the same trait. Those marked with a * were significant to at least the $p = 0.05$ level.

Table 28

Intercorrelates of the EPQ-R and the GPP-I Scales in the Officer Sample

Var	P	E	N	L	Ad	As	R	E	S	SE	C	O	P	V
P	1	0.32*	0.06	-0.22	0.30	0.32*	-0.34*	0.10	-0.10	0.09	-0.47*	-0.01	-0.31	0.06
E		1	-0.35*	-0.02	-0.08	0.64*	-0.51*	-0.07	0.48*	0.49*	-0.60*	0.21	0.002	0.13
N			1	-0.29	0.43	-0.49*	0.35*	-0.33*	-0.40*	-0.70*	0.13	-0.35*	-0.50*	0.17
L				1	-0.89*	-0.04	0.21	-0.04	0.11	0.17	0.17	-0.16	0.07	-0.06
Ad					1	-0.35	-0.24	0.67	-0.45	-0.44	0.02	0.46	0.15	0.49
As						1	-0.55*	-0.11	0.46*	0.63*	-0.52	0.27	0.02	0.11
R							1	-0.06	-0.38*	-0.18	0.46*	-0.12	-0.12	0.06
E								1	-0.46*	0.33	-0.04	0.12	0.10	-0.38*
S									1	0.46*	-0.30	0.20	0.20	-0.01
SE										1	-0.34*	0.35*	0.21	-0.22
C											1	-0.21	0.19	0.03
O												1	0.21	-0.13
P													1	-0.21
V														1

GPP-I

The EPQ-R manual did not provide intercorrelates for an adequate comparison group, however, the GPP-I manual provided intercorrelates on a sample of male and female managers for the GPP-I scales. The UK sample data is presented below in Table 29. Furthermore, the GPP-I manual provided intercorrelates for a sample of 130 managers with the EPI scales Neuroticism and Extraversion. These intercorrelates are also presented in Table 29.

Table 29

Intercorrelates for the GPP-I and EPI Scales

Scale	A	R	E	S	C	O	P	V
A		-0.25	-0.02	0.49	-0.28	0.43	0.05	0.40
R			0.46	-0.15	0.35	0.10	0.13	0.13
E				-0.29	0.24	0.19	0.28	0.07
S					-0.26	0.13	0.12	0.22
C						-0.11	0.28	-0.08
O							0.12	0.26
P								-0.07
V								
EPI E	0.46	-0.24	-0.14	0.72	-0.37	0.03	0.09	0.39
EPI N	-0.10	-0.20	-0.45	0.00	-0.05	-0.15	-0.14	-0.24

Several EPQ-R and GPP-I correlates in the current sample were different to those reported in the manual (Gordon, 1993). In particular, the Ascendancy/Cautiousness correlation at -0.52 in the current sample compared with -0.28 in the manual; Ascendancy/Original Thinking at 0.27 and 0.43 respectively; Responsibility/Emotional Stability at -0.06 and 0.46 respectively; and Emotional Stability/Vigor at -0.38 and 0.07 respectively. These differences may have been due to the presence of range restriction, differences in the sample such as different types of management style in the military compared with civilian management, or the sample size for the current study may have been too small.

Several EPQ-R correlates were also different to the EPI correlates reported in the GPP-I manual (Gordon, 1993). These were Extraversion/Responsibility at -0.51 and -0.24 respectively, Extraversion/Sociability at 0.48 and 0.72 respectively, Extraversion/Cautiousness at -0.60 and -0.37 respectively, Extraversion/Vigor at 0.13 and 0.39 respectively, Neuroticism/Responsibility at 0.35 and -0.20 respectively, Neuroticism/Sociability at -0.40 and 0.00 respectively, and Neuroticism/Personal Relations at -0.50 and -0.24 respectively. These differences were, more than likely, due to differences in the EPI and EPQ-R versions, as well as the small sample size of the current study.

Table 30 shows the uncorrected correlates of the EPQ-R and GPP-I scales with the corresponding WBQ scales in the officer sample. Again, intercorrelates for the Emotional Stability scale were based on a sample of 37 due to the missing data from two participants. The EPQ-R Addiction/WBQ Emotional Stability correlate was based on a sample of 6.

Table 30

Uncorrected Correlates of the EPQ-R and GPP-I Scales with the WBQ Scales

Var	QP	QE	QN	QL	QAd	QAs	QR	QE	QS	QSE	QC	QO	QP	QV
P	**-0.16**	0.01	0.21	-0.06	0.32*	-0.07	-0.06	0.03	0.09	-0.29	0.01	-0.02	0.12	-0.11
E	-0.10	**0.10**	0.25	0.02	0.22	-0.24	-0.01	0.12	-0.06	0.03	0.14	-0.17	0.18	-0.12
N	-0.19	0.20	**-0.12**	-0.07	-0.02	0.21	0.02	0.43*	0.22	-0.002	-0.11	0.05	0.19	0.10
L	-0.001	-0.33*	-0.17	**-0.04**	-0.21	-0.13	0.03	-0.41*	-0.07	0.09	0.15	0.01	0.09	0.05
Ad	-0.02	-0.21	0.77*	0.08	**0.38**	-0.46	-0.77*	-0.09	-0.82*	-0.70	0.25	-0.58	-0.20	-0.25
As	-0.02	-0.01	0.13	-0.14	0.18	**-0.19**	0.06	-0.26	0.04	0.03	0.05	-0.14	0.05	0.01
R	-0.23	0.11	-0.42*	-0.17	-0.29	0.20	**0.22**	0.30	0.18	0.21	0.09	0.13	0.18	0.32*
E	-0.04	-0.18	0.17	0.02	-0.13	0.03	-0.08	**-0.001**	-0.10	-0.25	0.34*	-0.16	-0.01	-0.31
S	0.26	-0.04	0.17	0.25	0.08	-0.22	-0.21	0.35*	**-0.22**	-0.05	-0.19	-0.04	-0.20	-0.04
SE	-0.10	-0.13	0.05	0.04	-0.03	-0.22	-0.03	-0.26	-0.10	**-0.08**	0.26	-0.20	0.05	-0.06
C	0.08	0.05	-0.23	-0.07	-0.32*	0.10	0.18	-0.01	-0.12	0.26	**-0.05**	0.04	-0.11	-0.12
O	0.06	-0.16	0.02	-0.08	0.12	0.02	-0.17	- 0.17	-0.08	-0.11	0.08	**-0.08**	-0.21	0.05
P	-0.01	-0.19	0.07	0.19	0.18	-0.09	-0.07	0.40*	-0.31	-0.003	-0.16	-0.13	**-0.04**	-0.17
V	0.02	-0.07	0.02	-0.02	-0.09	-0.25	0.42*	0.06	0.01	0.26	-0.25	-0.06	0.07	**0.28**

None of the correlates represented on the bold diagonal were significant. The Vigor scale correlation neared significance with the level at $p = 0.08$. Before any conclusion was made regarding the effectiveness of the questionnaires used in this sample, the reliability coefficients for the WBQ were examined.

Table 31 shows the alpha reliability coefficient for each scale of the WBQ in the officer sample.

Table 31

Alpha Coefficients for each Scale of the WBQ in the Officer Sample

Var	QP	QE	QN	QL	QAd	QAs	QR	QE	QS	QSE	QC	QO	QP	QV
α	0.72	0.38	0.44	0.72	N/A	0.33	N/A	0.18	-0.20	N/A	N/A	0.69	0.58	N/A

The alpha coefficients could not be calculated for the WBQ Addiction, Responsibility, Self-Esteem, Cautiousness, or Vigor scales as there was only one item measuring each scale. Only the Psychoticism and Lie scales showed reasonable reliability for the WBQ. Therefore, it can be said with reasonable confidence that the EPQ-R Psychoticism and Lie scales did not accurately measure behaviour associated with the officer role in the officer sample. The other scales did not reliably measure the behaviours to make an accurate interpretation possible.

Hypothesis 2: There should be a high correlation between EPQ-R N and EPQ-R L, and EPQ-R P and EPQ-R L if faking was present during the selection process.

As there was no control group used in the study, the Lie scale scores could not be compared to see whether these scores were elevated compared with a non-applicant sample. According to Eysenck and Eysenck (1991), the correlation between Neuroticism and the Lie scale should be examined, as this correlation gives a good indication of faking. In particular, if the correlation approached -0.50 or greater, faking was likely to be present; if the correlation was low, and where there was little motivation to fake, then this measured a conformity factor (Eysenck & Eysenck, 1991). Similarly, there was usually a high negative correlation between Psychoticism and Lie Scale scores if faking was present (Eysenck & Eysenck, 1991).

The correlation reported in the combined sample for Neuroticism and the Lie scale was -0.20. This correlation, although negative, was not significant and therefore, was likely to reflect a conformity factor. In the officer cadet sample, the correlation was 0.26; this also

was not significant. In the officer sample, the correlation was -0.29. Again, this correlation was not significant, and likely to measure a conformity factor.

The correlation between Psychoticism and the Lie scale in the whole sample was -0.19 and was not significant. In the officer cadet sample, the correlation was -0.01 and not significant. Finally, the correlation in the officer sample was -0.22 and not significant.

Hypothesis 3: Officers who have been in longer should show similar personality profiles than those who have only been in a short time.

Independent sample t-tests revealed no significant differences in scores on each scale of the EPQ-R and GPP-I scales between those who had served greater than 3 years and those who had served less than three years amongst the officers. Post hoc comparisons could not be carried out as some groups contained less than two cases.

Hypothesis 4: Immediate superiors of the same gender/ethnicity should rate more highly those of the same gender/ethnicity.

This analysis could not be carried out due to an insufficient sample size for each gender.

Hypothesis 5: If selection was effective, items on the WBQ measuring High Psychoticism, High Neuroticism, Low Emotional Stability, Low Ascendancy and Low Cautiousness, should not be endorsed.

This hypothesis was tested with the caveat that officer selection was not based purely on personality questionnaire data but, rather, on a range of data obtained from the OSB process. Therefore, these undesirable profiles may still occur if the officer performed well in other aspects of the OSB.

A high score on the WBQ is defined as either 4 or 5, and a low score is defined as either 1 or 2. Some items were reverse-scored.

High Psychoticism

For the Psychoticism scale in the combined sample, there were three items, two of which were reverse-scored. On the first Psychoticism question, 9% of subjects scored a 4 or greater. On the second question measuring Psychoticism, 5% of subjects scored a 2 or less (reverse-scored). On the third Psychoticism question, 11% scored a 2 or less (reverse-scored) in the combined sample. The reliability of the WBQ Psychoticism scale for the whole sample was 0.71, therefore it was reasonable to assume that, although there were some officers and cadets displaying High Psychoticism behaviours, this number was fairly small.

As officer cadets were still yet to graduate from officer training, it was possible that some of these High Psychoticism scorers may be selected out still, as these data were based on the combined sample. In the officer cadet sample, no officer cadets scored 4 or greater on the first, or 2 or less (reverse-scored) on the second Psychoticism questions. Two of the fifteen officer cadets (13%) in the sample scored 2 or less for the third Psychoticism question (reverse-scored). The reliability for the Psychoticism scale in the officer cadet sample was 0.76 so, again, it was reasonable to assume, despite the small sample size, that High Psychoticism scorers were selected out. In the officer sample, 5 of the 39 officers (12.8%) scored higher than 4 on the first Psychoticism question, 3 (7.7%) scored 2 or less on the second Psychoticism question, and 4 (10.3%) scored 2 or less on the third Psychoticism question. With the alpha coefficient at 0.72 for the officer sample, it was reasonable to assume that most people with High EPQ-R Psychoticism scores were successfully selected out.

High Neuroticism

Two items in the WBQ measured the Neuroticism scale. In the combined sample, 14 of the 54 (25.9%) participants scored 4 or greater on the first question measuring Neuroticism and 5 of the 54 (9.3%) participants scored 4 or greater on the second question measuring Neuroticism. The reliability for the WBQ Neuroticism scale in the combined sample was 0.66, therefore, it could not be said with any degree of certainty whether High EPQ-R Neuroticism scorers were successfully selected out.

In the officer cadet sample, 6 of the 15 officer cadets (40%) scored a 4 or greater on the first WBQ Neuroticism question and 3 of the 15 (20%) scored 4 or greater on the second WBQ Neuroticism item. With a reliability of 0.92 for this scale in the officer cadet sample, it could confidently be said that High Neuroticism scorers were not effectively selected out on the basis of their EPQ-R scores. In the officer sample, however, 8 of the 39 officers (20.5%) scored a 4 or greater on the first WBQ Neuroticism item, and 2 of the 39 (5.1%) scored a 4 or greater on the second WBQ Neuroticism item. The reliability for the WBQ Neuroticism scale in the officer sample was only 0.44, therefore it could not be said whether High Neuroticism scorers were effectively selected out.

Low Emotional Stability

Two questions, one reverse-scored, measured Emotional Stability in the WBQ. In the combined sample, 14 of the 54 participants (25.9%) scored either a 2 or less on the first item, and 4 of 52 participants (7.7%), (data was missing for two participants on this item) scored 4 or more on the second item (reverse-scored). The reliability for the combined sample for this scale was -0.47, therefore no conclusions could be drawn for these data. Similarly, in the officer cadet sample, 7 of the 15 (46.7%) scored 2 or less on the first item, and 3 of the 15 (20%) scored 4 or more on the second item (reverse-scored), however, as the correlation coefficient of the two items was -0.49, no conclusions could be drawn. In the officer sample, 7 of the 39 (17.9%) scored a 2 or less on the first item, and 1 of 37 (2.7%) scored 4 or more (reverse-scored) on the second item (data was missing for two officers on this item). The alpha reliability for the officer sample was 0.18, so no conclusions could be drawn from these data either.

Low Ascendancy

Two questions, one reverse-scored, measured Ascendancy in the WBQ. In the combined sample, 5 of the 54 participants (9.26%) scored a 4 on the first item (reverse-scored), and 4 of the 54 participants (7.40%) scored a 2 or less on the second item. The reliability for the WBQ Ascendancy scale was low at 0.38 for the combined sample, so no

firm conclusions could be drawn. In the officer cadet sample, 2 out of 15 officer cadets (13.3%) scored a 4 on the first item (reverse-scored), and 1 of the 15 officer cadets (6.6%) scored a 2 on the second item. The reliability, although higher at 0.50, was not large enough to make firm conclusions regarding these findings. Finally, 3 of the 39 officers (7.7%) scored a 4 on the first item (reverse-scored), and 3 of the 39 (7.7%) scored a 1 or a 2 on the second item. Again, with a low reliability of 0.33 for the WBQ Ascendancy scale, no firm conclusions could be drawn.

Low Cautiousness

Only one item in the WBQ measured Cautiousness in the WBQ, consequently, the reliability coefficient could not be calculated. For completeness, in the combined sample, 11 of the 54 participants (20.4%) scored 2 or less, in the officer cadet sample, 3 of the 15 participants (20%) scored 2 or less, and in the officer sample, 8 of the 39 participants (20.5%) scored 2 or less.

Interactions

Chi square analyses were performed to assess the relationship of these interactions measured by the behaviour questionnaire to the original questionnaires.

Officer Cadets

High Psychoticism/Low Cautiousness reported $\chi^2 = 4.61$, df = 3, $\alpha = 0.20$; Low Extraversion/Low Sociability showed a $\chi^2 = 5.30$, df = 2, $\alpha = 0.07$; High Psychoticism/High Ascendancy showed a $\chi^2 = 1.54$, df = 3, $\alpha = 0.67$; High Responsibility/High Vigor showed a $\chi^2 = 1.55$, df = 3, $\alpha = 0.67$; High Original Thinking/Low Vigor showed a $\chi^2 = 1.09$, df = 3, $\alpha = 0.78$; High Personal Relations/Low Ascendancy showed a $\chi^2 = 2.31$, df = 3, $\alpha = 0.51$; Low Responsibility/Low Vigor showed a $\chi^2 = 0.72$, df = 3, $\alpha = 0.87$; High Vigor/High Original Thinking/Low Responsibility had no data available; and High Sociability/Low Personal Relations showed a $\chi^2 = 1.99$, df = 3, $\alpha = 0.58$. None of these chi squares were significant, although the chi square for High Extraversion/Low Sociability neared significance at the 0.07 level.

Officers

In the officer sample, none of the chi squares were significant either. High Psychoticism/Low Cautiousness reported $\chi^2 = 1.95$, df = 3, $\alpha = 0.58$; Low Extraversion/Low Sociability showed a $\chi^2 = 3.17$, df = 3, $\alpha = 0.37$; High Psychoticism/High Ascendancy showed a $\chi^2 = 1.67$, df = 3, $\alpha = 0.64$; High Responsibility/High Vigor showed a $\chi^2 = 5.63$, df = 3, $\alpha = 0.13$; High Original Thinking/Low Vigor showed a $\chi^2 = 3.49$, df = 3, $\alpha = 0.32$; High Personal Relations/Low Ascendancy showed a $\chi^2 = 1.0$, df = 3, $\alpha = 0.80$; Low Responsibility/Low Vigor showed a $\chi^2 = 1.43$, df = 3, $\alpha = 0.70$; High Vigor/High Original Thinking/Low Responsibility showed a $\chi^2 = 8.33$, df = 7, $\alpha = 0.31$; and High Sociability/Low Personal Relations showed a $\chi^2 = 6.01$, df = 3, $\alpha = 0.11$. Again, none of these chi square analyses were significant.

DISCUSSION

Hypotheses

Hypothesis 1

The first hypothesis that scores on the EPQ-R and GPP-I would correlate highly with corresponding scores on the WBQ was not supported in any of the samples. None of the scales correlated significantly with their WBQ counterparts, showing no evidence for convergent validity. However, the Vigor scale correlated 0.28 with its corresponding scale in the WBQ, at a 0.08 level of significance in the officer sample. The reliability of the Vigor scale could not be assessed as there was only one item measuring Vigor in the WBQ. According to the GPP-I manual (Gordon, 1993), Vigor correlated 0.39 with the Eysenck Personality Inventory Extraversion scale, 0.48 with the Personality Research Form Achievement scale, and 0.25 with the Sociability scale of the California Psychological Inventory Rational scales. Therefore, Vigor could be said to measure a similar construct as Extraversion. Extraversion has been shown to correlate significantly with job performance. For example, in Black's (1997) study of personality testing in NZ Police selection, he found that Extraversion correlated 0.16 at the 0.01 level of significance with subsequent police performance. In Barrick and Mount's (1991) meta-analysis, they found that Extraversion predicted performance in jobs where contact with people was high, such as managers (0.18) and salespeople (0.15). However, in the current study, EPQ-R Extraversion correlated 0.12, 0.07, and 0.13 with Vigor in the combined sample, officer cadet sample, and officer sample respectively. Furthermore, EPQ-R Extraversion correlated 0.13, 0.31, and 0.10 respectively with WBQ Extraversion. Barrick and Mount's (1991) findings were not, therefore, supported in the present research. Given a larger sample size, these correlates, may have been higher and, therefore, significant.

Some of the correlates were negative, which is not surprising considering the small sample sizes (see Tett et al., 1991). For example, the correlation between GPP-I Emotional Stability and WBQ Emotional Stability for the combined sample was -0.13. There are two possible explanations for this finding. First, the EPQ-R and GPP-I scales may measure the behaviours they are supposed to measure. This means that the two questionnaires are not valid predictors of future behaviour and, therefore, are not effective selection tools.

However, a more likely explanation concerns the WBQ scales. The alpha coefficients ranged from -0.47 to 0.71 in the combined sample, -0.49 to 0.92 in the officer cadet sample, and -0.20 to 0.72 in the officer sample. These coefficients reflect a low to moderate reliability for the WBQ scales. As reliability and validity are closely intertwined, a scale with low reliability cannot be valid (Clark & Watson, 1995). Therefore, a more likely explanation for the nonsignificant findings is that the WBQ scales do not accurately measure the personality profiles and, hence, constructs they are supposed to measure.

The first hypothesis, then, may have been supported if the reliability of the WBQ scales was higher. One major problem was that some reliability estimates could not be calculated because only one item was written for that particular scale. Using Psychoticism as an example, the *Statistica* software programme calculated that, to have a reliability coefficient of 0.80 for the Psychoticism scale, six additional items would be required, to give a total of nine items. Wanous, Reichers, and Hudy (1997) claimed that single-item measures were discouraged because they often showed unacceptable levels of reliability. They argued that single-item measures could only be used for factual measures such as age, number of jobs, or where the psychological construct was narrow (Wanous et al., 1997). However, for complex constructs, particularly personality measures, multiple items should be used (Wanous et al., 1997). Wanous et al. (1997) performed a meta-analysis of job satisfaction studies that used single-item measures to assess their usefulness. They rearranged the attenuation formula to calculate the reliability estimate for the single-item measures by adding an assumed correlation between the two overall job satisfaction measures. This method could not be replicated in the current research as Wanous et al. (1997) did not provide enough information on the procedure. The mean observed correlation between the single-item measures and the overall job satisfaction measures was 0.63 and, when corrected for unreliability, was 0.67 (Wanous et al., 1997). Interestingly, though, they found that single-item measures were more robust than scales. Nevertheless, Wanous et al. (1997) recommended the use of single-item measures only when other constraints prevent the use of well-constructed scales with an acceptable number of items.

Simply increasing the number of items to increase the reliability estimate will not, on its own, solve the problem of low alpha coefficients. Clark and Watson (1995) argued that if the number of items on a scale was increased, avoiding high reliability estimates was difficult, as the alpha coefficient essentially was a function of the number of items and the average

intercorrelates amongst those items. Instead, Clark and Watson (1995) argued that, as the number of items was essentially irrelevant, the average inter-item correlation should be used as a measure of internal consistency as this did not rely on the number of items.

To increase reliability estimates, and for correction formulas to provide useful additional information, the most important factor, as Hunter et al (1982) pointed out, is that studies, particularly those involving scale development, need to be well-designed in the first place. In order to improve the reliability and validity of the WBQ, better conceptualisation of the workplace behaviour construct is needed including writing personality-relevant items. The WBQ was developed by writing behavioural items associated with each personality profile. Conceptual linkages between the personality profiles and the behavioural statements used were defined through interpretations of the profiles. In future, these conceptual linkages should be defined based on a literature review of relevant, well-designed studies that have linked personality profiles with specific, job-oriented behaviours as Tett et al. (1991) suggested. Furthermore, although a team of psychologists assisted with the process, subjective judgement played a key role in deciding items whereas the use of more objective measures such critical incidents used in Day and Silverman's (1989) study, may have produced a more reliable and valid questionnaire.

Hypothesis 2

The second hypothesis concerning the presence of a significant negative correlation between EPQ-R Neuroticism and the EPQ-R Lie scale, and between EPQ-R Psychoticism and the EPQ-R Lie scale was also not supported. In the combined sample, the correlation was -0.20 for the Neuroticism/Lie correlation, and -0.19 for the Psychoticism/Lie correlation. In the officer cadet sample, these correlates were 0.26 and -0.01 respectively. In the officer sample, these correlates were -0.29 and -0.22 respectively. The small sample sizes and lack of an adequate control group prevent a conclusive statement on these findings, however, it can tentatively be said that faking did not occur in these samples, despite carrying out the questionnaire administration under selection conditions. Instead, it is more likely that the Lie scale measured a conformity or naivety factor as suggested by Eysenck and Eysenck (1991). In Eysenck and Eysenck's (1991) norming and validation study for the EPQ-R, the Neuroticism/Lie correlation was found to be -0.25 for males and -0.26 for females in a

sample of male and female students, teachers, and other subjects aged 16 to 70 years. The Psychoticism/Lie correlation was found to be -0.34 for males and -0.16 for females in this same sample. These calculations were based on the entire sample, and not divided according to age. These findings, then, were consistent with research that found faking did not affect the validity of personality questionnaires under selection conditions, as they were likely to measure self-deceptive enhancement rather than impression management (Tett et al., 1991; Hough et al., 1990; and Barrick & Mount, 1996). However, Eysenck and Eysenck's (1991) study was not carried out on applicants, who may be more motivated to fake. It is recommended that individual item scores on the EPQ-R and the GPP-I be retained in a readily-usable form so that reliability information can be calculated and a more rigorous design can be employed.

Hypothesis 3

The third hypothesis, which stated that officers who had served greater than three years should show similar personality profiles as a result of organisation socialisation to those who had served three years or less was not supported. One reason may be that three years was not a long enough time frame, however, again, the small sample sizes meant that an accurate interpretation was not possible. Using a larger sample may produce different results. It is unlikely that no organisation socialisation took place due to the strong military culture, however, researchers have had trouble separating out homogenous personality characteristics due to organisation socialisation or due to initial attraction to the organisation. For example, Schneider et al (1998) argued that homogeneity of personality characteristics within organisations were affected by three processes: First, attraction, where a person may be attracted to an organisation of which they perceive the employees to possess similar personality characteristics to themselves. Second, the organisation tends to select employees whom they believe have similar personality characteristics to themselves. Finally, a person is more likely to leave an organisation if they believe they do not possess similar personality characteristics, in other words, if they perceive that they do not fit in (Schneider et al., 1998). Those who leave the organisation serve to increase the homogeneity of personality characteristics of the organisation, as the people left behind all "fit in." Similarly, Black (1998) argued that selection in the NZ Police was based more on perceived organisation fit than the personality characteristics related to successful performance. He claimed that

selection in this way is not necessarily bad, as often good performance is a function of how well new recruits accept NZ Police organisational practices and values (Black, 1998).

Similarly, Judge and Cable (1997) argued that people were less likely to leave an organisation if they perceived that they fit in. They tested components of the model reported in Schneider's (1998) study and found that job seekers' personality characteristics influenced the type of organisation culture they preferred, and job seekers sought organisations that contained values that matched their own (Judge & Cable, 1997). Schneider et al. (1998) also postulated that organisation socialisation processes moderated the perceived fit of an employee in an organisation. That is, if effective organisation processes were in place, such that the employee comes to adopt the organisation's values and philosophy, then this leads to an increased likelihood that a person will remain with an organisation.

In the current study, participants may only be representative of those who fit in and have chosen to stay. Therefore, it is possible that there was no difference between those who had served longer and those who had served only a short time because only people with similar personalities and values to the military's were selected. A more rigorous design would entail comparing personality questionnaire data of unsuccessful and successful applicants such as that employed by Black (1998).

Hypothesis 4

The fourth hypothesis, addressing the gender and ethnicity of the immediate superior and its relationship to ratings of respondents with the same gender and ethnicity, could not be tested due to insufficient sample size, particularly of female and non-NZ pakeha participants. In light of Martin and Kirkcaldy's (1998) study where gender differences were found on the EPQ-R scales, it was expected that some gender differences should have occurred on the EPQ-R scales.

Hypothesis 5

The hypothesis that scores on the WBQ reflecting High Psychoticism, High Neuroticism, Low Emotional Stability, Low Ascendancy, and Low Cautiousness should not be endorsed highly if selection was effective showed that, in general, a minimal number of items measuring these scales were endorsed. This would suggest that selection was mainly effective. However, as a note of caution, due to the low reliability and, hence, validity of the WBQ scales, the endorsement rate of these items may not provide an accurate picture of whether these selection tools were effective. Only for Neuroticism in the officer cadet sample, where the reliability was 0.92, could any conclusion be made. 40% of the cadets displayed high scores on the first Neuroticism item and 20% scored highly on the second Neuroticism item. This would suggest that, amongst these cadets, High EPQ-R Neuroticism scorers were not effectively selected out. However, once again, caution must be exercised in this interpretation, as the sample size was only fifteen in the officer cadet sample.

Interactions

Finally, the chi square for the EPQ-R Extraversion/GPP-I Sociability interaction in the officer cadet sample reached a level of 0.07. Once again, although not significant, a larger sample size may have rendered the chi square figure higher, and hence, significant. The EPQ-R Extraversion/GPP-I Sociability interaction item reflects someone who actively avoids interacting with others, that is, the person will go out of their way to avoid working or being with others; they prefer to do things by themselves. For this to be significant in the present sample would be surprising, considering that officers are generally meant to be team players. However, seeing as this result occurred in the officer cadet sample, it is possible that the cadets may develop their team-player skills further during training and subsequent job placement.

Methodological Effects on the Data Obtained

Sample Size

The main problem with the research was the small sample size. Out of 250 officers and officer cadets (reduced from 500 already), only 54 responded, despite sending a reminder notice. This yielded a response rate of only 21.6%. The sample on which the data was based may have given questionable results, as it was not truly representative of the officer population as a whole. In addition, the small sample size meant that only limited statistical analyses could be performed. Aside from the APEC conference and East Timor crisis, what might have influenced the return rate? It is possible that those who did respond may have had different personality characteristics to those that did not respond. Furthemore, officers in certain positions may not have been so deeply involved with APEC, East Timor, or other tasks so they may have had more time to complete the questionnaires.

Another possibility was the length of the questionnaire. Although the WBQ questionnaire contained only 52 items (including 21 T-SD items later excluded from analysis), the NZ Army sent out additional questionnaires to be completed that assessed the relationship between their intelligence test scores, OSB gradings and subsequent workplace behaviour. The officers and officer cadets were informed that, although the questionnaire looked long, it would only take 15 - 20 minutes to complete. Potential participants may have decided it looked too long. One solution would be to send out the questionnaire separately to the Army Psychology Service questionnaires. Officers (and other army personnel), however, are often the target of the Army Psychology Service and other external agencies for research. This means they frequently receive questionnaires and the like to complete. The Army Psychology Service tries to limit the number of times an officer is asked to complete questionnaires so that, in the long run, they will be more inclined to participate in studies. Consequently, if they can send their material out at the same time as external research, this reduces the number of times an officer is asked to participate in a study. The situation requires a balance between obtaining a high response rate through a shorter questionnaire and keeping the workload of officers at a manageable level. In the current case the research would, more than likely, have benefited had the package not contained as many questionnaires.

Furthermore, more than one officer or officer cadet had the same immediate superior. This meant that the immediate superior would have had to fill in several questionnaires. For example, in the officer cadet sample, seven officer cadets had the same immediate superior, and five of the remaining cadets had the same immediate superior. One officer commented that their immediate superior had five questionnaires to fill in. Again, one way to remedy this problem is to reduce the length of the questionnaire so that the immediate superiors would be more inclined to participate.

A further problem regarding immediate superiors concerned the officer cadet sample. Officer cadets gave the questionnaire to another officer cadet senior in rank to them, instead of their OCS instructor. In future studies of this nature, the instructions need to be clarified so that the officer cadets know to whom they are to give the questionnaire. It is not certain what effect this had on the results, however, it is reasonable to suggest that the senior officer cadets would have had more chance to observe the participants' behaviour as they work closely together with them. Therefore, the data may be more useful for the officer cadet sample with a larger sample size.

Range Restriction

A major issue worth pointing out is the presence of range restriction. The current sample consisted of officers and officer cadets who had already been selected from the OSB, although the officer cadets still had to pass their training before being deemed an officer. This sample meant that the results were subject to range restriction errors. Participants with high Psychoticism scores, high Neuroticism, low Emotional Stability, or high Lie scores were unlikely to be selected. The current sample, then, should contain very few of these people. Looking at the frequency of responses for the WBQ Psychoticism scale in the combined sample, 80% of participants scored either a 1 or a 2 for the item "Fits in well with others" while only 9% scored at either a 4 or a 5 for this WBQ item (Does not fit in well with others). However, 48% of participants scored either a 1 or a 2 for the item "Rarely becomes anxious" for WBQ Neuroticism. In contrast, 26% of the sample scored a 4 for this WBQ item (Often becomes anxious). No subject scored a 5. For the second item measuring WBQ Neuroticism, 61% of subjects scored either a 4 or a 5 for the item "Performs effectively under

pressure" while only 9% of subjects scored a 2 (Does not perform effectively under pressure). No subject scored a 1. Finally, 69% of subjects scored either a 4 or a 5 for the WBQ Lie scale item "Portrays self in a realistic manner and 11.1% scored a 2 for its reverse (Does not portray self in a realistic manner). For the second WBQ Lie scale item, 78% of subjects scored a 4 or a 5 for the item "Sincere in their support of colleagues and superiors" while 1.9% scored either a 1 or a 2 for its reverse (Overtly supports colleagues and superiors to earn favour). The standard deviations provided in the manuals compared with the standard deviations of the EPQ-R and GPP-I scores of the current sample indicate the presence of range restriction, particularly for the Psychoticism, Ascendancy, and Emotional Stability scales. Again, future research addressing differences between unsuccessful and successful applicant data would be helpful in highlighting range restriction issues.

Statistics

Some minor methodological details are worth mentioning. First, the initial data provided by the NZ Army of the EPQ-R and GPP-I scores were already entered into a database. While it was assumed that these scores were accurate, they could not be checked by the researcher. Second, as the NZ Army were unable to provide individual item scores for the EPQ-R and the GPP-I, the alpha coefficients for these scales could not be computed and compared with the population alpha coefficients given in the manuals. Although this has not affected the data in the present study, had the sample size and alpha coefficients of the WBQ been higher, the corrections for attenuation and range restriction require the use of these alpha coefficients. Finally, the intercorrelates of the EPQ-R and GPP-I scales were compared with those presented in the manuals. Ideally, they should be compared with intercorrelates in military normative studies.

Summary and Conclusions

The small sample size and poor reliability of the WBQ has prevented any firm conclusions regarding the current research. Therefore, whether personality questionnaires are useful selection tools, particularly in a military selection setting, needs to be researched further. In themselves, the EPQ-R and the GPP-I show good reliability and validity; the

GPP-I has also been well-validated in selection and military settings. However, the reliability of the EPQ-R and the GPP-I could not be determined for the current sample as individual item scores were not provided. While there were no available personality studies on NZ military samples, there were a few studies using NZ Police subjects and personality questionnaires in selection. As the NZ Police organisation has a similar structure to the NZ Army, and the two often work together, future studies could address similarities and differences between the NZ Police and NZ Army. Since the NZ Army use the EPQ-R primarily as a screening device rather than a predictor of future performance, it is still possible that the EPQ-R will remain an effective selection tool for the NZ Army's purposes. However, the EPQ-R and GPP-I scales should be conceptually-linked to specific job performance criteria developed through a careful job analysis. The questionnaires can then be validated more rigorously in the military selection setting. Careful attention also needs to be paid to faking, as applicants are more likely to distort responses than the general population. To assess the level of faking, future research should use a control group who are not officer applicants and compare the data with both unsuccessful and successful applicant data. Finally, personality questionnaires are but one part of officer selection, therefore, in the end, the validity of the decisions made from the officer selection process will be moderated by the weight given to each component of the assessment centre. Future research that has addressed the shortcomings of the methodology in the current research should be more successful in identifying the usefulness of the EPQ-R and GPP-I questionnaires for officer selection.

Recommendations

From the basis of the results and the methodology problems inherent in the design, the following recommendations are proposed to improve future research:

1. The database should include individual item scores so that reliability estimates can be calculated.

2. The NZ Army should ensure that personality-relevant criteria based on sound research demonstrating conceptual linkages of personality traits to performance criteria of the officer role are included in the job analysis of the officer role currently being undertaken.

3. The NZ Army should conduct research comparing differences between unsuccessful and successful applicants', and a control group's EPQ-R and GPP-I personality questionnaire data to help identify the level of faking and range restriction.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44(1), 1 - 26.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. Journal of Applied Psychology, 81(3), 261 - 272.

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), 478 - 494.

Black, J. (1997, August). Personality testing and police selection: Utility of the Big Five. Paper presented at the annual conference of the New Zealand Psychological Society, Palmerston North, NZ.

Black, J. (1998, September). Does a job applicant's personality contribute to a successful job application? Paper presented at the annual conference of the New Zealand Psychological Society, Wellington, NZ.

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. Journal of Applied Psychology, 67(1), 3 - 9.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.). Handbook of Industrial and Organizational Psychology (pp. 271 - 325). Palo Alto, CA: Consulting Psychologists Press.

Braun, J. R. (1963). Fakability of the Gordon Personal Inventory: Replication and extension. The Journal of Psychology, 55, 441 - 444.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309 - 319.

Corulla, W. J. (1987). A psychometric investigation of the Eysenck Personality Questionnaire (Revised) and its relationship to the I.7 Impulsiveness Questionnaire. Personality and Individual Differences, 8(5), 651 - 658.

Costa, Jr. P. T., & McCrae, R. R. (1997). Stability and change in personality assessment: The Revised NEO Personality Inventory in the year 2000. Journal of Personality Assessment, 68(1), 86 - 94.

Davies, M. F., French, C. C., & Keogh, E. (1998). Self-deceptive enhancement and impression management correlates of EPQ-R dimensions. Journal of Psychology, 132(4), 401 - 406.

Day, D. V., & Silverman, S. B. (1989). Personality and job performance: Evidence of incremental validity. Personnel Psychology, 42, 25 - 36.

Eatwell, J. (1998, August). Debate on validity is joined. Human Resources, 18 - 19.

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.

Eysenck, S. B. G., & Eysenck, H. J. (1991). Manual of the Eysenck Personality Scales. London: Hodder & Stoughton.

Friedman, H. S. (1983). On shutting one's eyes to face validity. Psychological Bulletin, 94(1), 185 - 187.

Friedman, H. S. (1984). Eysenck Personality Questionnaire. Test Critiques, 1, 158 - 161.

Gillis, J. S., & Lee, D. C. (1979). Relationships between the 16PF, GPP, and GPI. Educational and Psychological Measurement, 39, 7 - 12.

Goeters, K-M., Timmerman, B., & Maschke, P. (1993). The construction of personality questionnaires for selection of aviation personnel. The International Journal of Aviation Psychology, 3(2), 123 - 141.

Goodstein, L. D. & Lanyon, R. I. (1999). Applications of personality assessment to the workplace: A review. Journal of Business and Psychology, 13(3), 291 - 322.

Gordon, L. V. (1993). The Gordon Personal Profile-Inventory.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. Personnel Psychology, 18, 135 - 164.

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

Gulliksen, H. (1987). Theory of Mental Tests. New York: John Wiley & Sons.

Hershberger, S. L. (1999). Introduction to personality measurement. In S. E. Embretson & S. L. Hershberger. The New Rules of Measurement: What Every Psychologist and Educator Should Know (pp. 153 - 158). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hess, A. K. (1998). Gordon Personal Profile-Inventory (Revised). In J. C. Impara & B. S. Plake (Eds.). The Thirteenth Mental Measurements Yearbook. Buros Institute.

Hogan, R. T. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.). Handbook of Industrial and Organizational Psychology (pp. 873 - 917). Palo Alto, CA: Consulting Psychologists Press.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. Journal of Organizational Behavior, 17, 627 - 637.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions. American Psychologist, 51(5), 469 - 477.

Hough, L. M. (1992). The "Big Five" personality variables-construct confusion: Description versus prediction. Human Performance, 5(1&2), 139 - 155.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology, 75(5), 581 - 595.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating Research Findings Across Studies. Beverly Hills: Sage Publications.

Jackson, D. N. (1971). The dynamics of structured personality tests. Psychological Review, 78(3), 229 - 248.

Jackson, D. N., & Paunonen, S. V. (1980). Personality structure and assessment. Annual Review of Psychology, 31, 503 - 551.

Judge, T. A., & Cable, D. M. (1997). Applicant personality, organizational culture, and organization attraction. Personnel Psychology, 50, 359 - 394.

Kline, P. (1993). The Handbook of Psychological Testing. London: Routledge.

Knapp, R. R., & Fitzgerald, O. R. (1973). Comparative validity of the logically-developed versus "purified" research scales for the Personal Orientation Inventory. Educational and Psychological Measurement, 33, 971 - 976.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. Psychological Bulletin, 107(1), 34 - 47.

Mahar, D., Cologon, J., & Duck, J. (1995). Response strategies when faking personality questionnaires in a vocational selection setting. Personality and Individual Differences, 18(5), 605 - 609.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43(2), 335 - 354.

Martin, T., & Kirkcaldy, B. (1998). Gender differences on the EPQ-R and attitudes to work. Personality and Individual Differences, 24(1), 1 - 5.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Brown (Eds.), Test Validity (pp. 19 - 32). Hillsdale, NJ: Erlbaum.

Murphy, K. R., & Davidshofer, C. O. (1998). Psychological Testing: Principles and Applications. Upper Saddle River, NJ: Prentice-Hall Inc.

Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. Psychological Bulletin, 92(3), 572 - 580.

Orton, R. E. (1987). The foundations of construct validity: Towards an update. Journal of Research and Development in Education, 21(1), 23 - 35.

Ozer, D. J., & Reise, S. P. (1994). Personality assessment. Annual Review of Psychology, 45, 357 - 388.

Parker, R. M. (1993). Threats to the validity of research. Rehabilitation and Counselling Bulletin, 36(3), 130 - 138.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A Angleitner & J. J. Wiggins (Eds.). Personality Assessment Via Questionnaires: Current Issues in Theory and Measurement. New York: Springer-Verlag.

Pearson, P. R. (1989). A comparison of the Psychoticism scale of the EPQ and the EPQ-R. The Journal of Psychology, 122(6), 623 - 624.

Perkins, A. (1998). Personality and leadership. Unpublished manuscript, University of London at New Cross.

Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. Educational and Psychological Measurement, 56(2), 209 - 214.

Robertson, I. T. (1993). Personality assessment and personnel selection. European Review of Applied Psychology, 43(3), 187 - 194.

Robertson, I. T. (1994). Personality and personnel selection. In C. L. Cooper & Rousseau (Eds.), Trends in Organizational Behavior (pp. 75 - 89). Chichester: John Wiley & Sons Ltd.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. Journal of Applied Psychology, 83(4), 634 - 644.

Schippman, J. S., & Prien, E. P. (1989). An assessment of the contributions of general mental ability and personality characteristics to management success. Journal of Business and Psychology, 3(4), 423 - 437.

Schneider, B., Smith, D. B., Taylor, S., & Fleenor, J. (1998). Personality and organizations: A test of the homogeneity of personality hypothesis. Journal of Applied Psychology, 83(3), 462 - 470.

Schneider, R. J., Hough, L. M., & Dunnette, M. D. (1996). Broadsided by broad traits: How to sink science in five dimensions or less. Journal of Organizational Behavior, 17, 639 - 655.

Schwab, D. P. (1971). Issues in response distortion studies of personality inventories: A critique and replicated study. Personnel Psychology, 24, 637 - 647.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. Personnel Psychology, 44, 703 - 742.

Tokar, D. M., Fischer, A. R., & Subich, L. M. (1998). Personality and vocational behavior: A selective review of the literature, 1993 - 1997. Journal of Vocational Behavior, 53, 115 - 153.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? Journal of Applied Psychology, 82(2), 247 - 252.

Wilson, D. J., & Doolabh, A. (1992). Reliability, factorial validity and equivalence of several forms of the Eysenck Personality Inventory/Questionnaire in Zimbabwe. Personality and Individual Differences, 13(6), 637 - 643.

APPENDIX A: WORKPLACE BEHAVIOUR QUESTIONNAIRE

CODE:

We would like some background information on you.

SECTION 1 (*To be completed by the Officer or Officer Cadet*)

What is your age?_____

What is your gender? **Female**.....**1**

Male.....**2**

What is your present rank?_____

What is your corps?_____

What is your ethnic group? **NZ Maori**.....**1** **Asian**.....**4**

NZ European/Pakeha.....**2** **Other (Please Specify)**.....**5**

.....

Pacific Islander.....**3**

Total number of years served (RF)_____

Total number of years served (TF)_____

What was your highest formal educational qualification when you first joined the Army?

No school qualification.....1

School Certificate.....2

Sixth Form Certificate.....3

University Entrance (or equivalent).....4

Bursary or Scholarship.....5

Trade or Professional certificate or diploma.....6

Part-Degree or Diploma.....7

Bachelor Degree.....8

Postgraduate qualification.....9

Now please hand this questionnaire to your immediate superior to complete

We would like some background information on you.

SECTION 2 (***To be completed by immediate superior***)

What is your age?_____

What is your gender? **Female**.....**1**

Male.....**2**

What is your present rank?_____

What is your ethnic group? **NZ Maori**.....**1** **Asian**.....**4**

NZ European/Pakeha.....**2** **Other (Please Specify)**.....**5**

Pacific Islander.....**3** .....

Total number of years served (RF)_____

Total number of years served (TF)_____

SECTION 3 (***To be completed by immediate superior***)

For each of the following items please circle a number on the scale between 1 and 5 that you feel best describes the officer/officer cadet.

Fits in well with others	1	2	3	4	5	Does not fit in well with others
Reserved	1	2	3	4	5	Outgoing and sociable
Prefers working by self	1	2	3	4	5	Prefers working with others
Rarely becomes anxious	1	2	3	4	5	Often becomes anxious
Does not perform effectively under pressure	1	2	3	4	5	Performs effectively under pressure
Tends to portray self in an overly positive manner	1	2	3	4	5	Portrays self in a realistic manner
Overtly supports colleagues and superiors to earn favour	1	2	3	4	5	Sincere in their support of colleagues and superiors
Alcohol has not adversely affected performance	1	2	3	4	5	Alcohol has adversely affected performance
Often voices opinions confidently in front of colleagues and superiors	1	2	3	4	5	Rarely voices opinions confidently in front of colleagues and superiors
Takes a passive role within a group	1	2	3	4	5	Takes an active role within a group
Rarely works to the best of their ability	1	2	3	4	5	Often works to the best of their ability
Rarely expresses emotion	1	2	3	4	5	Freely expresses emotion

STAFF IN CONFIDENCE

Quickly returns to normal functioning after a stressful experience	1	2	3	4	5	Takes longer than most to return to normal functioning after a stressful experience
Overly concerned with gaining others' approval	1	2	3	4	5	Not at all concerned with gaining others' approval
Works well as part of a team	1	2	3	4	5	Does not work well as part of a team
Works well independently	1	2	3	4	5	Does not work well independently
Lacks confidence and assurance about themselves	1	2	3	4	5	Is confident and assured about themselves
Impulsive	1	2	3	4	5	Overly-cautious
Often generates creative solutions to challenging problems encountered at work	1	2	3	4	5	Rarely generates creative solutions to challenging problems encountered at work
Rarely seeks intellectual stimulation	1	2	3	4	5	Often seeks intellectual stimulation
Patient and tolerant of others	1	2	3	4	5	Impatient and intolerant of others
Distrustful of others	1	2	3	4	5	Trusting of others
"Rubs people up" the wrong way	1	2	3	4	5	Does not "rub people up" the wrong way
Lacks energy and drive	1	2	3	4	5	Has lots of energy and drive
Warm, sympathetic, understanding, kind and generous	1	2	3	4	5	Is not warm, sympathetic, understanding, kind or generous
Cheerful, friendly, and easy to get along with	1	2	3	4	5	Is not cheerful, friendly, or easy to get along with
Is considerate and kind to others	1	2	3	4	5	Is not considerate or kind to others
Is not warm and sensitive to the feelings of others	1	2	3	4	5	Is warm and sensitive to the feelings of others
Willing to help others when they are busy	1	2	3	4	5	Not willing to help others when they are busy
Not efficient, responsible or dependable	1	2	3	4	5	Efficient, responsible, and dependable
Persists in completing tasks to the best of their ability	1	2	3	4	5	Does not persist in completing tasks to the best of their ability
Organised and tidy	1	2	3	4	5	Not organised or tidy
Is not talkative or comfortable conversing with others	1	2	3	4	5	Talkative and comfortable conversing with others
Does not show much interest at social functions	1	2	3	4	5	Is the life and soul of the party

STAFF IN CONFIDENCE

STAFF IN CONFIDENCE

Not an active participant in social activities	1	2	3	4	5	Active participant in social activities
Does not enjoy intellectual discussion or working on complex problems that involve intellectual challenge	1	2	3	4	5	Enjoys intellectual discussions and working on complex problems that involve intellectual challenge
Shows an interest in theoretical analysis	1	2	3	4	5	Does not show an interest in theoretical analysis
Often in deep thought	1	2	3	4	5	Not often in deep thought
Demonstrates an interest in various cultural activities such as classical music, poetry, or visiting museums	1	2	3	4	5	Does not demonstrate an interest in various cultural activities such as classical music, poetry, or visiting museums
Displays adverse reactions under pressure and exhibits swings in emotions	1	2	3	4	5	Does not display adverse reactions under pressure or exhibit swings in emotions
Does not tend to worry more than most people	1	2	3	4	5	Tends to worry more than most people
Does not get upset easily, lose temper with others, or respond badly to criticism	1	2	3	4	5	Gets upset easily, loses temper with others, and responds badly to criticism
Accepting of others' achievements and opportunities	1	2	3	4	5	Not accepting of others' achievements and opportunities
Not able to come up with workable solutions to problems	1	2	3	4	5	Able to come up with workable solutions to problems
Not able to learn new information quickly and accurately	1	2	3	4	5	Able to learn new information quickly and accurately
Able to apply a logical thought process	1	2	3	4	5	Not able to apply a logical thought process
Not able to see relationships between different concepts	1	2	3	4	5	Able to see relationships between different concepts
Able to apply learned skills in unfamiliar situations	1	2	3	4	5	Not able to apply learned skills in unfamiliar situations
Able to think clearly and accurately with unfamiliar information and/or in unfamiliar situations	1	2	3	4	5	Not able to think clearly and accurately with unfamiliar information and/or in unfamiliar situations

STAFF IN CONFIDENCE

APPENDIX B: ITEMS OF EACH WBQ SCALE

DESCRIPTION OF ITEMS FOR EACH WBQ SCALE

EPQ-R

PSYCHOTICISM:	Items 1, 22, 23
EXTRAVERSION:	Items 2, 3
NEUROTICISM:	Items 4, 5
LIE:	Items 6, 7
ADDICTION:	Item 8

GPP-I

ASCENDANCY:	Items 9, 10
RESPONSIBILITY:	Item 11
EMOTIONAL STABILITY:	Items 12, 13
SOCIABILITY:	Items 14, 15, 16
SELF-ESTEEM:	Item 17
CAUTIOUSNESS:	Item 18
ORIGINAL THINKING:	Items 19, 20
PERSONAL RELATIONS:	Items 21, 22, 23
VIGOR:	Item 24

APPENDIX C: INFORMATION SHEETS FOR JUNIOR OFFICERS, OFFICER CADETS, AND IMMEDIATE SUPERIORS

EVALUATION OF PERSONALITY QUESTIONNAIRES AND OTHER OFFICER SELECTION BOARD INFORMATION

INFORMATION SHEET FOR OFFICERS AND OFFICER CADETS

Dear

My name is Charlotte Bowden and I am a postgraduate student in the School of Psychology at Massey University. As part of the requirements for my Masters of Science, I am conducting a research project. The research has been approved by the Massey University Ethics Committee and the appropriate authorities within the NZ Army. It is being supervised by Associate Professor Douglas Paton, of the School of Psychology at Massey University and is under the direction of Major Kate Mirfin, Senior Psychologist (Army) and Major Clare Bennett, Assistant Director Psychological Research.

The aim of the research is to assess the usefulness of personality questionnaires in officer selection. Reviewing these questionnaires is necessary as they are not used just for officer selection, but for many other purposes in the NZ Army. Because of this, the questionnaires need to be valid so that accurate decisions can be made. In addition to items which reflect what the personality questionnaires purport to measure, some extra items have been included so that the Army Psychology Service can look at how well the B90 test of general reasoning ability and the criteria currently used at Officer Selection Boards predict associated behaviour. The results from this research will be used to refine officer selection procedures as part of an ongoing officer selection review. Your participation in this study would contribute towards this.

In order to carry out this research, the following information will be needed:

a) data from your Officer Selection Board, including your personality questionnaire results (these are currently held on NZ Army Psychology Service files and access is normally limited to NZ Army psychologists).

b) behavioural data, which will be obtained with your consent from your immediate superior, will be obtained through the enclosed questionnaire specifically designed for the purposes of the research.

I have enclosed the performance questionnaire together with an information sheet for you to give to your immediate superior to complete regarding behaviour related to personality questionnaire scales, other Officer Selection Board information. You are welcome to look at the questions on the questionnaire if you wish. The questionnaire results are to be used **solely for the purposes of the research.** That is, for the purposes of my thesis and also for the Army Psychology Service to refine officer selection procedures. The questions will, in no way, affect the formal NZDF Performance Appraisal process that you undergo. All performance questionnaires will be destroyed following completion of the research.

All information provided by you will be treated in the strictest confidence. You will be identified only by a code number to protect your anonymity and privacy during the course of the research. Data will be published in the form of aggregate data and summaries and no individual data will be published in these summaries. The final thesis may be published in a peer-reviewed journal, and the results will contribute to NZDF efforts to improve selection procedures.

PRIVACY ACT (1993)

The research will also strictly adhere to the Privacy Act (1993) principles. Under this Act, you have the right to:

a. contact the researcher or her supervisor at any time during the research to discuss any aspects of the study

b. decline to answer any question

c. withdraw from the study at any time. If you choose to do so, any information you have already provided will be removed from the study and destroyed

d. provide information on the understanding that all responses will be held in complete confidence by the researchers, to be used only for the purposes outlined in the information sheet and consent form. It will not be possible to identify individuals in any reports of the results

e. receive information about the results of the study on its completion.

The research will be carried out strictly in accordance with the standards set by the Massey University Ethics Committee, the Ethical Guidelines for Personnel Research in the NZDF as outlined in DFO 2/1997, and the NZ Psychological Society Code of Ethics.

If you wish to obtain a copy of a summary of the results, please tick the appropriate check box on the consent form.

What am I being asked to do?

1. Complete the yellow pages of questionnaire (please note that this includes the consent form) if you would like to participate.

2. Pass the questionnaire, letter to immediate superiors and return envelope, to your immediate superior.

3. Encourage your superior to return the questionnaire in the return envelope by [redacted] (However, the questionnaires will be accepted until mid-August).

I can be contacted by either writing to the address given at the top of the page or by phoning [redacted] Alternatively, you can contact Associate Professor Douglas Paton at Massey University on 06 350 5799 x 2064, or Major Kate Mirfin, Senior Psychologist (Army) on DtelN [redacted] to discuss any aspect of the research.

Thank you for your time, I look forward to your participation.

Yours sincerely,

Charlotte Bowden.

EVALUATION OF PERSONALITY QUESTIONNAIRES AND OTHER OFFICER SELECTION BOARD INFORMATION

INFORMATION SHEET FOR IMMEDIATE SUPERIORS

Dear

My name is Charlotte Bowden and I am a postgraduate student in the School of Psychology at Massey University. As part of the requirements for my Master of Science, I am conducting a research project. The research has been approved by the Massey University Ethics Committee and the appropriate authorities within the NZ Army. It is being supervised by Associate Professor Douglas Paton, of the School of Psychology at Massey University and is under the direction of Major Kate Mirfin, Senior Psychologist (Army) and Major Clare Bennett, Assistant Director Psychological Research.

The aim of the research is to assess the usefulness of personality questionnaires in officer selection. Reviewing these questionnaires is necessary as they are not used just for officer selection, but for many other purposes in the NZ Army. Because of this, the questionnaires need to be valid so that accurate decisions can be made. In addition to items which reflect what the personality questionnaires purport to measure, some extra items have been included so that the Army Psychology Service can look at how well the B90 test of general reasoning ability and the criteria currently used at Officer Selection Boards predict associated behaviour. The results from this research will be used to refine officer selection procedures as part of an ongoing officer selection review. Your participation in this study would contribute towards this.

What am I being asked to do?

1. Complete the white pages of the enclosed questionnaire. (You can check that the officer/officer cadet has consented for you to do this from the consent form on the front). Please complete the questionnaire only if you feel you can make an accurate assessment (eg if you have observed the officer/officer cadet regularly for at least four weeks). While the questionnaire may look long it will only take about 20 minutes to complete.

2. Return the questionnaire (in the return envelope) by [redacted] (However, the questionnaires will be accepted until mid-August).

The questionnaire results are to be used **solely for the purposes of the research.** That is, for the purposes of my thesis and also for the Army Psychology Service to refine officer selection procedures. The questionnaire data obtained in this study will, in no way, affect the formal NZDF Performance Appraisal process for the officer concerned. All performance questionnaires will be destroyed following completion of the research.

All information provided by you will be treated in the strictest confidence. Data will be published in the form of aggregate data and summaries and no individual data will be published in these summaries. The final thesis may be published in a peer-reviewed journal, and the results will contribute to NZDF efforts to improve selection procedures.

PRIVACY ACT (1993)

The research will also strictly adhere to the Privacy Act (1993) principles. Under this Act, you have the right to:

a. contact the researcher or her supervisor at any time during the research to discuss any aspects of the study

b. decline to answer any question

c. withdraw from the study at any time. If you choose to do so, any information you have already provided will be removed from the study and destroyed

d. provide information on the understanding that all responses will be held in complete confidence by the researchers, to be used only for the purposes outlined in the information sheet and consent form. It will not be possible to identify individuals in any reports of the results

e. receive information about the results of the study on its completion.

The research will be carried out strictly in accordance with the standards set by the Massey University Ethics Committee, the Ethical Guidelines for Personnel Research in the NZDF as outlined in DFO 2/1997, and the NZ Psychological Society Code of Ethics.

If you wish to obtain a copy of a summary of the results, please contact me according to the details set out below.

I can be contacted by either writing to the address given at the top of the page or by phoning [redacted] Alternatively, you can contact Associate Professor Douglas Paton at Massey University on 06 350 5799 x 2064, or Major Kate Mirfin, Senior Psychologist (Army) on [redacted] to discuss any aspect of the research.

Returning the questionnaire constitutes your informed consent.

Thank you for your time, I look forward to your participation.

Yours sincerely,

Charlotte Bowden.

APPENDIX D: CONSENT FORM

EVALUATION OF PERSONALITY QUESTIONNAIRES AND OTHER OFFICER SELECTION BOARD INFORMATION

INFORMED CONSENT FORM (OFFICERS AND OFFICER CADETS)

1. I have read and understood the information sheet explaining the research and I understand my rights under the Privacy Act (1993) as listed on the information sheet.

2. I agree to information from my Officer Selection Board, including personality questionnaire data already collected by the NZ Army Psychology Service, being provided to the researcher on the understanding that it is viewed only by the researcher and NZ Army Psychology Service personnel and that it will remain strictly "in confidence."

3. I understand that the behavioural information provided by my immediate superior is to be used strictly for the purposes of the research and will be destroyed on completion of the research.

4. In signing this form, I agree to participate in the study under the conditions set out in the information sheet.

5. I would like a summary of the results on completion of the research

YES ☐ **NO** ☐

Signed:

Name:

Date:

Address so we know where to send the summary to: