Running head: CFAT Bias

Determining if the Canadian Forces Aptitude Test is Biased Against Canadian Aboriginal Peoples

Thesis, submitted in partial fulfilment of the Requirement for the Degree of Master of Science in Applied Psychology (Industrial/Organizational)

Michael A. Vanderpool

Saint Mary's University

Approved

Approved:

Approved:

Date: March 2003

Your file Votre référence
ISBN: 0-612-86577-0
Our file Notre référence
ISBN: 0-612-86577-0

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgement

This thesis is the result of significant effort by many individuals. I would like to thank Dan Highway who was instrumental in obtaining permission from the various Band Chiefs and Tribal Councils to conduct this study with their people. Dan was also responsible for promoting the study among the First Nations communities and overseeing the logistical side of the administration.

I am also grateful for the support that I received from the Director of Human Resources Research and Evaluation. Special thanks go to Lieutenant Colonel Boswell for his support and insightful comments on the final draft.

I would also like to thank my supervisor, Dr. Victor Catano, who provided context, advice and many stimulating and thought provoking discussions.

As well, I would also like to thank my family for their patience and understanding. To my wife Monica: I know there were times when you didn't know what to do with me, but just being there was enough. To my son Zachary: Although you are too young to understand, you provided me with the motivation and inspiration that I needed to complete this vast undertaking.

Most importantly, I would like to thank the many Aboriginal Peoples and members of the Canadian Forces who took the time and effort to complete the various tests. As in any major research project, it could not have been accomplished without their willing participation.

Determining if the Canadian Forces Aptitude Test is
Biased Against Canadian Aboriginal Peoples

Michael A. Vanderpool

March 2003

Abstract

The primary purpose of this study was to determine if any items on the Canadian Forces Aptitude Test (CFAT) possessed any degree of bias on the basis of Aboriginal status. A secondary goal was to investigate the possibility of using another well-established measure of cognitive ability to select Aboriginal Peoples for employment in the Canadian Forces (CF). To achieve these ends, the CFAT, Wonderlic Personnel Test (WPT), Raven's Standard Progressive Matrices (SPM) and Mill Hill Vocabulary Scale (MHV) were administered to Aboriginal Peoples (n = 101) living in special access and remote communities. The same four tests were also administered to a reference group composed of recruits (n = 108) undergoing basic training in the CF.

Aboriginal Peoples scored significantly lower than the recruits on all verbal measures of cognitive ability. However, both groups performed similarly on both nonverbal measures of cognitive ability, the CFAT Spatial Ability (SA) scale and SPM. Differential Item Functioning (DIF) analysis, using logistical regression, detected a few items from the CFAT, SPM and MHV that displayed DIF, but none from the WPT. The two CFAT items that displayed DIF came from the Verbal Skill scale.

The "four-fifths" rule was used to determine if the CFAT had an adverse impact on Aboriginal Peoples. The CFAT scores of the Aboriginal participants were compared against the scores of Anglophone Non-Commissioned Member applicants. Selection ratios for both groups, based on CFAT scores, indicated that the CFAT did have an adverse impact on Aboriginal Peoples for all military occupational families; adverse impact was more severe for the Administration, Operator, Technical and Mechanical occupations. Selection of Aboriginal Peoples into the CF, based on the SPM and the CFAT Spatial Ability scale, coupled with English or French language training, may offer an alternative procedure that will increase the number of qualified Aboriginal Peoples accepted into the CF.

Determining if the Canadian Forces Aptitude Test is Biased Against Canadian Aboriginal Peoples

Introduction

When selecting job applicants on the basis of test scores, it is critical to avoid bias that may unfairly influence applicants' scores (Hambleton & Rodgers, 1995). In order to meet this requirement, selection tests must be fair to all applicants and not be biased against a segment of the applicant population (Zumbo, 1999). Bias is the presence of some characteristic of an item that results in differential performance for two individuals of the same ability but from different ethnic, sex, cultural, or religious groups (Hambleton & Rodgers, 1994). Test bias can result in systematic errors that distort inferences made in selection or placement.

Testing the learning ability of potential recruits is a cornerstone of the Canadian Forces (CF) selection process. The Canadian Forces Aptitude Test (CFAT) is a cognitive ability measure used by the CF to screen and place suitable applicants. The 60-item test is comprised of two 15-item subscales (Verbal Skills and Spatial Ability) and one 30-item subscale (Problem Solving). The Verbal Skills (VS) scale assesses a candidate's ability to understand the meanings and uses of words. The Spatial Ability (SA) scale measures a candidate's ability to mentally manipulate a variety of complex three-dimensional figures. Because no reading is required to complete the SA, it is essentially a nonverbal measure. Finally, the Problem Solving (PS) scale measures a candidate's ability to use mathematical and deduction skills in solving number and word problems. The CFAT is a timed test arranged in ascending order of difficulty.

In order to enrol in the CF, each applicant must write the CFAT and achieve a predetermined minimum cut off score set at the tenth percentile. CFAT scores are also used to determine which military occupations recruits are suitable for. Different occupations require different cut-off scores (see Table 1).

Table 1

CFAT Cut-off Score for the Various Military Families

Military Family	Minimum VSPS	Minimum PS	Minimum CFAT Total
Steward	10 %ile		
Cook	20 %ile		
General Military		20 %ile	
Administration	40 %ile		
RMS Clerk	40 %ile	30 %ile	
Mechanical			40 %ile
Operator			40 %ile
Technical			40 %ile

Note. VSPS = [VS %ile + PS %ile] / 2

The CFAT is not biased against gender or language (Zumbo & Hubley, 1998a). However, there are some concerns that the CFAT may be biased against Aboriginal Peoples (Boswell, R. A., personal communication, June 22, 2001).

The controversy concerning cognitive ability tests and Aboriginal Peoples is not unique to the CF. Although the literature relating to Aboriginal Peoples and test bias is

meagre (Osborne, 1985), there is evidence that many of the cognitive measures being used for selection are biased against Aboriginal Peoples (Brescia & Fortune, 1989; Darou, 1992; Kleinfield & Nelson; 1991; McShane & Plas; 1984). Aboriginal students tend to score 20 points lower than Caucasian students on verbal tests of cognitive ability (McShane & Plas, 1984). Likewise, Aboriginal Peoples have been found to have lower mean test scores, often as much as one standard deviation, in comparison to the majority population (McShane & Berry, 1988). Selection tests consistently underestimate the ability of Aboriginal Peoples. Consequently, Aboriginal Peoples may be denied opportunities or may be relegated to low paying jobs (Brescia & Fortune, 1989).

In an effort to overcome the bias of verbal cognitive ability tests, some researchers have proposed the use of nonverbal tests. Aboriginal students tend to score about five points higher on nonverbal tests of cognitive ability than Caucasian students (McShane & Plas, 1984). However, the use of nonverbal cognitive ability test in selection has not been without criticism. Parmar (1989) reviewed the literature on the relationship between cultural bias and tests of nonverbal intelligence and found inconsistent results.

Unfortunately, despite the consensus that cognitive ability tests are biased against Aboriginal Peoples, there is no general agreement or proposal on methodology for treatment of this problem (Schwartz, 1999). The purpose of this study is to determine if any items on the CFAT are biased against Aboriginal Peoples. If the CFAT is biased against Aboriginal Peoples, the second purpose of this study is to determine if another well established verbal or nonverbal test of cognitive ability could be used in lieu of the CFAT.

General Cognitive Ability

General cognitive ability tests have been used in personnel selection for more than 80 years (Outtz, 2002). Testing for general cognitive ability is the best way to classify a large number of applicants in terms of probable success in job performance (Schmidt & Hunter, 2000). General cognitive ability is the ability to grasp and reason correctly with concepts and solve problems (Schmidt & Hunter, 2000). To put it simply, general cognitive ability is the ability to learn (Hunter, 1986; Hunter & Schmidt, 1996). People who are more intelligent learn more job knowledge and learn it faster. Conversely, people cannot perform a job well if they don't know how to do it. Even jobs that most people would consider simple such as truck driver or machine operator require considerable job knowledge (Schmidt & Hunter, 2000). General cognitive ability is also related to people's ability to adapt to novel, complex or changing situations (Gottfredson, 1986).

General cognitive ability is probably the best measured and most studied human trait (Gottfredson, 2002). It is one of the best predictors of trainability and job performance (Jensen, 1986; Ree & Carreta, 1997; Ree & Earles, 1992; Schmidt & Hunter, 1998). When job performance is measured objectively using carefully constructed work sample tests, the correlation with general cognitive ability is .70 and when performance is measured using supervisor ratings, the correlation with general cognitive ability is over .60 for all jobs (Schmidt & Hunter, 1998). Regardless of the job, general cognitive ability predicts amount learned in training with validity of about .56 (Hunter & Hunter, 1984).

Aboriginal Peoples and Cognitive Ability

In a review of research on cognitive ability among Aboriginal Peoples, Osborne (1985) found that over a period of 10 years, only 28 studies have been conducted (16 in the U.S., 12 in Canada). Despite the limited number of studies on cognitive ability among Aboriginal Peoples (Osborne, 1985), some important discoveries have emerged from that research. Inuits appear to have an uncanny ability to comprehend rotated visual configurations (Klienfeldt, 1973) and perform well on nonverbal measures of spatial ability and inductive reasoning (McArthur, 1973). Well-developed visual perception skills have been found among other Aboriginal groups. Aboriginal Peoples across North America tend to perform well on visual and spatial components of cognitive ability tests (McShane & Berry, 1988). In contrast, Aboriginal Peoples tend to perform poorly on verbal measures (McShane & Berry, 1988).

There is no conclusive explanation for the disparity of Aboriginal Peoples performance on verbal and nonverbal measures of cognitive ability. However some researchers have suggested that testing in one's secondary language may contribute to these results (Krywanuik & Das, 1976; Zarske & Moore, 1982). For many Aboriginal Peoples, English is a second language. Sattler (1982) notes that language related factors often confound attempts to accurately measure the cognitive ability of people from various cultures. Even if individuals have an adequate level of English reading and writing skills, testing in their first language is preferable. Li (1999) provides evidence that, even when two culturally different groups are using the same language at the same level of proficiency, inter-cultural communications conveys two-thirds less information than that of intra-cultural communication.

Nonverbal Measures of Cognitive Ability

Verbal measures of cognitive ability underestimate the performance of Aboriginal Peoples (Brescia & Fortune, 1989; Darou, 1992; Kleinfield & Nelson; 1991; McShane & Plas; 1984). On the other hand, Aboriginal Peoples tend to do well on nonverbal measure of intelligence (McShane & Berry, 1988; McShane & Plas, 1984). Although there are many types of nonverbal measures of cognitive ability, the majority of these tests tend to measure spatial ability and/or inductive reasoning.

Spatial ability instruments measure a candidate's ability to generate, retain, and transform a variety of complex three-dimensional figures (Allen, Kirasic, Dobson, Long, & Beck, 1996). Spatial ability has long been recognized as a factor contributing to success in mathematics, natural sciences, engineering, architecture, and other fields of study (Miller & Bertoline, 1991; Rhoades, 1981). It predicts an individual's ability to do jobs that require visual analysis and assembly.

There seem to be three main ways in which spatial ability might contribute to mathematics: (1) Geometry emphasizes spatial relationships (Brown & Wheatley, 1989); (2) Some degree of spatial ability is necessary for the correct placement and alignment of digits, and as such must play a part in multi-digit arithmetic (Dahmen, Hartje, Büssing, & Sturm, 1982); and (3) It is possible that spatial representations of the mathematical relationships in a word problem can facilitate its solution (Wheatley, 1991).

Nonverbal measures of inductive analytical reasoning measures the ability to discern meaning in confusion, and the ability to perceive and identify relationships (Raven, Raven & Court, 1998a). In other words, inductive or analytical reasoning involves the ability to reason and solve problems involving new information, without

relying on a base of knowledge derived from pervious experience or schooling (Carpenter, Just, & Shell, 1990). Tests of inductive or analytical reasoning are considered to be measures of higher order cognitive ability and are thought to be one of the finest measures of general intelligence (Stough, Nettlebeck, & Cooper, 1993).

Test Bias

According to classical test theory, the observable test score is made up of a true score and an error score. True scores are the score an individual should receive on the test if there were no errors. Error scores are random errors and exist in all psychological tests. Random errors are unrelated to the individual's true score and can either increase or decrease an individual's true score. In the long run, these increases and decreases will even out so that there is no effect. Unfortunately, test scores can also be affected by an error that is not random called test bias. Test bias refers to a systematic or constant error of measurement in a specified direction, as opposed to random error, associated with group membership (Reynolds & Brown, 1997).

Test fairness is often confused with test bias, but they are not the same thing (Campbell & Cotton, 1994). Test fairness relates to how a test is used, while test bias refers to statistical properties of the test (Cronshaw, 1991). The Society for Industrial and Organizational Psychology (SIOP; 1987) states, "Fairness or lack of fairness is not a property of the selection procedure, but rather a joint function of the procedure, the job, the population, and how the scores from it are used." (p. 49). A test is considered fair if it allows all test takers the same chance to demonstrate their abilities (Fairweather, 1986). A biased test may be used fairly. One acceptable method is to generate separate cut off

scores for different groups based on separate prediction formulas (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

There are a number of different forms in which test bias can present itself such as construct bias, method bias, or item bias (van de Vijver & Tanzer, 1998). Construct bias occurs when the construct or trait that is being measured is not identical across cultural groups. Method bias can be related to group differences on a latent factor that is not related to the factor being studied. Item bias occurs when one group of examinees are less likely to answer one or more items correctly than another group of examinees because of some characteristic of the test or testing situation that is not relevant to the test purpose (Clauser & Mazor, 1998).

Employment Equity

In recent years, the CF has made a considerable effort to increase the representation of Aboriginal Peoples in the CF (Murray, 1999). This drive for diversity was fuelled primarily by amendments that were made to the Employment Equity Act in 1996 (Bussiere, 1997). Prior to 1996, the CF was exempt from the Act that required all federal government agencies to increase the representation of designated minority groups (women, Aboriginal Peoples, visible minorities, and people with disabilities) in their employee pool until it attained a level that is reflective of the Canadian workforce.

In compliance with the Employment Equity Act, the CF conducted an analysis of the composition of its service members (Ewins, 1997). The representation of designated minority groups in the CF was far below that of the Canadian population. Specifically,

14.1% of CF members identified themselves as female, 2.1% as visible minorities and 1.4% as Aboriginal Peoples, compared to the composition of the Canadian workforce where 50.7% of Canadians were female, 9.4% were visible minorities and 3.2% were Aboriginal Peoples (Smith, 1995).

Initially, researchers believed that the discrepancy between the representation of the Canadian population and the CF was a consequence of the CF's inability to attract members of designated minority groups. However, the Environics Research Group Limited (1997) demonstrated that the CF was indeed attracting individuals from designated minority groups. Nonetheless, the number of interested members of designated minority groups largely outweighed the actual number of minorities enrolled in the CF.

A review of the entire recruitment and selection process was undertaken to determine why members of designated minority groups were not enrolling in the CF despite their stated interests. The results of this analysis indicated that one of the two tests used to screen suitable applicants, the General Classification (GC) Test was biased against members of designated minority groups (Guelph Centre for Occupational Research, 1997). Although the CF no longer administers the GC, its successor, the Canadian Forces Aptitude Test (CFAT), was derived in part from the GC.

Prior to the implementation of the CFAT, the GC was used in conjunction with the Canadian Forces Classification Battery (CFCB). The administration of both the GC and CFCB was very time consuming. The CFAT was designed to streamline the selection process. The items of the CFAT were derived from a combination of items from both the GC and CFCB. Therefore, there were some concerns that the CFAT might also be biased

against minority groups. In light of this concern, the CF commissioned two separate studies to analyse the cultural fairness of the CFAT. Although both studies used the same data set, each study employed different methods to detect item bias. Using the four-fifths rule of adverse impact, Bussiere (1997) found that the CFAT was adversely biased against Aboriginal Peoples. In a separate study, Zumbo and Hubley (1998b) used Differential Item Functioning (DIF) analysis to demonstrate that the CFAT was not biased against Aboriginal Peoples. Zumbo and Hubley (1998b) argue that the contradictory findings between the two studies were due to methodological and analytical differences.

In an attempt to determine whether or not the CFAT was indeed biased against Aboriginal Peoples and/or other designated minority groups, the CF commissioned yet another study using the four-fifths rule of adverse impact (Organization and Management Solutions & Myklebust, 2000). This study concluded that there was no evidence of adverse impact against Aboriginal Peoples, visible minorities and females. However, it did find that non-Aboriginal Peoples were 1.5 times more likely to be enrolled in the CF than Aboriginal Peoples.

Determining if the CFAT is Biased

The primary objective of this study is to test whether the CFAT is biased against Aboriginal Peoples. To determine this, two approaches will be used. First logistical regression will be used to determine if any items of the CFAT display DIF. The second approach will entail the use of the four-fifths rule to determine if the CFAT adversely

impacts against Aboriginal Peoples. Multivariate and univariate analysis will also be used to assess group differences on tests means.

In order for the CFAT to be considered a fair and unbiased measure of cognitive ability for Aboriginal Peoples, the following conditions should be met:

1. There should not be significant group differences on test means between Aboriginal and non-Aboriginal Peoples on the CFAT total score, VS scale, SA scale and PS scale;
2. None of the items of the CFAT should display DIF; and
3. The CFAT should not adversely impact against Aboriginal Peoples.

Differential Item Functioning (DIF)

Differential Item Functioning (DIF) is one the most effective methods of detecting test bias (Zumbo, 1999). DIF statistical techniques are based on the assumption that examinees that have the same amount of an underlying trait that is being measured should perform similarly on different items of the test regardless of their group membership (Hambleton, Swaminathan, & Rogers, 1991). If one group of examinees performs differently on any item, then DIF is said to be present. Ackerman (1992) contends that DIF is due to the presence of a secondary nuisance dimension that intrudes on the ability being measured. For example, a word problem on a test of mathematical ability may inadvertently measure verbal ability as well. Consequently, examinees with low verbal ability may perform differently from examinees with high verbal ability, even though examinees from both groups may have the same mathematical ability.

In DIF analysis, the goal is to compare the performance of two groups. The Focal group is usually composed of the subpopulation of interest to the researcher, whereas the Reference group serves as the standard for comparison (Stark, Chernyshenko, Chuah, Lee & Wadlington, 2001). DIF is a necessary but not sufficient condition for item bias (Zumbo, 1999). In other words, if DIF is not present, then there is no item bias. On the other hand, the presence of DIF is not sufficient to pronounce that an item is biased. In order to determine if an item is indeed biased one would need to conduct a follow up study using content or empirical analysis.

Logistical Regression. Although a variety of DIF analysis methods exist, logistical regression is the most recommended and effective method (Robie, Mueller, & Campion, 2001; Clauser & Mazor, 1998; Zumbo, 1999). Several studies have demonstrated that, compared to other popular procedures like the Mantel-Haenszel (MH) and Simultaneous Item Bias Test (SIBTEST), logistical regression has comparable power to detect uniform DIF and superior power to detect non-uniform DIF (uniform and non-uniform DIF are described on page 13; Li & Stout, 1996; Narayanan & Swaminathan, 1994: Rogers & Swaminathan, 1993; Whitmore & Schumacker, 1999). Power is defined as the rate of correct identification of items that display DIF (Jodoin & Huff, 2001).

Another advantage of logistical regression is that it is less sensitive to sample size than Item Response Theory (IRT) methods of DIF. Consequently, logistical regression is the method of choice when sample sizes are less than 200 per group (Robie et al., 2001; Swaminathan & Rogers, 1990).

Logistical regression is based on a statistical modeling of the probability of correctly responding to an item by group membership (i.e. Focal group and Reference group) and a criterion variable (usually scale or subscale score; Zumbo, 1999; see Appendices A, B and C for a more complete explanation on how to conduct and interpret DIF analysis using logistic regression). In logistical regression, a model comparison is performed in which an item response (0 = incorrect, 1 = correct) is predicted by the scale score under investigation, group membership (0 = Reference, 1 = Focal), and the interaction between scale score and group membership (Robie et al., 2001). These variables are entered hierarchically in the following order:

Step 1: Scale score,

Step 2: Group membership, and

Step 3: Interaction between scale score and group membership.

For an item to be classified as displaying DIF, two criteria must be met. The DIF must be statistically significant and the magnitude of the significance (effect size) must be substantial and meaningful (Robie et al., 2001: Zumbo, 1999). To assess significance, a likelihood ratio Chi-squared (χ^2) test is computed. In order to meet the first criterion, the p-value for the two-degree of freedom χ^2-value must be ≤ 0.01 (Zumbo, 1999). An alpha level of .01 is used to control for the number of statistical tests being conducted (Robie et al., 2001).

The second criterion requires effect size to be substantial and meaningful. To meet the second criterion, the entire model (scale score, group membership, scale score $\times$ group membership) must account for at least 13% of the variance in the outcome variable (Robie et al., 2001; Zumbo & Thomas, 1997). Zumbo and Thomas (1997) have

devised a simple procedure to examine effect size. The Zumbo-Thomas effect size measure is correlated with other DIF procedures like the MH and SIB test (Gierl, Rogers, & Klinger, 1999).

DIF analysis is based on the assumption that both groups have the same amount of an underlying trait that is being measured. However, in reality, unequal ability distribution between Focal and Reference group is quite common (Jodoin & Huff, 2001). Type I error rate increases and power decreases when the ability distribution of the Focal and Reference groups have unequal means (Narayanan & Swaminathan, 1994). Type I errors occurs when items are identified as exhibiting DIF, when, in fact, DIF is not present (Jodoin & Huff, 2001). When using logistical regression, the use of tests of significance that consider effect size reduces the Type I error rate associated with ability distribution differences (Jodoin & Gierl, 2001: Narayanan & Swaminathan, 1994; Rogers & Swaminathan, 1993); even when the sample is small (Jodoin & Huff, 2001).

Uniform vs. Non-uniform DIF. In the absence of DIF, the item characteristics curves (ICC) for the two groups would be the same (see Figure 1). ICC is an s-shaped curve that represents the relationship between the item and total score. Figure 2 presents an example of an item that displays a substantial uniform DIF. The DIF is uniform because the ICCs for both groups are similar in shape, however one ICC is shifted to the right or to the left of the other; the ICCs do not cross. A uniform DIF may indicate that the item is not an equivalent measure of the same variable for both groups (Zumbo, 1999). In other words, the probability of getting the item correct is different for both groups and these differences are fairly stable across score levels (Robie et al., 2001). An example of an

item displaying a non-uniform DIF is presented in Figure 3. The DIF is non-uniform because the ICC of one group is different in shape and crosses over the ICC of the other group. For those individuals who score at or above the mean (i.e. $z \geq 0$), Group 2 is favoured whereas for those scoring below the mean (i.e. $z < 0$) Group 1 is favoured (Zumbo, 1999). In other words, the probabilities for getting the item correct are different for the two groups and the differences are not necessarily stable across score levels (Robie et al., 2001).

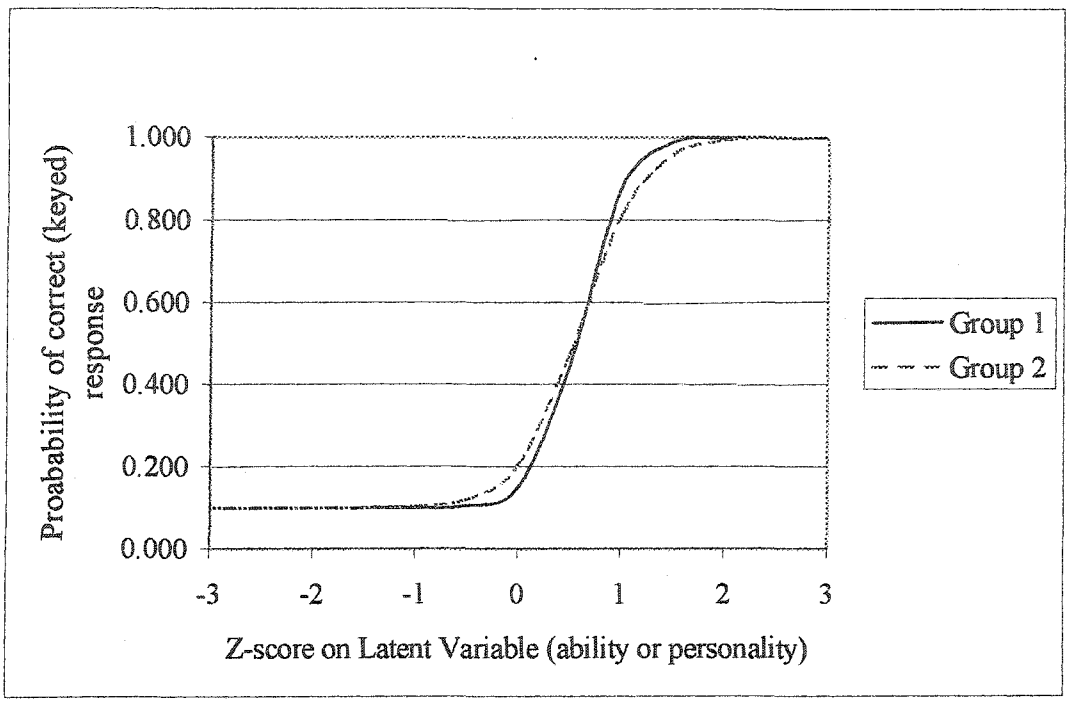

Figure 1. An example of an item that does not display DIF.

Note. From *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistical Regression Modelling as a Unitary Framework for Binary and Likert Type (Ordinal) Item Scores,* by B. D. Zumbo, 1999, Ottawa, ON: Director Human Resources Research and Evaluation. Copyright 1999 by Her Majesty the Queen in Right of Canada. Reprinted with permission.

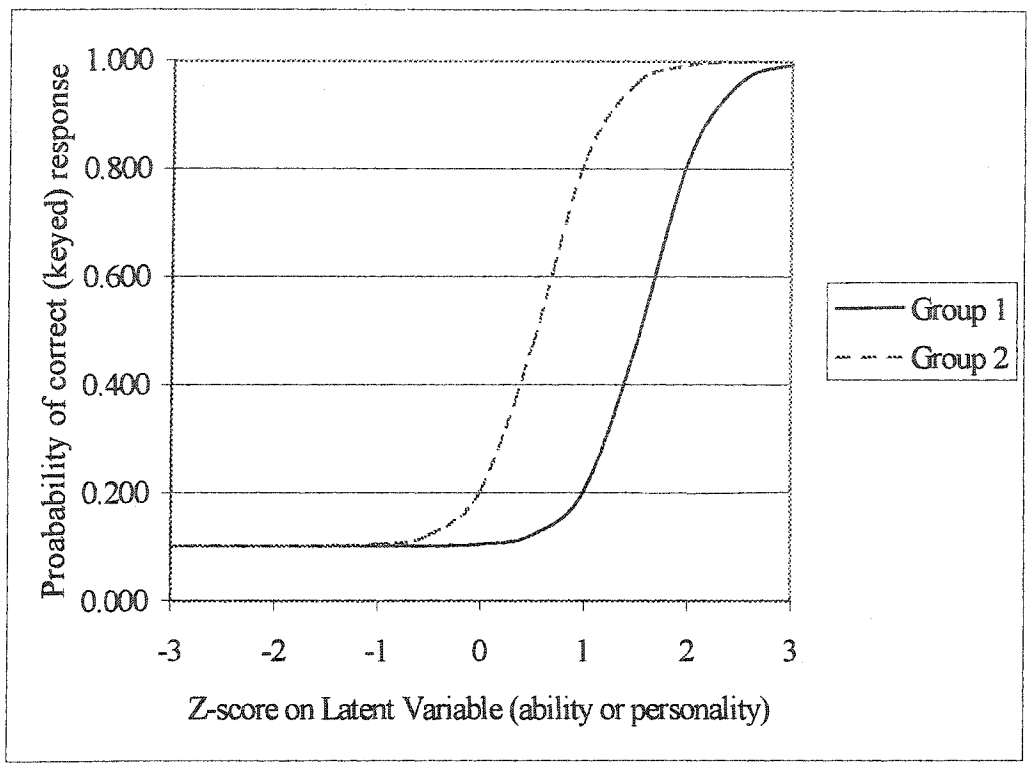

Figure 2. An example of an item that displays substantial uniform DIF.

Note. From *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistical Regression Modelling as a Unitary Framework for Binary and Likert Type (Ordinal) Item Scores*, by B. D. Zumbo, 1999, Ottawa, ON: Director Human Resources Research and Evaluation. Copyright 1999 by Her Majesty the Queen in Right of Canada. Reprinted with permission.

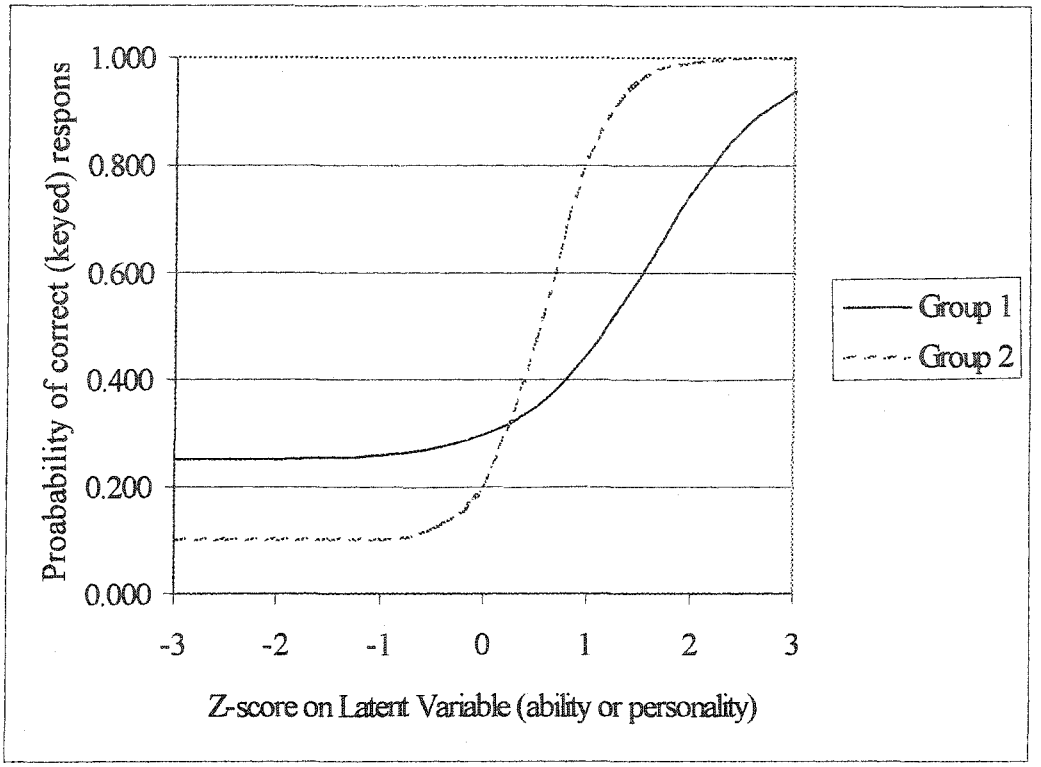

Figure 3. An example of an item that displays substantial non-uniform DIF.

Note. From *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistical Regression Modelling as a Unitary Framework for Binary and Likert Type (Ordinal) Item Scores*, by B. D. Zumbo, 1999, Ottawa, ON: Director Human Resources Research and Evaluation. Copyright 1999 by Her Majesty the Queen in Right of Canada. Reprinted with permission.

Purifying the Matching Variable. Holland and Thayer (1988) note that in the process of conducting a DIF analysis one should "purify the matching criterion". That is, items that are identified as DIF are omitted, and the total or scale score is recalculated. This recalculated score is then used as the matching criterion for a second logistical regression DIF analysis. Again, all items are assessed. In addition, Holland and Thayer (1988) suggest that the item under examination should be included in the matching criterion even if it was identified as displaying DIF and excluded from the criterion for all other items. This procedure decreases Type I errors (Zumbo, 1999).

Adverse Impact

The term adverse impact has legal implications and refers to a situation in which group differences in test performance results in a disproportionate selection of members of a protected group (Clauser & Mazor, 1998). The four-fifths rule is an operational procedure that is often used to detect adverse impact. According to the four-fifths rule, adverse impact occurs when the selection rate for designated minority groups is less than four-fifths that of the comparison group (Catano, Cronshaw, Wiesner, Hackett & Methot, 2001). The four-fifths rule is based on the Impact Ratio, which is the ratio of the selection rate for the minority group to the selection rate for the majority group (EEOC, 1978).

There are some criticisms towards the use of the four-fifth rule to identify adverse impact (Vining, McPhilips, & Boardman, 1986). The four-fifth rule ignores the concepts of chance and statistical significance (Organization and Management Solutions & Myklebust, 2000). More specifically, the four-fifth rule is susceptible to Type I and Type II errors (Boardman, 1979; Greenberg 1979). Type I error is the risk of falsely identifying

cases of adverse impact where none exists. Type II error is the likelihood of failing to identify cases of adverse impact when it does exists. When sample sizes are small, there is an increased risk of Type I errors. Conversely, when sample sizes are large, there is an increased risk of Type II errors.

There is also some criticism towards applying the four-fifths rule of adverse impact to detect item bias (Camilli & Sheppard, 1994). Although the detection of adverse impact may raise some concerns about the selection process it does not mean that a test is biased (Camilli & Shepard, 1994). Zumbo and Hubley (1998b) argue that adverse impact is not related to test performance and that its use in item bias detection ignores sampling variability.

Despite these criticisms, the four-fifths rule of adverse impact is still the most favoured method for determining adverse impact in employment discrimination cases (Morris & Lobsenz, 2000). This popularity is credited to the fact that the four-fifths rule is straightforward and easy to implement.

Finding a Suitable Replacement for the CFAT

If the CFAT is biased, the second objective of this study is to determine if another valid and reliable verbal or nonverbal measure of cognitive ability could be used in lieu of the CFAT. Unfortunately researchers cannot agree on any one test as being suitable for Aboriginal Peoples. One thing researchers can agree on is that verbal measures of general cognitive ability are biased against Aboriginal Peoples (Brescia & Fortune, 1989; Darou, 1992; Kleinfield & Nelson; 1991; McShane & Plas; 1984). On the other hand, nonverbal measures of general cognitive ability are believed to biased towards Aboriginal Peoples

(McShane & Plas, 1984). However, this relationship between nonverbal measures and cultural bias has not been consistent (Parmar, 1989).

To accomplish the second objective, two well-established verbal and nonverbal measures of cognitive ability were selected for use in this study. The Wonderlic Personnel Test (WPT) is a reliable verbal measure of general cognitive ability that has been used extensively in industrial and organizational psychology (Dodrill, 1983). The Raven's Standard Progressive Matrices (SPM) is generally regarded as a culture fair non–verbal test of cognitive ability (Marshalek, Lohman, & Snow, 1983). However, by itself, the SPM does not provide a complete assessment of an individual's cognitive ability (Raven, et al., 2000). To better assess cognitive ability, the Mill-Hill Vocabulary scale (MHV) is often administered with the SPM (Raven, et al., 2000). Therefore, it was decided to include the MHV in this study.

A DIF analysis will be conducted to evaluate the suitability of selecting Aboriginal Peoples with the WPT, SPM and MHV. Prior to the DIF analysis, the reliabilities of the CFAT, SPM, MHV and WPT will be examined. Multivariate and univariate analysis will also be used to assess group differences on tests means.

Correlation analysis will be performed to examine the relationships between each measure. Of particular interest is the relationship of CFAT, SPM and MHV to the WPT. Of all the cognitive ability measures being used in this study, the WPT is the most widely used and highly regarded. A high correlation between the WPT and the remaining measures will provide evidence for convergent validity. Convergent validity, which is a type of construct validity, refers to the principle that measures that should be related are in reality related (Trochim, 2000). Construct validity refers to how well the test measures

ability, characteristics or other attributes of the test taker (Cronbach & Meehl, 1955). Convergent validity can be demonstrated with a correlational analysis between the measures in question with an established measure of the ability under study. Correlations between theoretically similar measures should be "high" while correlations between theoretically dissimilar measures should be "low" (Trochim, 2000).

Wonderlic Personnel Test

The WPT is a 50-item test designed to measure verbal, numerical, analytical and spatial abilities (Murphy, 1984). The items are arranged in order of difficulty, beginning at a modest level and gradually increasing (Wonderlic, 1997). The WPT measures the level at which an individual learns, understands instructions and solves problems. The test produces an overall score based on the number of items answered correctly in 12 minutes. The WPT comprises both multiple choice and short answer questions and requires a sixth grade reading level (Frisch & Jessop, 1989).

The publisher of WPT claims that the test is culture free and has few effects on minorities. However, Chan (1997) found that African Americans had lower predictive validity perceptions towards the WPT than Caucasians. Since predictive validity perceptions are positively correlated with performance on the WPT, Chan concluded that African Americans would receive lower scores. On the other hand, Dodrill (1981) found that the predictive validity of the WPT was not influenced by variables such as sex, age, emotional adjustment and years of education.

In order for the WPT to be an unbiased and reliable measure of cognitive ability for Aboriginal Peoples, the following conditions should be met:

1. There should not be significant group differences on test means between Aboriginal and non-Aboriginal Peoples on the WPT; and
2. None of the items of the WPT should display DIF.

Raven's Standard Progressive Matrices

Raven's Standard Progressive Matrices (SPM) is considered one of the best available measures of general intelligence and complex reasoning (Marshalek, Lohman, & Snow, 1983). The SPM is a nonverbal test designed to assess inductive or analytical reasoning (Bors & Stokes, 1998).

The SPM meets all the criteria for what is generally considered to be a culture fair test (Albert, 1998a). The test is not timed, is free of culturally laden material, does not involve language, can be administered individually or to groups and the instructions can be pantomimed (Vincent, 1991). The SPM is unbiased against a culturally diverse population including Aboriginal Peoples (Raven, Raven & Court, 2000). Furthermore, researchers have found that the primary language of Navajo adolescents (Navajo or English) did not influence SPM scores (Sidles, MacAvoy, Bernstone, & Kuhn, 1987). The SPM was designed for use to assess adults of average intelligence.

The SPM consists of 60 items divided into five sets of 12. Each set and the items within each set get progressively harder. The easier problems at the beginning of each set provide training for solving the more difficult subsequent problems (Matthews, 1988). Even though the stimuli themselves are completely nonverbal, the SPM correlates highly with verbal measures of cognitive ability (Saccuzzo & Johnson, 1995).

In order for the SPM to be an unbiased and reliable measure of cognitive ability for Aboriginal Peoples, the following conditions should be met:

1. There should not be significant group differences on test means between Aboriginal and non-Aboriginal Peoples on the SPM;
2. None of the items of the SPM should display DIF; and
3. To display construct validity, the SPM should correlate positively and highly with the WPT and CFAT.

Mill Hill Vocabulary Test

Because the SPM measures only nonverbal ability, it is not a complete measure of general intelligence (Raven et al., 2000). In order to get a more complete picture, the publishers' of SPM recommend using the test in conjunction with the Mill Hill Vocabulary scale (MHV). The MHV is a measure of reproductive ability. Reproductive ability is the ability to recall and use verbal knowledge (Raven et al., 2000).

The MHV comes in a variety of forms. For the purpose of this study, the Senior Multiple Choice form was used. The MHV consists of 68 items divided into two sets of 34. Each item within each set gets progressively harder.

There is no evidence that the MHV is a fair or unbiased test. However, without the MHV, an individual's ability to understand or comprehend verbal or written language would not be assessed. In order for recruits to successfully complete basic and occupational training they must be able to function in either an English or French environment. Including the MHV with the SPM, allows for an exploration of the fairness

of the MHV with a sample of Aboriginal Peoples. This study will focus solely on the ability to comprehend and use English words.

In order for the MHV to be an unbiased and reliable measure of cognitive ability for Aboriginal Peoples, the following conditions should be met:

1. There should not be significant group differences on test means between Aboriginal and non-Aboriginal Peoples on the MHV;
2. None of the items of the MHV should display DIF; and
3. To display construct validity, the MHV should correlate positively with the WPT and CFAT.

Method

Participants

Prior to the administration of the tests, participants were briefed on the purpose of the study and were asked to read and sign an informed consent form. Participation in this study was voluntary. Participants were informed that they could refuse to participate and that they could terminate their participation at any time. Aboriginal Peoples received fifty dollars for their participation in this study.

Focal Group

The Focal group was composed of 101 Aboriginal Peoples (63.4% male v. 36.6% females) living in three special access and one remote community in northern Manitoba. According to Indian and Northern Affairs Canada (INAC) special access refers to a First Nation community that does not have a year round road access to the nearest service access (2002a). A remote community refers to a First Nation community that is over 350 kilometre from the nearest service access and has year round road access (INAC, 2002a). A slight majority of participants indicated that English was their primary language (50.5%) while the remainder used Ojicree (33.7%) or a local dialect (15.8%). Although participants ranged from 18 to 28 years of age, the majority (69.3%) were 22 years or younger. The majority of participants (71.3%) had completed grade 10 or higher. Table 2 presents a break down of participants by last grade completed.

Several of the Aboriginal Peoples who participated in the study (28.7%) had not completed grade 10 and were not eligible to join the CF. Of the eligible participants, several Aboriginal Peoples (29.2%) did not meet the CFAT minimum score required for

Anglophone Non-commissioned Members (NCM) applicants. Consequently, a new group was created by dropping the non-eligible participants from the Focal group. This new group was created to allow more appropriate comparisons with the Reference group, which consisted of CF recruits who had completed grade 10 and met the required minimum score on the CFAT.

Eligible Group

The Eligible group was composed of 51 Aboriginal participants including 28 males (54.9%) and 23 females (45.1%). A slim majority of participants indicated that English was their primary language (51%) while the remainder used Ojicree (27.5%) or a local dialect (21.6%). Although participants ranged from 18 to 27 years of age, the majority (62.1%) were 22 years or younger. A break down of participants by last grade completed is presented in Table 2.

Reference Group

The Reference group was composed of 108 CF recruits undergoing basic training at the Canadian Forces Leadership and Recruit School in Saint Jean, Quebec. Eighty-three (76.9%) males and 25 (23.1%) females participated in this study. Although participants ranged from 18 to 45 years of age, the majority (68.5%) were 25 years or younger (48.1% were 22 years or younger). The minimum education requirement for enrolment in the CF is grade 10. Therefore, unlike the Focal group, all participants in the Reference group had completed grade 10 and had met the minimum score required on the CFAT. A break down of participants by last grade completed is presented in Table 2.

Table 2

Breakdown of Focal Group, Eligible and Reference Group by Last Grade Completed

	Focal		Eligible		Recruits	
Last Grade Completed	N	%	N	%	N	%
8	7	6.9	0	0	0	
9	22	21.8	0	0	0	
10	19	18.8	9	17.6	1	0.9
11	23	22.8	15	29.4	10	9.3
12	27	26.7	24	47.1	50	46.3
13	0		0		1	0.9
Some College	3	2.97	3	5.9	23	21.3
College Diploma	0		0		10	9.3
Some University	0		0		9	8.4
University Degree	0		0		4	3.7
Total	101	100	51	100	108	100

Measures

Canadian Forces Aptitude Test

The CFAT is a speeded test arranged in ascending order of difficulty. This test is composed of 60 items that measures three facets of cognitive ability: verbal skills (VS), spatial ability (SA) and problem solving (PS). Each scale produces a score, which is combined into one overall score. Items not completed are scored as incorrect.

The VS, SA, and PS scales have internal consistency reliabilities of .87, .88 and .91 respectively (Black, 1999). The ability of the CFAT to predict occupational performance has been demonstrated by numerous studies (Campbell, 2001; Ibel & Cotton, 1994; MacLennan, 1997; Woycheshin, 1999).

Wonderlic Personnel Test

The WPT is a group-administered test of general cognitive ability consisting of 50 short answer and multiple-choice items that must be completed in twelve minutes. The WPT correlates highly with the WAIS-R (r = .92; Dodrill & Warner, 1988) and has an internal consistency reliability ranging from .83 to .89 (McKelvie, 1989). Test re-test reliabilities range from .82 to .94 (Wonderlic, 1997). In an earlier study, the WPT correlated moderately with the CFAT total score among two different samples (r = .63 and r = .69: Albert, 1998b).

Raven's Standard Progressive Matrices

The SPM consists of 60 items divided into five sets of 12 diagrammatic puzzles (Raven et al., 2000). Each two-dimensional puzzle has a part missing that the test taker

must identify among the options provided. Concurrent validity with the Wechsler Adult Intelligence Adult Intelligence Scale-Revised (WAIS-R) ranges from .65 to .88 (Burke, 1985; Sheppard, Fiorentino, Collins, 1968). The SPM has an internal consistency reliability of .96 (Burke, 1985). Test re-test reliabilities range from .87 to .92 (Nkaya, Huteau, & Bonnet, 1994). In a previous study (Albert, 1998b), the SPM correlated with the WPT (r = .49) and with the CFAT (r = .49). The SPM correlated most strongly with spatial ability subscale of the CFAT (r = .51).

Mill Hill Vocabulary Test

The MHV also comes in a variety of forms. For the purpose of this study, the MHV Senior Multiple Choice form was used. The MHV Senior form consists of 68 multiple choice items dived into two sets. The published internal consistency reliability of the MHV is .90 (Raven, Raven & Court, 1998b). Test re-test reliabilities range from .74 to .90 (Watts, Baddeley, & Williams, 1982). Previous research has shown that the MHV correlated modestly with the SPM (r = .34 to r = .46; Deary, 1995), which suggest that both tests measure different aspects of cognitive functioning.

Procedure

Focal Group and Eligible Group

Aboriginal participants were asked to complete the CFAT, WPT, SPM, and MHV in a single session. The CFAT was administered first, followed by the WPT, SPM and MHV. Participants were given a fifteen-minute break between the administration of the CFAT and the WPT, and a five-minute break between the WPT and the remainder of the

tests. The administration of the tests followed the procedures outlined in the respective tests manual (Director Recruiting Education and Training, 1998; Raven, Raven & Court, 2000; Raven, Raven & Court, 1998b; Wonderlic, 1997).

Reference Group

The same procedures used in administering the tests to the Focal group were used with the recruits with one exception. The recruits were only asked to write the WPT, SPM and MHV. With the express permission of the recruits, the CFAT data were obtained from archival databases. The recruits had already written the CFAT prior to enrolling in the CF.

Data Analysis

Descriptive Statistics

The means and standard deviations for each measure were examined to determine the distribution of sample and to examine group differences. Cronbach's alphas were calculated to assess the reliability of each scale.

Comparison of Means

It was predicted that there would be no significance differences in test means on the nonverbal measures (CFAT SA, SPM) between the Aboriginal and non-Aboriginal Peoples. Conversely, it was expected that there would significant differences on the verbal measures of cognitive ability (CFAT VS and PS, WPT, MHV) between the Aboriginal and non-Aboriginal Peoples.

The rationale for forming the Eligible group was that the Focal group and Reference group were not equal because of education and CFAT scores. These differences would, thus, confound the results. In order to verify this prediction, test means were compared separately for the Focal and Reference groups and for the Eligible and Reference groups using multivariate and univariate analyses.

Almost half of the Aboriginal Peoples in both the Focal and Eligible groups indicated that English was not their primary language. Testing individuals in their second language may confound test results of verbal measures. Consequently, separate univariate analyses of covariance were performed for each test to determine if the difference in test means between groups was significant after controlling for language. Language was expected to influence performance on the verbal measures of cognitive ability (CFAT VS and PS, WPT, MHV) but not on the nonverbal measures (CFAT SA, SPM).

In addition, a multivariate analysis of covariance was performed to determine if the difference in group means between the Focal and Eligible groups were significant after controlling for education and CFAT total score. It was anticipated that the results would be similar to those found in the multivariate and univariate analyses of Eligible and Reference groups. If the results were similar, than it could be argued that the creation of the Eligible group was justified.

Prior to conducting the analysis, the data for each group were examined to determine whether or not they met the necessary assumptions. The data for each scale, with the exception of the CFAT VS, met the assumptions of homogeneity of variances, normality, linearity, homogeneity of regression and reliability of covariance. The CFAT VS scale did not have equal variance. However, this was not considered an issue because

the Reference group (n = 108) produced more variances than the Focal and Eligible groups (Tabachnick & Fidell, 2001)[1]. Although several outliers were detected, they were kept in the analysis[2].

Correlation Analysis

A correlation analysis examined the relationship between the CFAT, WPT, SPM and MHV. Separate analyses were conducted for the Reference, Focal, Eligible and combined groups.

Differential Item Functioning (DIF) Analysis

To detect DIF, logistical regression was performed following the methods outlined by Zumbo (1999) and endorsed by Swaminathan and Rogers (1990). For an item to be classified as displaying DIF, the p-value for the two-degree of freedom χ^2 in logistical regression had to be ≤ 0.01 (Robie et al., 2001; Zumbo, 1999). Furthermore, the effect size (R^2) had to be > 0.130 (Zumbo & Thomas, 1997).

Prior to conducting a DIF analysis, the data sets for each group were examined to determine whether or not they met the necessary assumptions required for DIF analysis. The underlying assumption of DIF statistical techniques is that examinees of both groups have the same amount of the underlying trait that is being measured. Violating this assumption increases the rate of Type I error and decreases the power (Narayanan &

---

[1] A Mann-Whitney test was performed on the CFAT VS data because it did no meet the assumption of homogeneity of variances. However, the results of the nonparametric test were similar to that of the univariate analysis of variance.

[2] A MANCOVA of the CFAT, SPM, MHV and WPT was conducted with outliers left in and with outliers deleted. There was no difference in the results.

Swaminathan, 1994). A review of the means in Table 2 suggests that the CFAT SA and SPM data meets this assumption. On the other hand, there is a significant difference in the distribution of means between the groups among the remaining CFAT scales, WPT and MHV. However, several studies have demonstrated that Type I error rate associated with ability distribution differences decreases when effect size is taken into consideration. (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). With this in mind a DIF analysis of the CFAT, SPM, WPT and MHV items was still conducted.

Several items from the SPM and WPT did not meet the assumption of adequacy of expected frequencies and power. According to this assumption, the frequencies of responses for each item must be greater than one. In addition, to ensure adequate power, no more than 20% of all of the frequencies must be less than five. The frequencies for four SPM and two WPT items were not greater than one. Logistic regression could not be performed on these items because there was no variance in item response patterns. Two items from the MHV scale (A1 and B1) were practice items and were not included in the analysis. With regards to all the tests, there was no evidence of multicollinearity or singularity. Although several outliers were detected, it was decided to keep them in the analysis[3].

Logistic regression was used to compare the Focal group against the Reference group as well as the Eligible group against and the Reference group. The results were the same for both analyses. Therefore, only the results of the DIF analysis for the Focal and Reference groups will be reported.

---

[3] A DIF analysis of the CFAT, SPM, MHV and WPT were conducted with outliers left in and with outliers deleted. There was no difference in the results.

Adverse Impact

The four-fifths rule was used to determine if the CFAT adversely impacted against Aboriginal Peoples. According to this rule, adverse impact is established when the selection rate for designated minority groups is less than four-fifths of the of the non-designated minority group. For example, imagine that 60 % of Aboriginal Peoples met the criteria for enrolment in the CF based on CFAT score, while 80% of non-Aboriginal Peoples met the requirement. The Impact Ratio of Aboriginal Peoples who meet the criteria to that of non-Aboriginal Peoples is .75 (60/80), which is less than four fifths. Consequently, it can be alleged that the CFAT was adversely impacting against Aboriginal Peoples.

To determine if the CFAT was adversely impacting against Aboriginal Peoples, the selection rates for recruits enrolling in the CF were calculated from the entire Anglophone NCM recruit applicant CFAT scores that have been collected since 1997 (n = 53169). It was not possible to determine the actual selection rates for the recruit applicant population because it was not known which recruit applicant was actually selected or which occupation the applicant was selected for. Also, there are other requirements and assessment tools used during the selection process that were not taken into account. Instead, the term selection rate is used in this study to describe the percentage of Anglophone NCM recruit applicants who achieved the minimum CFAT cut score regardless of whether or not they had been selected. The same is true of the Aboriginal Peoples selection rate. For the purpose of this study, the Aboriginal Peoples selection rate refers to the percentage of Aboriginal Peoples who achieved the minimum CFAT cut score. Selection rates were also calculated for the percentage of Anglophone

NCM recruit applicants and Aboriginal Peoples who meet the minimum CFAT score required for selection into the various families of military occupations. Once again, these percentages were determined based on CFAT score regardless of which occupations Anglophone NCM recruit applicants had been selected for. All of the Anglophone NCM recruit applicants had completed, as a minimum, grade 10. Therefore, only Aboriginal Peoples who had completed grade 10 (n = 72) were included in this analysis.

Results

Descriptive Statistics

The means and standard deviations for each measure are presented in Table 3. There are considerable differences in Focal and Reference group means among the CFAT, WPT, SPM and MHV. However, there was no significant difference between the Eligible and Reference groups on the SPM. Furthermore, a closer look at the CFAT reveals that although there were differences in means among the VS and PS scales there was no significant difference between the Eligible and Reference groups on the SA scale. This is not surprising since the SA scale is a nonverbal measure.

There was no significance difference in mean scores between males and females on the WPT, SPM, MHV and all subscales of the CFAT. This was true for both Reference and Aboriginal Peoples groups.

Canadian Forces Aptitude Test

The means and standard deviations for each item of the CFAT are presented in Tables 4 through 6. As can be seen in Table 4, the Reference group scored considerably higher than the Focal and Eligible groups on the majority of items from the VS scale. On the other hand, all three groups performed similarly on the SA scale (see Table 5). The Eligible group scored higher than the Reference group on the majority of the SA items. The Reference group scored noticeably higher on the majority of the PS scale items (Table 6). A closer inspections reveals that the Focal and Eligible groups obtained their highest score on items PS2, PS3, PS4, PS5 and PS13. These items consisted of questions measuring basic math ability (PS3, PS4, PS5) and the ability to discern simple patterns (P2, P13). Conversely, the Focal and Eligible group performed poorly on questions

composed of word problems (PS29, PS28) and mathematical equations involving fractions and/or decimal points (PS19, PS22, PS24).

Table 3

Descriptive statistics of CFAT, SPM, MHV and WPT Scales for Focal, Eligible and Reference Groups

		Focal Group (N= 101)		Eligible Group (N = 51)		Reference Group (N=108)	
Scale	Subscales	Mean	SD	Mean	SD	Mean	SD
CFAT	VS	4.78[1]	2.46	5.88[3]	2.54	8.60	3.56
	SA	8.35[2]	2.84	9.84	2.32	9.19	2.72
	PS	9.20[1]	5.39	12.06[3]	5.63	17.73	5.36
	Total	22.33[1]	8.70	27.78[3]	8.13	35.52	8.95
SPM	Set A	11.44	.82	11.53	.86	11.25	1.43
	Set B	9.93	2.05	10.49	1.74	10.69	1.26
	Set C	8.63	1.93	9.00	1.92	9.11	1.65
	Set D	8.17	2.48	8.76	2.27	8.70	2.18
	Set E	4.23	2.66	4.94	2.67	5.15	2.98
	Total	42.47[2]	7.49	44.86	6.70	45.15	6.68
MHV	Set A	20.82	4.46	22.41	4.87	25.81	3.65
	Set B	20.46	4.61	22.51	4.33	25.44	4.86
	Total	41.39[1]	8.68	45.31[3]	8.32	51.07	7.31
WPT	Total	14.21[1]	5.71	16.82[3]	5.06	23.08	5.42

[1] Difference between the Focal and Reference group significant at .001
[2] Difference between the Focal and Reference group significant at .05
[3] Difference between the Eligible and Reference group significant at .001

Table 4

Descriptive statistics of CFAT Verbal Skill Items

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
VS1	.44	.50	.45	.503	.56	.50
VS2	.20	.40	.27	.45	.61	.49
VS3	.36	.48	.51	.51	.81	.39
VS4	.17	.38	.20	.40	.68	.47
VS5	.19	.39	.27	.45	.47	.50
VS6	.51	.50	.55	.50	.63	.49
VS7	.23	.42	.29	.46	.58	.50
VS8	.26	.44	.35	.48	.58	.50
VS9	.74	.44	.80	.40	.67	.47
VS10	.36	.48	.47	.50	.71	.45
VS11	.30	.46	.41	.50	.69	.47
VS12	.29	.46	.43	.50	.43	.50
VS13	.19	.39	.22	.42	.52	.50
VS14	.19	.39	.20	.40	.27	.45
VS15	.38	.49	.45	.50	.39	.49

Table 5

Descriptive statistics of CFAT Spatial Ability Items

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
SA1	.75	.43	.88	.32	.81	.40
SA2	.78	.42	.88	.33	.79	.41
SA3	.69	.46	.78	.42	.70	.46
SA4	.69	.46	.90	.30	.79	.41
SA5	.64	.48	.75	.44	.78	.42
SA6	.68	.47	.82	.39	.67	.47
SA7	.64	.48	.76	.43	.69	.46
SA8	.55	.50	.67	.48	.78	.42
SA9	.51	.50	.59	.50	.64	.48
SA10	.64	.48	.65	.48	.73	.45
SA11	.57	.50	.73	.45	.52	.50
SA12	.29	.46	.35	.48	.37	.49
SA13	.16	.37	.20	.40	.27	.45
SA14	.37	.48	.39	.49	.31	.48
SA15	.36	.48	.49	.51	.34	.48

Table 6

Descriptive statistics of CFAT Problem Solving Items

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
PS1	.42	.50	.49	.51	.79	.41
PS2	.54	.50	.67	.48	.81	.39
PS3	.46	.50	.61	.49	.76	.43
PS4	.48	.50	.59	.50	.75	.44
PS5	.49	.50	.71	.46	.78	.42
PS6	.30	.46	.43	.50	.67	.47
PS7	.39	.49	.55	.50	.59	.49
PS8	.42	.50	.45	.50	.77	.42
PS9	.42	.50	.59	.50	.76	.43
PS10	.36	.48	.51	.51	.70	.46
PS11	.28	.45	.39	.49	.56	.50
PS12	.19	.39	.22	.42	.74	.44
PS13	.46	.50	.61	.49	.72	.45
PS14	.41	.49	.55	.50	.68	.47
PS15	.32	.47	.39	.49	.59	.49
PS16	.35	.48	.39	.49	.67	.47
PS17	.23	.42	.39	.49	.46	.50
PS18	.40	.49	.55	.50	.60	.49

Table 6 continued

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
PS19	.15	.36	.16	.37	.40	.49
PS20	.22	.42	.33	.48	.59	.49
PS21	.34	.48	.47	.50	.72	.45
PS22	.15	.36	.24	.43	.38	.49
PS23	.21	.41	.29	.46	.40	.49
PS24	.12	.33	.14	.35	.36	.48
PS25	.22	.42	.31	.47	.61	.49
PS26	.19	.39	.18	.39	.52	.50
PS27	.30	.46	.31	.47	.29	.45
PS28	.15	.36	.16	.37	.39	.49
PS29	.14	.35	.14	.35	.30	.46
PS30	.17	.38	.25	.44	.38	.49

Wonderlic Personnel Test

The item means and standard deviations for the WPT are presented in Table 7. The Reference group scored considerably higher than the Focal and Eligible groups on the majority of items. Although the differences between means were smaller, the Eligible group tended to score higher than the Focal group. Aboriginal Peoples had particular

difficulty with items involving word comparisons, sentence parallelisms, and word problems requiring mathematics and logic.

Table 7

Descriptive statistics of WPT

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
WPT1	.97	.17	1.00	.00	.95	.21
WPT2	.87	.34	.90	.30	1.00	.00
WPT3	.88	.33	.92	.27	.98	.14
WPT4	.88	.33	.96	.20	.97	.17
WPT5	.58	.50	.76	.43	.93	.26
WPT6	.57	.50	.67	.48	.83	.37
WPT7	.72	.45	.76	.43	.88	.33
WPT8	.53	.50	.65	.48	.82	.38
WPT9	.61	.49	.71	.46	.71	.45
WPT10	.39	.49	.49	.50	.83	.37
WPT11	.44	.50	.65	.48	.83	.37
WPT12	.37	.48	.55	.50	.74	.44
WPT13	.56	.50	.61	.49	.78	.42
WPT14	.22	.41	.33	.48	.63	.49
WPT15	.25	.43	.43	.50	.56	.50
WPT16	.53	.50	.75	.44	.73	.45
WPT17	.15	.36	.25	.44	.50	.50
WPT18	.31	.46	.49	.50	.61	.49

Table 7 continued.

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
WPT19	.34	.47	.37	.49	.44	.50
WPT20	.44	.50	.59	.50	.81	.40
WPT21	.05	.22	.06	.24	.41	.49
WPT22	.51	.50	.49	.50	.34	.48
WPT23	.16	.37	.24	.43	.56	.50
WPT24	.20	.40	.31	.47	.36	.48
WPT25	.44	.50	.50	.50	.80	.40
WPT26	.40	.49	.55	.50	.81	.40
WPT27	.04	.20	.06	.24	.28	.45
WPT28	.26	.44	.20	.40	.56	.55
WPT29	.08	.27	.14	.35	.28	.45
WPT30	.07	.26	.10	.30	.44	.50
WPT31	.03	.17	.02	.14	.10	.30
WPT32	.27	.44	.29	.46	.35	.48
WPT33	.25	.43	.24	.43	.37	.49
WPT34	.07	.26	.10	.30	.19	.40
WPT35	.00	.00	.00	.00	.02	.14
WPT36	.04	.20	.08	.27	.09	.29
WPT37	.00	.00	.00	.00	.06	.23

Table 7 continued.

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
WPT38	.02	.14	.02	.14	.15	.36
WPT39	.17	.38	.18	.39	.21	.4
WPT40	.07	.26	.06	.24	.21	.41
WPT41	.03	.17	.04	.20	.10	.30
WPT42	.00	.10	.02	.14	.13	.34
WPT43	.09	.29	.10	.30	.30	.46
WPT44	.05	.22	.04	.20	.09	.29
WPT45	.00	.00	.00	.00	.01	.10
WPT46	.02	.14	.02	.14	.08	.28
WPT47	.02	.14	.02	.14	.06	.23
WPT48	.00	.00	.00	.00	.00	.00
WPT49	.00	.00	.00	.00	.07	.26
WPT50	.00	.00	.00	.00	.01	.10

Raven's Standard Progressive Matrices

The item means and standard deviations for the SPM are presented in Tables 8 through 12. The Reference group tended to score slightly higher than the Aboriginal Peoples groups. However, for the most part the differences in means were quite small.

Table 8

Descriptive statistics of SPM Set A

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
A1	.98	.14	.98	.14	.99	.10
A2	.99	.10	.98	.14	.99	.10
A3	.99	.10	.98	.14	.99	.10
A4	.99	.10	.98	.14	.99	.10
A5	1.00	.00	1.00	.00	1.00	.00
A6	1.00	.00	1.00	.00	1.00	.00
A7	.98	.14	1.00	.00	.97	.17
A8	.96	.20	.98	.14	.94	.23
A9	1.00	.00	1.00	.00	.99	.10
A10	.97	.17	1.00	.00	.95	.21
A11	.93	.26	.96	.20	.90	.30
A12	.64	.48	.67	.48	.63	.49

Table 9

Descriptive statistics of SPM Set B

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
B1	.99	.10	.98	.14	.99	.10
B2	.97	.17	.96	.20	.98	.14
B3	.98	.14	.98	.14	.98	.14
B4	.99	.10	1.00	.00	.98	.14
B5	.98	.14	.98	.14	1.00	.00
B6	.82	.39	.80	.40	.93	.26
B7	.72	.45	.75	.44	.80	.41
B8	.72	.45	.88	.33	.84	.37
B9	.71	.46	.80	.40	.85	.36
B10	.86	.35	.94	.24	.95	.21
B11	.66	.48	.76	.43	.71	.45
B12	.52	.50	.65	.48	.63	.49

Table 10

Descriptive statistics of SPM Set C

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
C1	.97	.17	1.00	.00	1.00	.00
C2	.94	.24	.96	.20	.99	.10
C3	.93	.26	.94	.24	.96	.19
C4	.90	.30	.90	.30	.82	.38
C5	.98	.14	1.00	.00	.96	.19
C6	.77	.42	.80	.40	.85	.36
C7	.91	.29	.88	.33	.93	.26
C8	.66	.48	.73	.45	.62	.49
C9	.64	.48	.69	.47	.78	.42
C10	.47	.50	.47	.50	.47	.50
C11	.31	.46	.37	.49	.45	.50
C12	.15	.36	.25	.44	.27	.45

Table 11

Descriptive statistics of SPM Set D

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
D1	.98	.14	1.00	.00	.99	.10
D2	.86	.35	.88	.33	.94	.23
D3	.83	.38	.88	.33	.92	.29
D4	.82	.39	.88	.33	.90	.30
D5	.87	.34	.92	.27	.93	.26
D6	.76	.43	.84	.37	.83	.37
D7	.74	.44	.80	.40	.76	.43
D8	.70	.46	.82	.39	.65	.48
D9	.68	.47	.75	.44	.65	.48
D10	.64	.48	.71	.46	.73	.45
D11	.29	.46	.29	.46	.29	.45
D12	.07	.26	.08	.27	.10	.30

Table 12

Descriptive statistics of SPM Set E

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
E1	.75	.43	.76	.43	.82	.38
E2	.62	.49	.71	.46	.73	.45
E3	.53	.50	.63	.49	.69	.46
E4	.38	.49	.41	.50	.45	.50
E5	.38	.49	.43	.50	.60	.49
E6	.42	.50	.55	.50	.44	.50
E7	.36	.48	.43	.50	.37	.49
E8	.24	.43	.33	.48	.43	.50
E9	.20	.40	.25	.44	.24	.43
E10	.11	.31	.14	.35	.20	.41
E11	.05	.22	.04	.20	.09	.29
E12	.02	.14	.04	.20	.06	.25

Mill Hill Vocabulary Scale

The item means and standard deviations for the MHV are presented in Tables 13 and 14. The Reference group scored considerably higher than the Focal and Eligible groups on the majority of items. Although the differences between means were smaller, the Eligible group tended to score higher than the Focal group.

Table 13

Descriptive statistics of MHV Set A

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
MA1	1.00	.000	1.00	.000	1.00	.00
MA2	.58	.50	.67	.48	.97	.17
MA3	.97	.17	1.00	.00	.99	.10
MA4	.85	.36	.94	.24	.97	.17
MA5	.34	.47	.45	.50	.93	.26
MA6	.36	.48	.49	.50	.78	.42
MA7	.54	.50	.65	.48	.90	.30
MA8	.56	.50	.61	.49	.93	.26
MA9	.64	.48	.80	.40	.95	.21
MA10	.14	.35	.22	.42	.43	.50
MA11	.35	.49	.45	.50	.81	.40
MA12	.25	.43	.41	.50	.61	.49
MA13	.19	.39	.24	.43	.31	.47
MA14	.51	.50	.69	.47	.79	.41
MA15	.15	.36	.18	.39	.18	.38
MA16	.24	.43	.27	.45	.53	.50
MA17	.21	.41	.22	.42	.19	.39
MA18	.28	.45	.31	.47	.59	.49

Table 13 continued.

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
MA19	.29	.45	.25	.44	.13	.34
MA20	.13	.34	.20	.40	.20	.40
MA21	.15	.36	.18	.39	.07	.26
MA22	.27	.44	.31	.47	.46	.50
MA23	.16	.37	.22	.42	.28	.45
MA24	.12	.33	.22	.42	.38	.49
MA25	.26	.44	.27	.45	.29	.45
MA26	.25	.43	.27	.45	.20	.40
MA27	.20	.40	.16	.37	.13	.34
MA28	.10	.30	.10	.30	.09	.29
MA29	.23	.42	.25	.44	.16	.37
MA30	.21	.41	.25	.44	.18	.38
MA31	.09	.29	.00	.27	.11	.32
MA32	.16	.37	.12	.33	.09	.29
MA33	.16	.37	.16	.37	.06	.25
MA34	.19	.39	.16	.37	.14	.35

Table 14

Descriptive statistics of MHV Set B

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
MB1	1.00	.00	1.00	.00	1.00	.00
MB2	.69	.46	.80	.40	.88	.33
MB3	.83	.38	.92	.27	.94	.25
MB4	.34	.47	.45	.50	.62	.49
MB5	.64	.48	.75	.44	.93	.26
MB6	.56	.50	.71	.46	.81	.39
MB7	.62	.49	.80	.40	.88	.33
MB8	.55	.50	.73	.45	.94	.25
MB9	.45	.50	.53	.50	.77	.42
MB10	.44	.50	.57	.50	.86	.35
MB11	.47	.50	.55	.50	.91	.29
MB12	.26	.44	.37	.49	.63	.49
MB13	.28	.45	.33	.48	.69	.46
MB14	.11	.31	.20	.41	.27	.45
MB15	.24	.43	.33	.48	.31	.47
MB16	.33	.47	.39	.49	.48	.50
MB17	.13	.34	.12	.33	.17	.37
MB18	.15	.36	.16	.37	.12	.33

Table 14 continued.

	Focal Group (N= 101)		Eligible Group (N= 51)		Reference Group (N=108)	
Scale	Mean	SD	Mean	SD	Mean	SD
MB19	.13	.34	.16	.37	.30	.46
MB20	.13	.34	.16	.37	.17	.37
MB21	.17	.38	.14	.35	.35	.48
MB22	.25	.43	.31	.47	.52	.50
MB23	.16	.37	.22	.42	.09	.29
MB24	.12	.33	.16	.37	.16	.37
MB25	.18	.38	.24	.43	.19	.40
MB26	.23	.42	.25	.44	.18	.38
MB27	.19	.39	.18	.39	.32	.47
MB28	.19	.40	.25	.44	.15	.36
MB29	.11	.31	.14	.35	.10	.30
MB30	.12	.33	.16	.37	.16	.37
MB31	.11	.31	.16	.37	.20	.40
MB32	.10	.30	.14	.35	.12	.33
MB33	.06	.24	.08	.27	.19	.40
MB34	.21	.40	.16	.37	.08	.28

Reliabilities

Cronbach alphas (α) for each test are presented in Table 15. Overall, the reliabilities found in this study for the CFAT, SPM, MHV and WPT were much lower than then those previously reported. This was true of both the Aboriginal Peoples groups and the Reference group. According to Nunnally and Bernstein (1994) tests should have at a minimum a reliability coefficient of .80. However, when important decisions are to be made with test scores, an internal consistency coefficients of .90 is the minimum with .95 or higher a desirable standard. With a reliability of .90 for the combined group of participants, the CFAT was the only measure to meet the higher standards. Nonetheless, the CFAT PS scale, SPM and MHV did have reliability coefficients greater than .80. The reliabilities of CFAT VS and SA scales and WPT fell below the .80 guideline. However, when all participants are combined into one group the WPT reliability coefficient increased to .87.

There was a considerable difference in the reliability of the VS between both Aboriginal Peoples groups (Focal: $\alpha = .55$; Eligible: $\alpha = .53$) and the Reference group ($\alpha = .78$). Deleting six items form the VS scale would increase the reliabilities for the Focal and Eligible group to .65 and .60, respectively; however, deleting the six items causes the reliabilities for the Reference and combined groups to drop.

The reliability coefficient for the SA scale ranged from .51 to .64. Deleting four items from the SA scale would only increase reliability to .56 for the Eligible group; however, deleting the four items causes the reliability to drop substantially for the Focal, Reference and combined groups.

Table 15

Alpha Reliabilities for the CFAT, SPM, MHV and WPT

Scale	Subscales	Reliability (α) Focal Group (n= 101)	Eligible Group (n = 51)	Reference Group (n = 108)	Combined
CFAT	Verbal Skill	.55	.53	.78	.78
	Spatial Ability	.64	.51	.63	.64
	Problem Solving	.82	.82	.80	.88
	Total	.85	.82	.85	.90
SPM		.88	.87	.87	.87
MHV		.85	.85	.82	.88
WPT		.79	.75	.75	.87

Comparison of Aboriginal and Reference Groups

Multivariate analyses of variances were used to determine if there were significant differences between Focal and Reference groups, and Eligible and Reference groups with respect to the following set of dependent variables: CFAT VS scale, CFAT SA scale, CFAT PS scale, CFAT total, WPT, SPM and MHV.

Focal Group vs. Reference Group

There was a significant multivariate difference (Wilks' Lambda = .52, F(1,209) = 30.86, p = .00) between the Focal and Reference groups. Subsequently, univariate F-tests

showed that the between group differences for all dependent measures were significant (see Table 16).

Table 16

Analysis of Variances of CFAT, SPM, MHV and WPT means for Focal and Reference Group

Source	SS	DF	MS	F	P
CFAT Verbal Skill	761.47	1	761.47	80.54	.000
CFAT Spatial Ability	36.71	1	36.71	4.76	.030
CFAT Problem Solving	3800.58	1	3800.58	131.53	.000
CFAT Total	9082.51	1	9082.51	116.39	.000
SPM	378.24	1	378.24	7.55	.007
MHV	4898.48	1	4898.48	76.51	.000
WPT	4111.27	1	4111.27	132.71	.000

Eligible Group vs. Reference Group

There was a significant multivariate difference (Wilks' Lambda = .66, F(1,159) = 13.29, p = .00) between the Eligible and Reference groups. Univariate F-tests showed significant between group differences for all dependent measures except for the CFAT SA and SPM (see Table 17).

Table 17

Analysis of Variances of CFAT, SPM, MHV and WPT means for Eligible and Reference Group

Source	SS	DF	MS	F	P
CFAT Verbal Skill	256.20	1	256.20	23.98	.000
CFAT Spatial Ability	15.00	1	15.00	2.22	.138
CFAT Problem Solving	1114.73	1	1114.73	37.62	.000
CFAT Total	2072.18	1	2072.18	27.38	.000
SPM	3.01	1	3.01	.07	.796
MHV	1149.46	1	1149.46	19.67	.000
WPT	1357.43	1	1357.43	48.11	.000

Testing individuals in their second language may confound test results of verbal measures. Consequently, separate univariate analyses of covariance were performed for each test to determine if the difference in test means between groups were significant after controlling for language (see Table 18). There were significant differences between the Eligible and Reference groups on all tests except for the CFAT SA, SPM and WPT.

Table 18

Analysis of Covariance of CFAT, SPM, MHV and WPT means for Eligible and Reference Group

Source	SS	DF	MS	F	P
CFAT Verbal Skill	164.14	1	164.14	15.27	.000
CFAT Spatial Ability	.26	1	.26	.04	.844
CFAT Problem Solving	526.89	1	526.89	17.84	.000
CFAT Total	1243.17	1	1243.17	16.33	.000
SPM	6.41	1	6.41	.14	.706
MHV	457.62	1	457.62	7.87	.006
WPT	893.49	1	893.49	31.48	.168

MANCOVA

A multivariate analysis of covariance (MANCOVA) was conducted to determine if there were statistically reliable mean differences on the SPM, MHV and WPT between the Focal and Reference groups after adjusting for differences on education and CFAT total score. The results of the MANCOVA (Wilks' Lambda = .95, $F(3,202) = 3.57$, $p = .02$) indicated the presence of a significant multivariate difference. Subsequently, univariate F-ratios were obtained for all three dependent variables (see Table 19). After controlling for education and performance on the CFAT, there was still a significant difference in performance on the WPT. However, there were no significant differences between the Focal and Reference groups on the SPM and MHV. These results justify the

creation of the Eligible group. The differences in education and CFAT score do confound the results.

Table 19

Analysis of Variance of SPM, MHV and WPT means for Focal and Reference Group

Source	SS	DF	MS	F	P
SPM	134.03	1	134.03	3.58	.060
MHV	65.77	1	65.77	1.50	.222
WPT	88.90	1	88.90	6.04	.015

Correlation Analysis

Tables 20 through 23 present the correlations between the tests for all three groups, as well as the groups combined. For the most part, correlations among the different tests were similar across the groups. However, the correlations for the Reference group tended to be weaker than the Focal and Eligible group.

Correlations between the CFAT and WPT (Combined: $r = .82$; Focal: $r = .72$; Eligible: $r = .76$; Reference: $r = .71$) indicated a strong relationship between the two. The CFAT and SPM displayed a weak to moderate association for the Reference group ($r = .47$) and for both Aboriginal Peoples groups (Focal: $r = .61$; Eligible: $r = .68$). The MHV displayed a moderate to strong relationship to the CFAT for the Reference group ($r = .50$) and for both Aboriginal Peoples groups (Focal: $r = .73$; Eligible: $r = .76$). The CFAT VS scale displayed a moderate to strong correlation with the MHV (Combined: $r = .73$;

Focal: $r = .65$; Eligible: $r = .65$; Reference: $r = .65$), while the CFAT SA demonstrated a moderate correlation with the SPM (Combined: $r = .47$; Focal: $r = .49$; Eligible: $r = .40$; Reference: $r = .42$).

Correlations between the SPM and WPT among the Focal and Reference groups were moderately weak (Focal: $r = .46$; Reference: $r = .42$). In contrast, the correlation between the SPM and WPT for the Eligible group was considerably higher ($r = .61$). However, the differences in correlations among the groups were not significant ($z = 1.20$, $p = .11$; Steel, Torrie, & Dickey, 1997). There was a large difference in correlations between the SPM and MHV among both Aboriginal Peoples groups (Focal: $r = .53$; Eligible: $r = .57$) and the Reference group ($r = .13$). The differences in correlations for the Aboriginal groups and the Reference group were significant (Focal and Reference: $z = 3.82$, $p = .00$; Eligible and Reference: $z = 2.82$, $p = .00$). A similar pattern appeared among the correlations between the WPT and MHV. The WPT and MHV displayed a strong correlation among the Aboriginal Peoples groups (Focal: $r = .76$; Eligible: $r = .74$) and only a moderate correlation with the Reference group ($r = .43$). The differences in correlations for the Aboriginal groups and the Reference group were significant (Focal and Reference: $z = 3.28$, $p = .00$; Eligible and Reference: $z = 2.97$, $p = .00$).

Table 20

Correlations Among Measures – Combined Group (N=209)

	1	2	3	4	5	6	7
1. CFAT	1.00						
2. VS	.80**	1.00					
3. SA	.60*	.30**	1.00				
4. PS	.93**	.64**	.39**	1.00			
5. SPM	.53**	.35**	.47**	.48**	1.00		
6. MHV	.73**	.73**	.34**	.64**	.38**	1.00	
7. WPT	.82**	.71**	.41**	.77**	.49**	.74**	1.00

**$p < .01$ (2-tailed)

Table 21

Correlations Among Measures – Focal Group (N=101)

	1	2	3	4	5	6	7
1. CFAT	1.00						
2. VS	.75**	1.00					
3. SA	.67**	.30**	1.00				
4. PS	.92**	.60**	.42**	1.00			
5. SPM	.61**	.41**	.49**	.54**	1.00		
6. MHV	.73**	.65**	.45**	.64**	.53**	1.00	
7. WPT	.72**	.64**	.39**	.66**	.46**	.76**	1.00

**$p < .01$ (2-tailed)

Table 22

Correlations Among Measures - Eligible Group (N=51)

	1	2	3	4	5	6	7
1. CFAT	1.00						
2. VS	.76**	1.00					
3. SA	.47**	.22	1.00				
4. PS	.91**	.55**	.16	1.00			
5. SPM	.68**	.50**	.40**	.58**	1.00		
6. MHV	.76**	.65**	.30*	.68**	.57**	1.00	
7. WPT	.76**	.62**	.30*	.70**	.61**	.74**	1.00

**$p < .01$ (2-tailed), * $p < .05$ (2-tailed)

Table 23

Correlations Among Measures –Reference Group (N=108)

	1	2	3	4	5	6	7
1. CFAT	1.00						
2. VS	.71**	1.00					
3. SA	.63**	.26**	1.00				
4. PS	.86**	.39**	.35**	1.00			
5. SPM	.47**	.24*	.42**	.41**	1.00		
6. MHV	.45**	.65**	.15	.24*	.13	1.00	
7. WPT	.71**	.55**	.43**	.60**	.42**	.43**	1.00

**$p < .01$ (2-tailed), * $p < .05$ (2-tailed)

DIF Analysis

The items of the CFAT, WPT, SPM and MHV were analyzed for the presence of DIF. For an item to be classified as displaying DIF, the p-value for the two-degree of freedom χ^2 in logistical regression had to be ≤ 0.01 (Robie et al., 2001; Zumbo, 1999). Furthermore, the effect size (R^2) had to be > 0.130 (Zumbo & Thomas, 1997).

CFAT Verbal Skill (VS)

All of the items on the CFAT VS scale were examined for the presence of DIF. The results of the logistical regression analysis are presented in Table 24. Two items from the CFAT VS scale displayed DIF. The χ^2 (2-df) p-value for item VS4 was .00 and the effect size was .17. Furthermore, the difference in R-squared from step 2 to step 3 (.02) was quite small indicating that the DIF was uniform in nature (see Figure 4). Item VS9 had a χ^2 (2-df) p-value of .00 and an effect size of .21. The difference in R-squared from step 2 to step 3 (.00) was quite small indicating that the DIF was uniform in nature (see Figure 5). The ICC in Figure 5 indicates that that item VS9 may be biased towards Aboriginal Peoples.

The two DIF items were omitted from their respective sets and a "purifying" DIF analysis was conducted on the remaining data. The result of the second logistical regression is presented in Table 25. No new items were identified as displaying DIF during the "purifying" DIF analysis.

Table 24

DIF Analysis of CFAT Verbal Skill (VS) Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
VS1	.55	.63	.64	5.78 (.06)	.09	No
VS2	.69	.71	.74	4.45 (.11)	.06	No
VS3	.64	.69	.67	4.41 (.11)	.04	No
VS4	.65	.80	.82	11.74 (.00)	.17	Yes
VS5	.64	.65	.70	3.61 (.17)	.06	No
VS6	.54	.57	.57	1.39 (.50)	.04	No
VS7	.65	.66	.66	.74 (.69)	.01	No
VS8	.63	.63	.65	1.51 (.47)	.02	No
VS9	.16	.36	.36	16.48 (.00)	.21	Yes
VS10	.70	.72	.72	1.24 (.54)	.02	No
VS11	.62	.66	.69	4.47 (.11)	.07	No
VS12	.59	.65	.65	3.67 (.16)	.06	No
VS13	.70	.73	.73	2.12 (.35)	.04	No
VS14	.35	.39	.39	2.50 (.29)	.04	No
VS15	.31	.43	.45	5.21 (.07)	.14	No

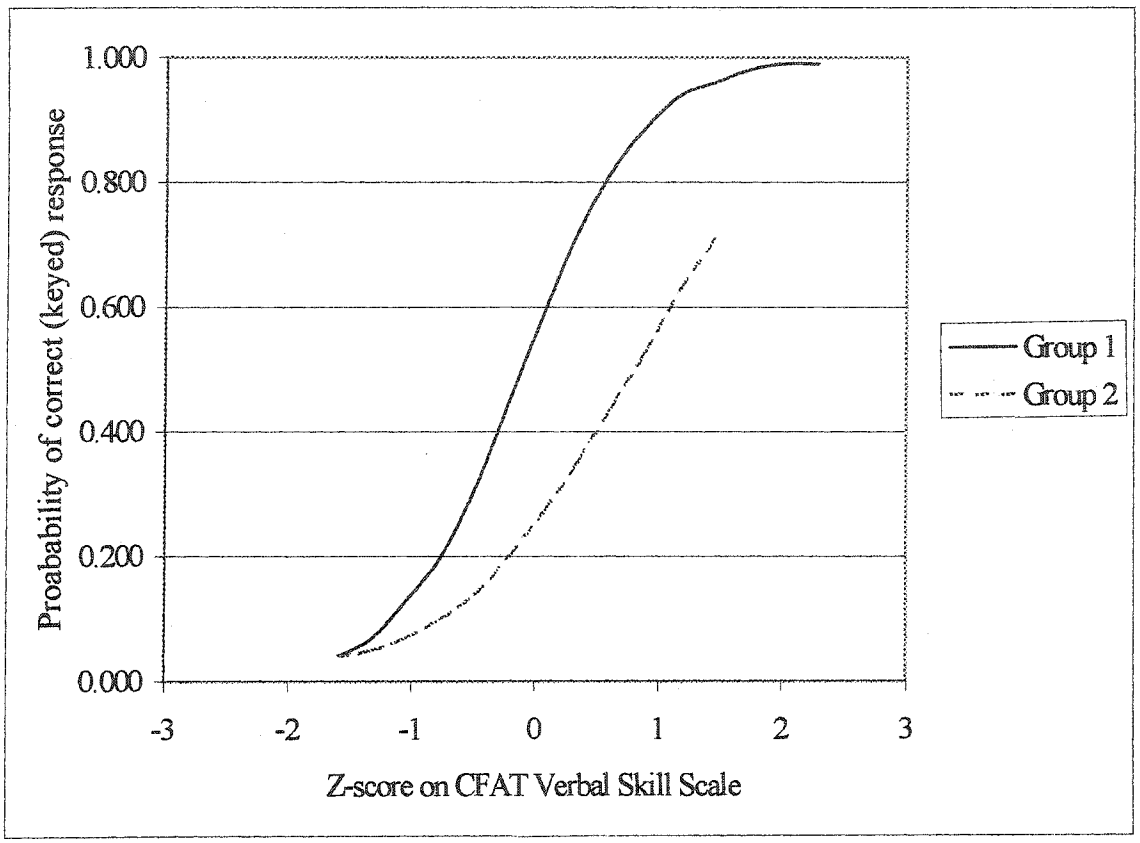

Figure 4. ICC of CFAT VS item 4 displaying uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group.

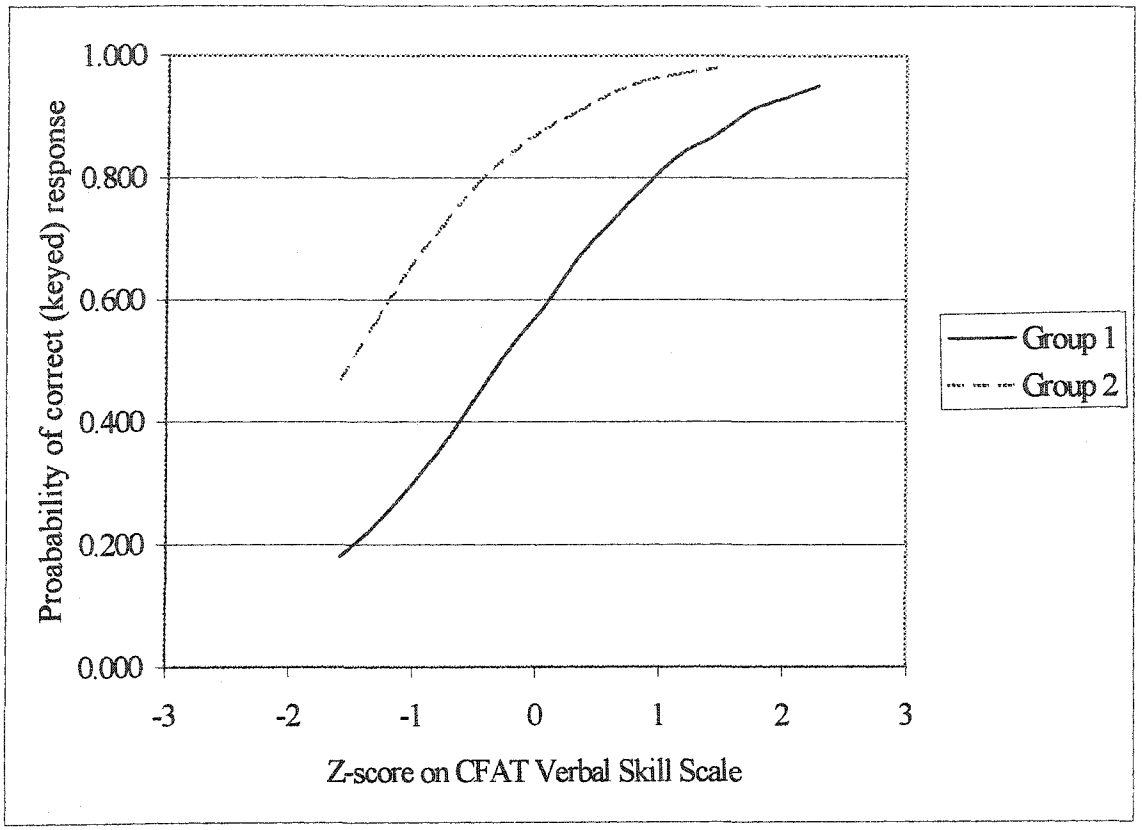

Figure 5. ICC of CFAT VS item 9 displaying uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group.

Table 25

DIF Analysis of CFAT Verbal Skill (VS) Subscale with DIF Items Omitted

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
VS1	.58	.68	.69	7.15 (.03)	.11	No
VS2	.76	.78	.81	4.08 (.13)	.06	No
VS3	.71	.78	.78	5.09 (.08)	.08	No
VS5	.64	.64	.69	3.70 (.16)	.06	No
VS6	.73	.78	.79	1.64 (.44)	.06	No
VS7	.67	.68	.68	.79 (.67)	.01	No
VS8	.68	.68	.70	1.12 (.57)	.02	No
VS10	.62	.63	.64	1.51 (.47)	.02	No
VS11	.68	.74	.78	5.61 (.06)	.10	No
VS12	.63	.69	.70	3.93 (.14)	.07	No
VS13	.69	.72	.73	2.29 (.32)	.04	No
VS14	.31	.34	.35	2.50 (.29)	.04	No
VS15	.28	.38	.40	6.17 (.05)	.13	No

CFAT Spatial Ability (SA)

All of the items on the CFAT SA scale were examined for the presence of DIF. The results of the logistical regression analysis are presented in Table 26. None of the CFAT SA scale items displayed DIF.

CFAT Problem Solving (PS)

All of the items on the CFAT PS scale were examined for the presence of DIF. The results of the logistical regression analysis are presented in Table 27. None of the CFAT PS scale items displayed DIF.

Table 26

DIF Analysis of CFAT Spatial Ability (SA) Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
SA1	.47	.47	.47	.04 (.98)	.00	No
SA2	.27	.28	.28	.89 (.64)	.01	No
SA3	.61	.63	.74	7.01 (.03)	.13	No
SA4	.67	.67	.67	.01 (1.00)	.00	No
SA5	.70	.74	.74	1.80 (.41)	.04	No
SA6	.63	.67	.75	7.01 (.03)	.12	No
SA7	.68	.68	.69	.57 (.75)	.01	No
SA8	.40	.53	.54	8.21 (.02)	.14	No
SA9	.59	.60	.60	.81 (.67)	.01	No
SA10	.41	.43	.44	.94 (.63)	.03	No
SA11	.69	.77	.77	5.28 (.07)	.09	No
SA12	.51	.52	.54	1.04 (.60)	.03	No
SA13	.42	.45	.45	1.91 (.39)	.03	No
SA14	.05	.07	.08	1.03 (.60)	.03	No
SA15	.54	.56	.56	1.49 (.48)	.03	No

Table 27

DIF Analysis of CFAT Problem Solving (PS) Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
PS1	.49	.50	.53	3.44 (.18)	.04	No
PS2	.53	.53	.59	3.75 (.15)	.06	No
PS3	.41	.41	.41	.03 (.99)	.00	No
PS5	.51	.51	.54	2.44 (.30)	.03	No
PS5	.50	.51	.53	2.01 (.37)	.02	No
PS6	.62	.62	.62	.18 (.92)	.00	No
PS7	.39	.44	.44	4.68 (.10)	.05	No
PS8	.58	.58	.59	1.04 (.60)	.02	No
PS9	.54	.54	.62	7.46 (.02)	.09	No
PS10	.30	.30	.33	4.14 (.13)	.03	No
PS11	.38	.39	.39	.45 (.80)	.01	No
PS12	.47	.58	.60	12.27 (.00)	.13	No
PS13	.42	.44	.44	1.87 (.39)	.02	No
PS14	.49	.50	.50	.98 (61)	.01	No
PS15	.37	.38	.38	.44 (.80)	.01	No
PS16	.36	.37	.39	3.15 (.21)	.03	No
PS17	.35	.35	.41	4.45 (.11)	.06	No
PS18	.46	.51	.53	5.60 (.06)	.07	No

Table 27 continued.

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
PS19	.38	.38	.47	8.77 (.01)	.10	No
PS20	.50	.48	.50	.782 (.68)	.01	No
PS21	.44	.45	.47	2.77 (.25)	.03	No
PS22	.46	.46	.48	1.63 (.44)	.02	No
PS23	.51	.54	.55	3.51 (.17)	.04	No
PS24	.40	.40	.51	10.77 (.01)	.12	No
PS25	.49	.49	.50	.93 (.63)	.01	No
PS26	.40	.42	.42	1.94 (.38)	.02	No
PS27	.10	.17	.18	6.49 (.04)	.08	No
PS28	.43	.43	.45	1.21 (.55)	.02	No
PS29	.34	.35	.38	3.05 (.22)	.04	No
PS30	.43	.44	.45	1.62 (.44)	.02	No

Wonderlic Personnel Test

All of the items on the WPT scale were examined for the presence of DIF. The results of the logistical regression analysis are presented in Table 28. Two items did not have frequencies greater than one and therefore could not be analysed (WPT45, WPT48). None of the WPT scale items displayed DIF.

Table 28

DIF Analysis of WPT

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
WPT1	.07	.09	.14	7.93 (.02)	.08	No
WPT2	.27	.27	.28	3.71 (.16)	.00	No
WPT3	.22	.23	.23	.08 (.96)	.00	No
WPT4	.32	.33	.33	.15 (.93)	.00	No
WPT5	.50	.51	.51	1.30 (.52)	.02	No
WPT6	.33	.33	.33	.43 (.81)	.00	No
WPT7	.20	.21	.22	.95 (.62)	.02	No
WPT8	.36	.37	.43	7.63 (.02)	.07	No
WPT9	.23	.26	.30	4.46 (.11)	.07	No
WPT10	.36	.40	.40	5.76 (.06)	.05	No
WPT11	.55	.55	.55	.36 (.84)	.00	No
WPT12	.41	.41	.41	.85 (.65)	.01	No
WPT13	.29	.29	.32	2.39 (.30)	.03	No
WPT14	.45	.46	.47	1.79 (.41)	.02	No
WPT15	.39	.40	.42	2.33 (.31)	.03	No
WPT16	.20	.20	.24	5.15 (.08)	.04	No
WPT17	.45	.46	.49	3.62 (.16)	.04	No
WPT18	.52	.52	.52	.13 (.94)	.00	No

Table 28 continued.

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
WPT19	.26	.29	.32	4.04 (.13)	.06	No
WPT20	.44	.51	.52	5.77 (.06)	.08	No
WPT21	.37	.43	.46	14.01 (.00)	.10	No
WPT22	.04	.11	.11	3.92 (.14)	.07	No
WPT23	.46	.50	.50	3.44 (.18)	.04	No
WPT24	.34	.35	.36	1.97 (.37)	.03	No
WPT25	.45	.48	.48	2.53 (.28)	.03	No
WPT26	.45	.47	.51	4.56 (.10)	.05	No
WPT27	.34	.36	.38	3.83 (.15)	.05	No
WPT28	.32	.34	.43	7.94 (.02)	.11	No
WPT29	.44	.45	.46	1.37 (.50)	.02	No
WPT30	.52	.55	.55	2.30 (.32)	.03	No
WPT31	.30	.30	.31	.30 (.86)	.01	No
WPT32	.28	.33	.33	3.24 (.20)	.05	No
WPT33	.20	.20	.22	1.33 (.52)	.03	No
WPT34	.42	.44	.45	1.33 (.52)	.02	No
WPT35	.39	.39	.39	.38 (.83)	.00	No
WPT36	.43	.48	.49	2.83 (.24)	.06	No
WPT37	.54	.54	.54	1.17 (.56)	.00	No

Table 28 continued.

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
WPT38	.37	.38	.40	2.02 (.37)	.03	No
WPT39	.22	.27	.29	3.38 (.19)	.06	No
WPT40	.18	.19	.23	3.39 (.18)	.05	No
WPT41	.28	.30	.30	1.65 (.44)	.02	No
WPT42	.28	.30	.30	1.79 (.41)	.02	No
WPT43	.38	.36	.39	.35 (.84)	.01	No
WPT44	.02	.03	.05	1.48 (.48)	.04	No
WPT45	n/a	n/a	n/a	n/a	n/a	n/a
WPT46	.16	.17	.17	.58 (.75)	.02	No
WPT47	.03	.04	.05	.80 (.67)	.01	No
WPT48	n/a	n/a	n/a	n/a	n/a	n/a
WPT49	.07	.07	.07	4.64 (.10)	.00	No
WPT50	.09	.09	.09	2.76 (.25)	.00	No

n/a: These values cannot be calculated because there was no variance in the item response patterns.

Raven's Standard Progressive Matrices

All of the items on the SPM were examined for the presence of DIF using logistical regression. The results of the DIF analyses are presented in Tables 29 through 33. Several items did not have frequencies greater than one and therefore could not be analysed (A5, A6, A9, D1).

None of the items in Set A, B, or C displayed DIF. However, one item in Set D and one item in Set E displayed DIF. The χ^2 (2-df) p-value for item D11 was .00 and the effect size was .19. Furthermore, the difference in R-squared from step 2 to step 3 (.19) was quite large indicating that the DIF was non-uniform in nature (see Figure 6). Item E10 had a χ^2 (2-df) p-value of .01 and an effect size of .15. The difference in R-squared from step 2 to step 3 (.14) was quite large indicating that the DIF was non-uniform in nature (see Figure 7).

The two DIF items were omitted from their respective sets and a "purifying" DIF analysis was conducted on the remaining data. The results of the second logistical regression for sets D and E are presented in Tables 34 and 35. No new items were identified as displaying DIF during the "purifying" DIF analysis.

Table 29

DIF Analysis of SPM A Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
A1	.14	.20	.34	5.64 (.06)	.20	No
A2	.02	.02	.02	8.62 (.01)	.01	No
A3	.02	.02	.02	8.62 (.01)	.01	No
A4	.02	.02	.02	8.62 (.01)	.01	No
A5	n/a	n/a	n/a	n/a	n/a	n/a
A6	n/a	n/a	n/a	n/a	n/a	n/a
A7	.21	.22	.23	.75 (.69)	.02	No
A8	.19	.19	.22	1.42 (.49)	.02	No
A9	n/a	n/a	n/a	n/a	n/a	n/a
A10	.27	.27	.29	.84 (.66)	.03	No
A11	.14	.14	.18	4.39 (.11)	.05	No
A12	.35	.35	.35	.02 (.99)	.00	No

n/a: These values cannot be calculated because there was no variance in the item response patterns.

Table 30

DIF Analysis of SPM B Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
B1	.33	.38	.39	1.07 (.59)	.06	No
B2	.43	.44	.49	1.08 (.58)	.06	No
B3	.34	.81	.82	6.04 (.05)	.48	No
B4	.11	.12	.24	2.04 (.36)	.13	No
B5	.89	.89	.89	.04 (.98)	.00	No
B6	.55	.58	.58	1.38 (.50)	.03	No
B7	.55	.56	.62	4.47 (.11)	.07	No
B8	.61	.61	.64	2.20 (.33)	.03	No
B9	.77	.77	.77	.28 (.87)	.01	No
B10	.50	.50	.53	1.86 (.40)	.03	No
B11	.41	.43	.43	2.61 (.27)	.02	No
B12	.44	.44	.46	2.08 (.35)	.02	No

Table 31

DIF Analysis of SPM C Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
C1	.72	.72	.72	3.15 (.21)	.00	No
C2	.57	.58	.65	4.67 (.10)	.08	No
C3	.65	.65	.67	.59 (.75)	.02	No
C4	.33	.46	.47	7.03 (.03)	.14	No
C5	.37	.46	.46	1.73 (.42)	.10	No
C6	.68	.69	.78	3.73 (.16)	.10	No
C7	.75	.77	.79	.94 (.63)	.03	No
C8	.78	.87	.89	6.15 (.05)	.11	No
C9	.75	.79	.79	1.87 (.39)	.04	No
C10	.75	.76	.79	2.56 (.28)	.05	No
C11	.47	.49	.49	2.22 (.33)	.02	No
C12	.10	.11	.12	4.11 (.13)	.02	No

Table 32

DIF Analysis of SPM D Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
D1	n/a	n/a	n/a	n/a	n/a	n/a
D2	.63	.68	.71	3.58 (.17)	.09	No
D3	.72	.75	.75	1.31 (.52)	.03	No
D4	.79	.80	.82	1.21 (.55)	.03	No
D5	.68	.68	.74	3.12 (.21)	.07	No
D6	.63	.63	.64	.68 (.71)	.01	No
D7	.77	.79	.83	4.07 (.13)	.07	No
D8	.48	.54	.57	7.77 (.02)	.09	No
D9	.50	.53	.56	4.92 (.09)	.06	No
D10	.65	.66	.66	.07 (.97)	.00	No
D11	.20	.20	.40	19.39 (.00)	.19	Yes
D12	.21	.21	.29	4.41 (.11)	.08	No

n/a: These values cannot be calculated because there was no variance in the item response patterns.

Table 33

DIF Analysis of SPM E Subscale

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
E1	.13	.13	.14	2.09 (.35)	.01	No
E2	.18	.18	.18	1.56 (.46)	.01	No
E3	.35	.36	.39	6.21 (.05)	.04	No
E4	.53	.53	.56	3.51 (.17)	.03	No
E5	.44	.49	.49	5.49 (.06)	.05	No
E6	.55	.57	.59	4.36 (.11)	.04	No
E7	.52	.54	.54	1.59 (.45)	.02	No
E8	.53	.56	.58	4.43 (.11)	.05	No
E9	.54	.55	.55	.10 (.95)	.00	No
E10	.45	.47	.60	10.05 (.01)	.15	Yes
E11	.28	.29	.35	2.81 (.25)	.07	No
E12	.32	.37	.37	1.40 (.50)	.06	No

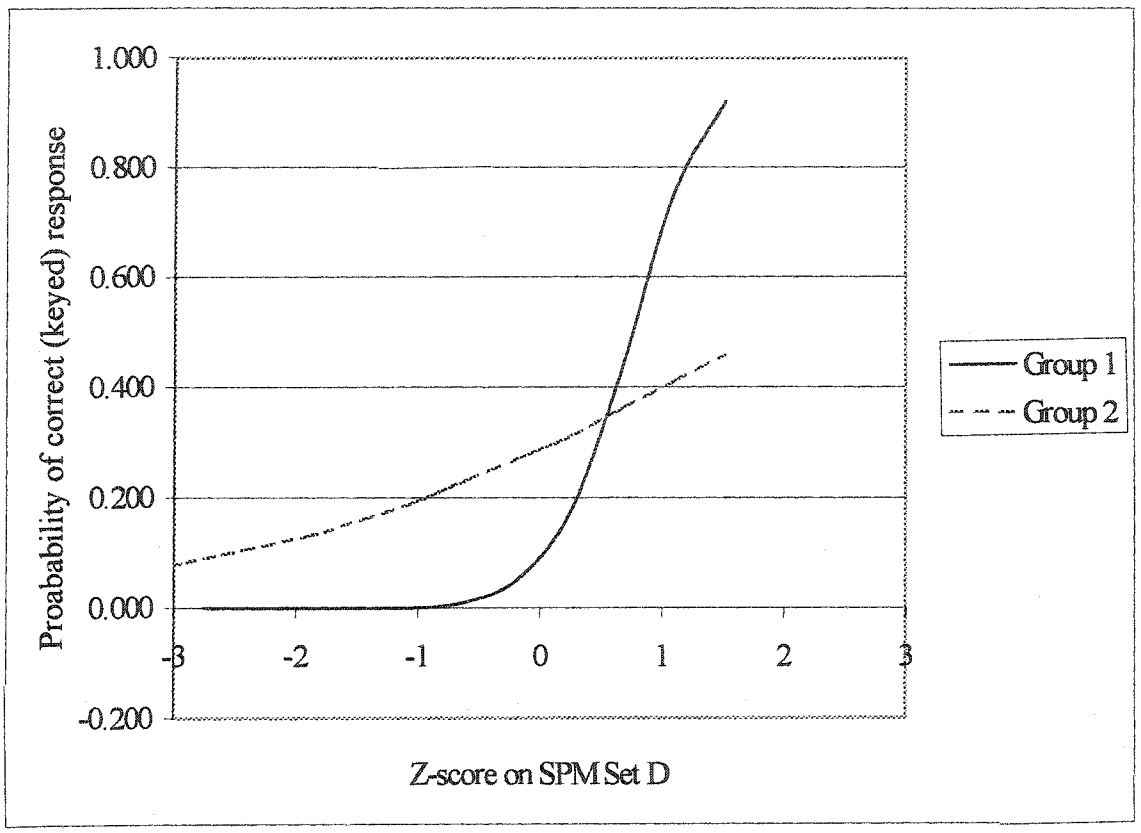

Figure 6. ICC of SPM item D11 displaying non-uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group.

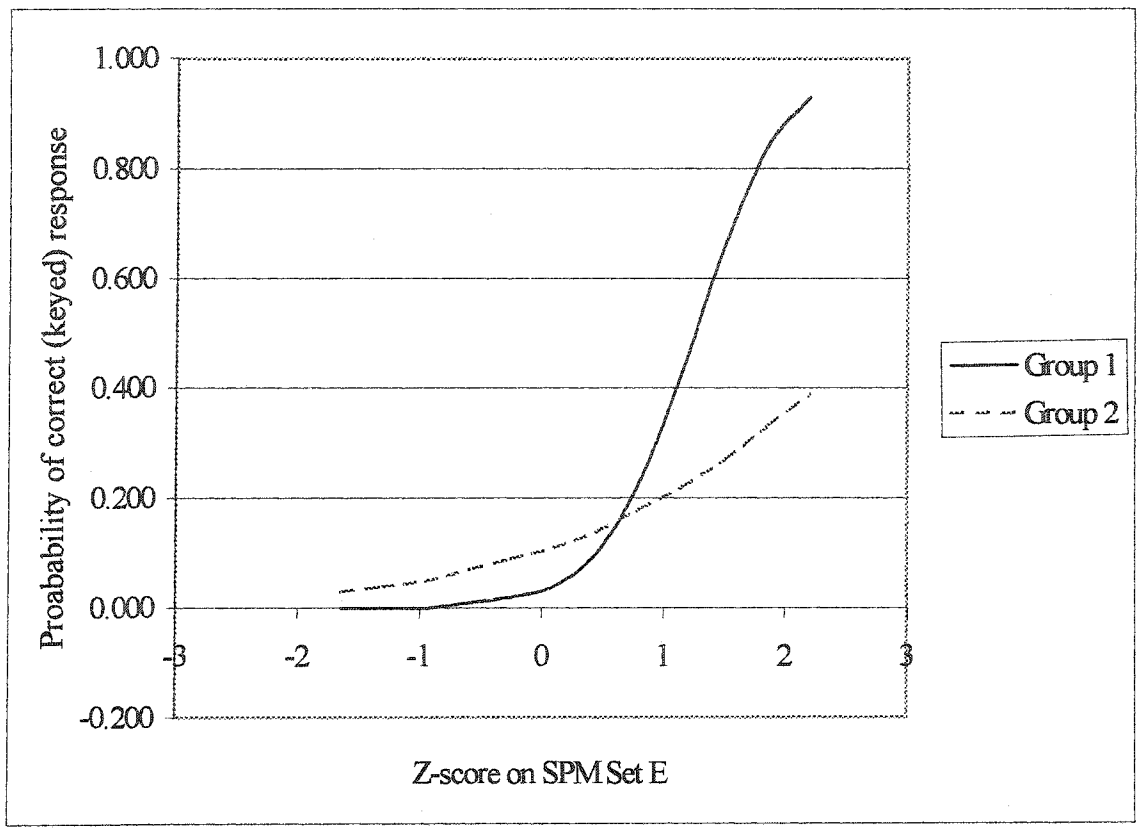

Figure 7. ICC of SPM item E10 displaying non-uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group.

Table 34

DIF Analysis of SPM Set D with DIF Item Omitted

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
D1	n/a	n/a	n/a	n/a	n/a	n/a
D2	.66	.70	.72	2.41 (.30)	.06	No
D3	.73	.75	.75	.87 (.65)	.02	No
D4	.75	.76	.78	1.13 (.57)	.02	No
D5	.64	.64	.70	3.41 (.18)	.07	No
D6	.53	.53	.53	.14 (.93)	.00	No
D7	.67	.71	.76	6.03 (.05)	.07	No
D8	.49	.55	.60	9.43 (.01)	.12	No
D9	.40	.43	.47	6.74 (.03)	.07	No
D10	.57	.57	.57	.01 (.99)	.00	No
D12	.09	.10	.19	9.06 (.01)	.10	No

n/a: These values cannot be calculated because there was no variance in the item response patterns

Table 35

DIF Analysis of SPM Set E with DIF Item Omitted

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
E1	.13	.13	.14	2.25 (.33)	.01	No
E2	.19	.19	.20	1.73 (.42)	.01	No
E3	.33	.34	.37	6.30 (.04)	.04	No
E4	.51	.51	.55	4.28 (.12)	.04	No
E5	.42	.46	.47	5.68 (.06)	.04	No
E6	.56	.57	.61	5.02 (.08)	.05	No
E7	.47	.48	.49	2.06 (.36)	.02	No
E8	.51	.55	.57	5.27 (.07)	.06	No
E9	.51	.51	.51	.31 (.86)	.01	No
E11	.25	.30	.38	3.20 (.20)	.13	No
E12	.24	.30	.30	1.69 (.43)	.07	No

Mill Hill Vocabulary Scale

All of the items on the MHV were examined for the presence of DIF using logistical regression. The results of the DIF analyses are presented in Tables 36 and 37. Three items in Set A displayed DIF. The χ^2 (2-df) p-value for item A2 was .00 and the effect size was .19. Furthermore, the difference in R-squared from step 2 to step 3 (.03) was quite small indicating that the DIF was uniform in nature (see Figure 8). Item A5 had a χ^2 (2-df) p-value of .00 and an effect size of .18. The difference in R-squared from step 2 to step 3 (.00) was quite small indicating that the DIF was uniform in nature (see Figure 9). The χ^2 (2-df) p-value for item A13 was .00 and the effect size was .17. The difference in R-squared from step 2 to step 3 (.17) was quite large indicating that the DIF was non-uniform in nature (see Figure 10). The ICC in Figure 10 indicates that that item A13 may be biased towards Aboriginal Peoples.

Two items in Set B displayed DIF. Item B11 had a χ^2 (2-df) p-value of .00 and an effect size of .18. The difference in R-squared from step 2 to step 3 (.03) was quite small indicating that the DIF was uniform in nature (see Figure 11). The χ^2 (2-df) p-value for item B23 was .01 and the effect size was .13. Furthermore, the difference in R-squared from step 2 to step 3 (.02) was quite small indicating that the DIF was uniform in nature (see Figure 12). The ICC in Figure 12 indicates that that item B23 may be biased towards Aboriginal Peoples.

The five DIF items were omitted from their respective sets and a "purifying" DIF analysis was conducted on the remaining data. The results of the second logistical regression for sets A and B are presented in Tables 38 and 39. No new items were identified as displaying DIF during the "purifying" DIF analysis.

Table 36

DIF Analysis of MHV Set A

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
A2	.32	.49	.52	16.54 (.00)	.19	Yes
A3	.31	.32	.39	1.93 (.38)	.08	No
A4	.34	.34	.39	3.48 (.18)	.05	No
A5	.30	.48	.48	23.18 (.00)	.18	Yes
A6	.42	.43	.44	2.99 (.22)	.03	No
A7	.56	.62	.62	4.53 (.10)	.06	No
A8	.41	.46	.46	5.26 (.07)	.05	No
A9	.56	.62	.64	5.56 (.06)	.07	No
A10	.46	.48	.52	5.05 (.08)	.06	No
A11	.51	.56	.60	8.53 (.01)	.09	No
A12	.39	.40	.48	12.37 (.00)	.09	No
A13	.37	.37	.54	15.47 (.00)	.17	Yes
A14	.64	.65	.65	.11 (.95)	.00	No
A15	.17	.21	.22	2.42 (.30)	.05	No
A16	.59	.59	.59	.30 (.86)	.00	No
A17	.12	.16	.18	1.97 (.37)	.05	No
A18	.58	.58	.59	.66 (.72)	.01	No

Table 36 continued.

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
A19	.05	.17	.17	4.87 (.09)	.12	No
A20	.14	.15	.15	.21 (.90)	.01	No
A21	.01	.05	.08	4.56 (.10)	.08	No
A22	.35	.36	.38	2.95 (.23)	.030	No
A23	.22	.24	.27	10.85 (.00)	.05	No
A24	.12	.12	.12	.08 (.96)	.00	No
A25	.09	.10	.12	.90 (.64)	.02	No
A26	.02	.07	.11	3.25 (.20)	.09	No
A27	.00	.03	.04	1.66 (.44)	.04	No
A28	.06	.08	.09	1.19 (.55)	.04	No
A29	.09	.23	.28	7.40 (.03)	.19	No
A30	.07	.13	.14	2.49 (.29)	.06	No
A31	.21	.22	.22	.33 (.85)	.01	No
A32	.07	.18	.21	6.43 (.04)	.14	No
A33	.00	.18	.18	5.49 (.06)	.18	No
A34	.03	.08	.16	5.70 (.06)	.13	No

Table 37

DIF Analysis of MHV Set B

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
B2	.09	.09	.10	3.45 (.18)	.01	No
B3	.42	.42	.43	.42 (.81)	.01	No
B4	.51	.51	.53	1.54 (.46)	.02	No
B5	.54	.60	.62	4.86 (.09)	.08	No
B6	.72	.73	.74	1.38 (.50)	.02	No
B7	.59	.60	.61	1.37 (.50)	.02	No
B8	.64	.72	.75	9.42 (.01)	.11	No
B9	.57	.61	.61	2.65 (.27)	.04	No
B10	.69	.76	.76	4.98 (.08)	.07	No
B11	.46	.60	.63	15.21 (.00)	.18	Yes
B12	.63	.66	.66	2.22 (.33)	.03	No
B13	.64	.69	.70	4.24 (.12)	.06	No
B14	.42	.42	.42	.18 (.91)	.00	No
B15	.35	.37	.39	1.82 (.40)	.04	No
B16	.44	.45	.45	1.13 (.57)	.02	No
B17	.18	.19	.30	7.32 (.03)	.13	No
B18	.14	.25	.27	4.67 (.10)	.13	No

Table 37 continued.

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
B19	.26	.27	.38	9.78 (.01)	.12	No
B20	.07	.08	.18	10.17 (.01)	.11	No
B21	.25	.25	.37	11.28 (.00)	.12	No
B22	.47	.47	.51	3.45 (.18)	.04	No
B23	.02	.08	.15	9.87 (.01)	.13	Yes
B24	.23	.24	.29	2.80 (.25)	.07	No
B25	.24	.29	.30	2.40 (.30)	.06	No
B26	.14	.31	.32	8.20 (.02)	.18	No
B27	.32	.33	.38	2.44 (.30)	.05	No
B28	.01	.07	.08	2.53 (.28)	.07	No
B29	.01	.01	.01	.19 (.91)	.01	No
B30	.03	.04	.08	1.43 (.49)	.05	No
B31	.33	.33	.35	.58 (.75)	.02	No
B32	.39	.50	.51	5.46 (.07)	.12	No
B33	.34	.35	.35	.71 (.70)	.01	No
B34	.19	.24	.26	2.73 (.26)	.07	No

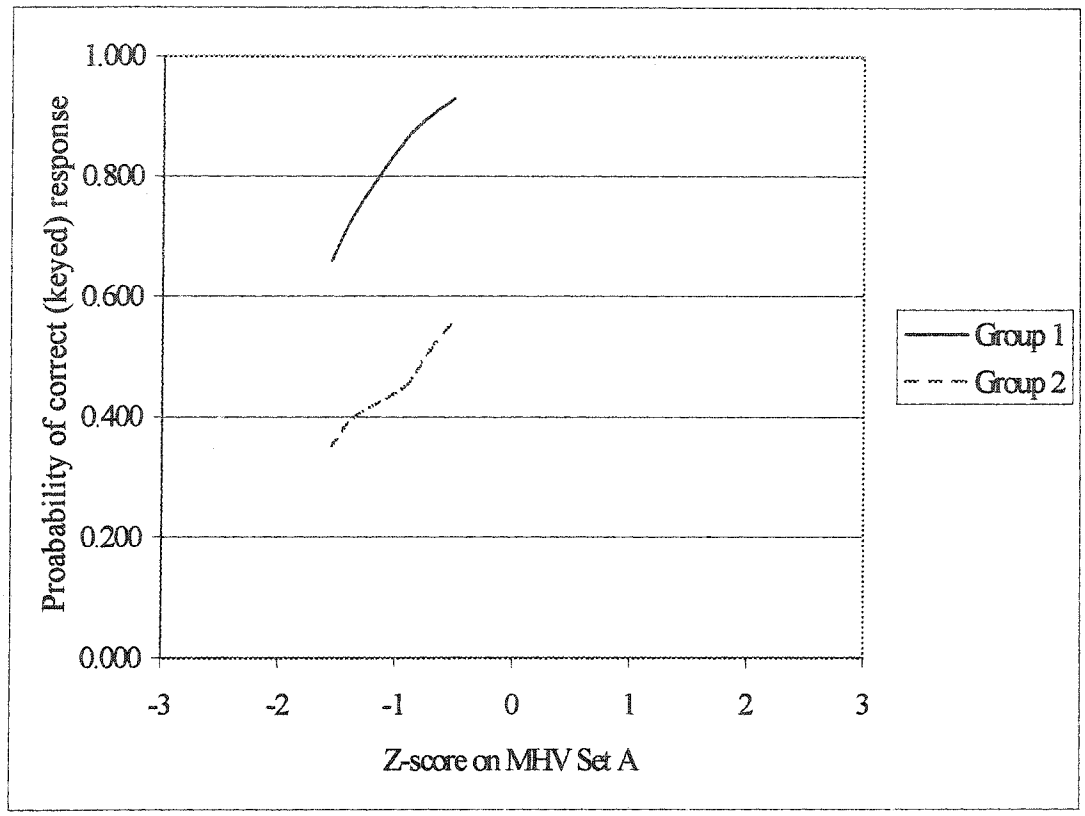

Figure 8. ICC of MHV item A2 displaying uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group

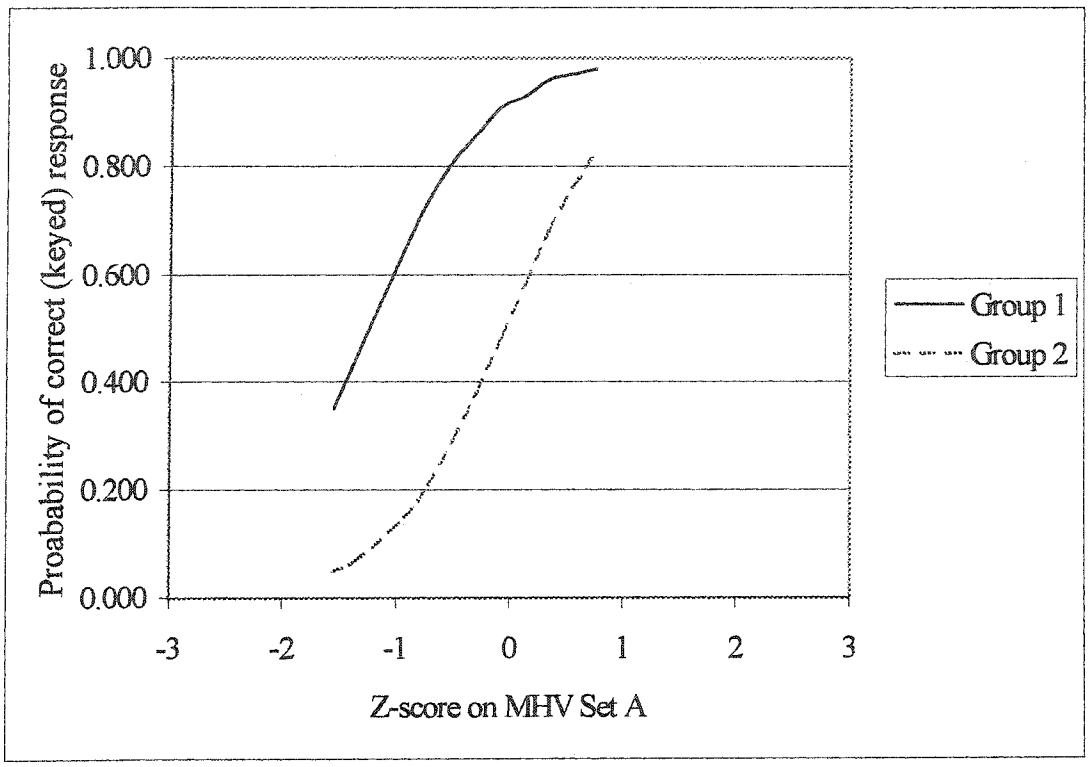

Figure 9. ICC of MHV itemA5 displaying uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group

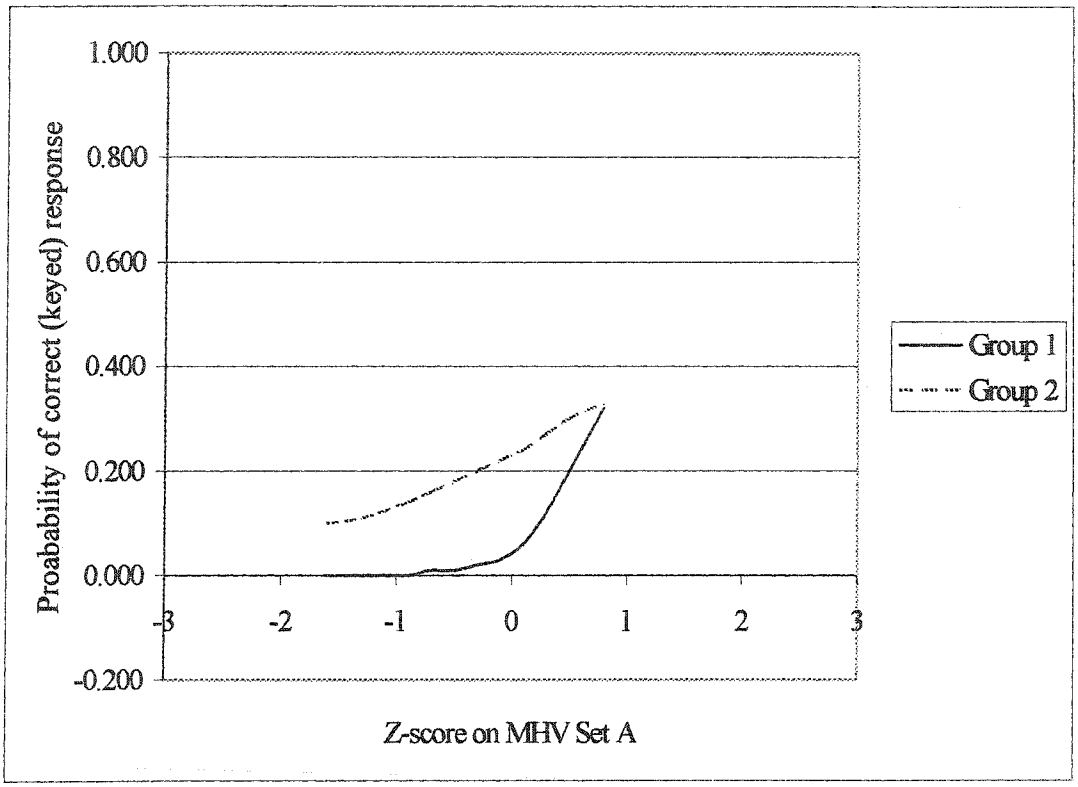

Figure 10. ICC of MHV item A13 displaying non-uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group

Table 38

DIF Analysis of MHV Set A with DIF Items Omitted

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
A3	.29	.30	.34	1.21 (.55)	.05	No
A4	.38	.39	.44	3.16 (.21)	.05	No
A6	.25	.28	.28	5.90 (.05)	.03	No
A7	.42	.51	.51	7.73 (.02)	.09	No
A8	.37	.45	.45	8.85 (.01)	.08	No
A9	.56	.66	.68	8.31 (.02)	.12	No
A10	.47	.51	.53	4.96 (.08)	.06	No
A11	.36	.44	.47	13.24 (.00)	.11	No
A12	.28	.29	.35	12.07 (.00)	.07	No
A14	.57	.58	.58	.31 (.86)	.01	No
A15	.18	.22	.23	2.06 (.36)	.04	No
A16	.64	.65	.65	.78 (.68)	.01	No
A17	.16	.20	.21	1.66 (.44)	.05	No
A18	.64	.66	.66	1.49 (.48)	.02	No
A19	.04	.23	.23	5.93 (.05)	.19	No
A20	.15	.16	.16	.21 (.90)	.01	No
A21	.01	.07	.11	4.57 (.10)	.11	No

Table 38 continued.

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
A22	.43	.43	.46	2.32 (.31)	.03	No
A23	.25	.26	.29	7.36 (.03)	.04	No
A24	.13	.13	.13	.71 (.70)	.00	No
A25	.12	.13	.14	.62 (.74)	.02	No
A26	.02	.06	.09	3.09 (.21)	.07	No
A27	.00	.06	.06	1.83 (.40)	.06	No
A28	.07	.09	.10	1.12 (.57)	.04	No
A29	.10	.21	.25	7.85 (.02)	.16	No
A30	.13	.20	.21	2.67 (.26)	.08	No
A31	.26	.27	.27	.12 (.94)	.01	No
A32	.10	.24	.27	5.91 (.05)	.02	No
A33	.00	.24	.25	6.33 (.04)	.24	No
A34	.04	.09	.16	5.06 (.08)	.12	No

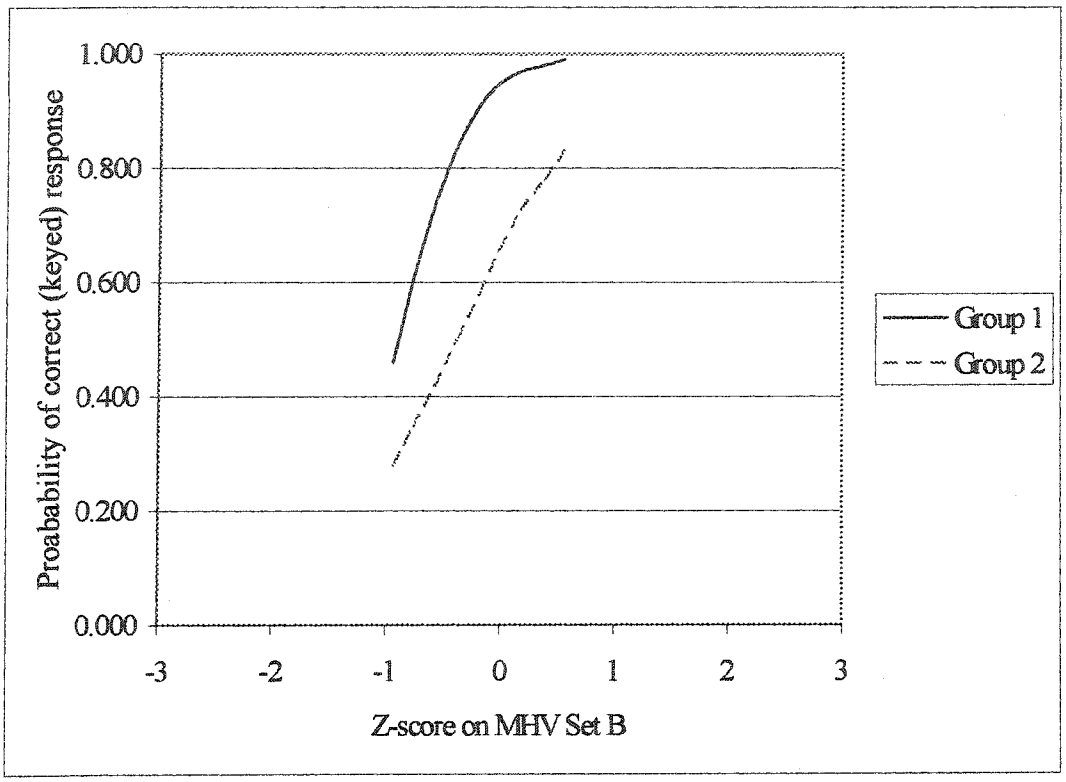

Figure 11. ICC of MHV item B11 displaying uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group

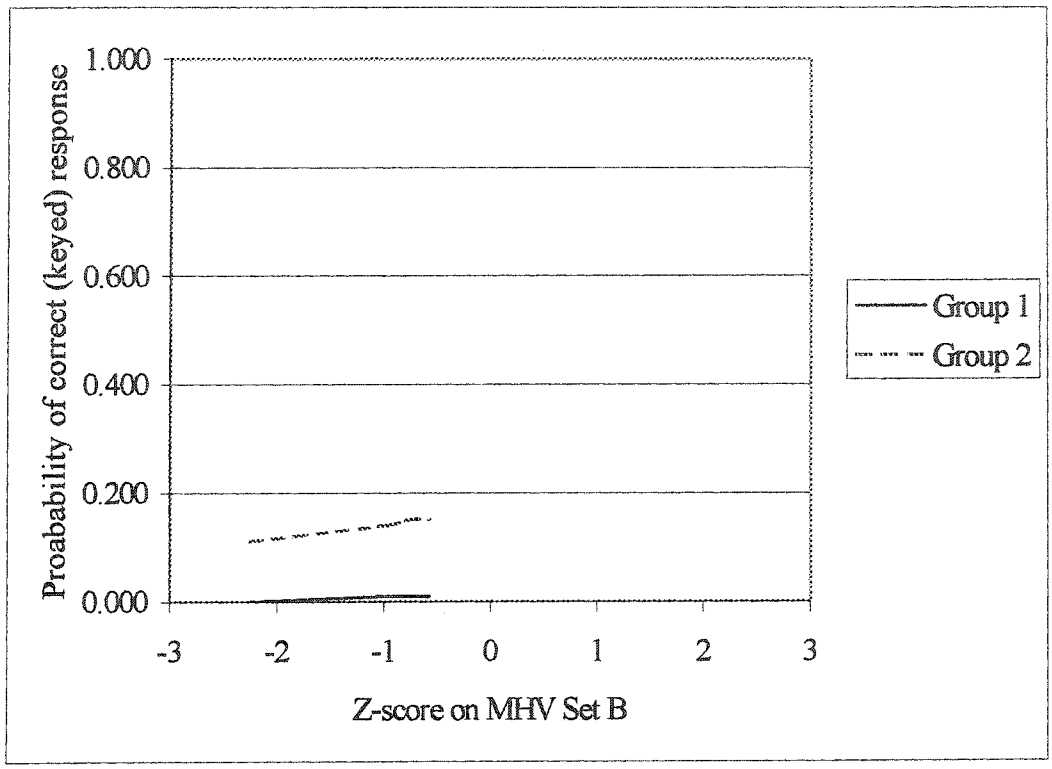

Figure 12. ICC of MHV item B23 displaying uniform DIF.

Note. Group 1 is the Reference group and Group 2 is the Focal group

Table 39

DIF Analysis of MHV Set B with DIF Items Omitted

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
B2	.07	.07	.08	4.09 (.13)	.01	No
B3	.40	.40	.40	.29 (.86)	.01	No
B4	.46	.47	.49	2.52 (.28)	.03	No
B5	.44	.49	.51	6.01 (.05)	.07	No
B6	.68	.68	.71	1.99 (.39)	.03	No
B7	.62	.62	.64	1.58 (.45)	.02	No
B8	.67	.78	.80	10.72 (.01)	.13	Yes
B9	.51	.55	.55	3.07 (.22)	.04	No
B10	.71	.81	.81	6.40 (.04)	.10	No
B12	.62	.65	.66	2.67 (.26)	.04	No
B13	.53	.58	.60	5.56 (.06)	.06	No
B14	.45	.46	.46	.22 (.89)	.00	No
B15	.33	.35	.36	1.85 (.40)	.04	No
B16	.46	.47	.48	1.35 (.51)	.02	No
B17	.17	.18	.28	6.61 (.04)	.11	No
B18	.11	.20	.22	4.15 (.13)	.11	No
B19	.28	.29	.41	9.45 (.01)	.12	No

Table 39 continued.

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
B20	.10	.10	.22	9.43 (.01)	.12	No
B21	.28	.28	.39	10.19 (.01)	.11	No
B22	.47	.47	.51	3.21 (.20)	.03	No
B24	.25	.27	.31	2.59 (.27)	.06	No
B25	.28	.34	.35	2.32 (.31)	.07	No
B26	.16	.34	.36	8.73 (.01)	.21	No
B27	.27	.27	.32	2.90 (.23)	.06	No
B28	.01	.07	.08	2.42 (.30)	.07	No
B29	.01	.01	.01	.17 (.92)	.01	No
B30	.03	.03	.05	.88 (.65)	.03	No
B31	.33	.33	.35	.65 (.72)	.02	No
B32	.37	.47	.47	5.39 (.07)	.11	No
B33	.24	.24	.25	1.07 (.59)	.01	No
B34	.13	.17	.18	2.97 (.23)	.05	No

Adverse Impact

The results of the adverse impact analysis are presented in presented in Table 40. The selection ratio for each family of military occupations is displayed in hierarchical order. That is, the occupational families that require the lower CFAT cut-off scores are presented first. In order to demonstrate that the CFAT is not adversely impacting against Aboriginal Peoples, the selection ratio against the comparison group must be at least .80. According to the four-fifths rule, the CFAT is adversely impacting against Aboriginal Peoples. The ratio of Aboriginal Peoples meeting the minimum CFAT requirement for employment in the CF compared to the general Anglophone NCM applicant population falls just short of four-fifths at .77.

Evidence of adverse impact against Aboriginal Peoples can also be seen across all military families. In fact, the effect of adverse impact increases dramatically with each family. The selection ratio for the occupations in the Administrative, Mechanical, Operator and Technical families is .39.

Table 40

Assessment of Adverse Impact Against Aboriginal Peoples

	Total CFAT Data n =53169		Aboriginal Peoples n = 72		Impact Ratio	Adverse Impact?
	Pass % (n)	Fail % (n)	Pass % (n)	Fail % (n)		
CFAT Cut off	91.2 (48478)	8.8 (4691)	70.8 (51)	29.2 (21)	.78	Yes
Military Family						
Steward	94.6 (50298)	5.4 (2871)	72.2 (52)	27.8 (20)	.76	Yes
Cook	86.4 (45938)	13.6 (7231)	47.2 (34)	52.8 (38)	.55	Yes
General	80.0 (42535)	20.0 (10634)	45.8 (33)	54.2 (39)	.57	Yes
Military						
Administration	63.5 (33762)	36.5 (19407)	25.0 (18)	75.0 (54)	.39	Yes
RMS Clerk	58.9 (31312)	41.1 (21857)	0 (0)	100 (72)	.00	Yes
Mechanical	60.4 (32114)	39.6 (21055)	23.6 (19)	76.4 (53)	.39	Yes
Operator	60.4 (32114)	39.6 (21055)	23.6 (19)	76.4 (53)	.39	Yes
Technical	60.4 (32114)	39.6 (21055)	23.6 (19)	76.4 (53)	.39	Yes

Discussion

The results of this study suggest that the CFAT is not a fair tool to use when selecting Aboriginal Peoples living in special access and remote communities. Although the findings of the DIF analysis indicate that only one item on the CFAT may be biased against Aboriginal Peoples, the results of the adverse impact analysis indicate that there is a problem. In other words, although the test appears to be unbiased, the results suggest that the CFAT may be an unfair test with respect to Aboriginal Peoples.

There were significant differences in mean scores between the Aboriginal Peoples and the Reference group on the WPT and the MHV. However, the Aboriginal Peoples and the recruits performed similarly on the SPM. Although a couple of items of the SPM displayed DIF, the results suggest that it would be a more appropriate measure to use with Aboriginal Peoples.

The differences in mean scores between Aboriginal Peoples and the Reference group on verbal measures of cognitive ability do not reflect differences in cognitive ability. In fact, the lack in significant differences on nonverbal measures indicates that Aboriginal Peoples have the same level of cognitive ability as the Reference group. The differences in mean scores on verbal measures may the result of differences in language ability and education.

Determining if the Canadian Forces Aptitude Test is Biased

The primary goal of this study was to determine if the CFAT was biased against Aboriginal Peoples. To achieve this goal, three approaches were used. First, the means for the CFAT and its three scales were examined to determine if there were significant

group differences. The second approach entailed the use of logistical regression to determine if any items of the CFAT displayed DIF. Finally, the four-fifths rule was used to determine if the CFAT adversely impacted against Aboriginal Peoples. In order for the CFAT to be considered fair and unbiased, three conditions had to be met.

The first condition stated that there should not be any significant group differences on test means between Aboriginal Peoples and the Reference group on the CFAT total score and its three scales (VS, SA and PS). This condition was not met. There were significant differences between Aboriginal Peoples and the Reference group on the CFAT total, VS scale and PS scale. This was true for both the Focal and Eligible groups. Although the group differences in test means were not as dramatic as those found in other studies, they do support previous findings that Aboriginal Peoples tended to score lower than non-Aboriginal Peoples on verbal tests of cognitive ability (McShane & Plas, 1984).

After equating the Aboriginal Peoples with the Reference group on the basis of education and the CFAT minimum requirements, there was no significant difference in group means on the CFAT SA scale. This result supports previous findings that established that Aboriginal Peoples tended to score slightly higher on nonverbal tests of cognitive ability than non-Aboriginal Peoples (McShane & Plas, 1984).

The second condition stated that none of the items of the CFAT should display DIF. This condition was not met. Two items of the CFAT VS scale displayed uniform DIF. The presence of uniform DIF indicates that the two items are not equivalent measures of the same variable for both groups (Zumbo, 1999). In other words, the probability of getting the item correct is different for both groups and these differences are fairly stable across score levels (Robie et al., 2001). However, only one of the two

items displayed potential bias against Aboriginal Peoples. The results suggest that item VS9 may be advantageous to Aboriginal Peoples.

The presence of DIF does not mean that the items are biased. As stated earlier, DIF is a necessary but not sufficient condition for item bias. Content experts should review these two items in an effort to determine why they are performing in this manner.

The third condition, which stated that the CFAT should not adversely impact against Aboriginal Peoples, was not met. The application of the four-fifths rule indicates that the CFAT adversely impacts against Aboriginal Peoples. The ratio of Aboriginal Peoples meeting the minimum CFAT requirement for employment in the CF compared to the general Anglophone NCM applicant population falls just short of four-fifths. In other words, although there is evidence of adverse impact, the effect is quite small. More disturbing is the ratio of Aboriginal Peoples to general Anglophone NCM applicants who meet the CFAT requirement for employment in the various military occupation families. As the results indicate, based on CFAT scores alone, very few of the Aboriginal Peoples tested in this study would have been selected for employment in the Administrative, Mechanical, Operator and Technical families. Unfortunately the majority of the Aboriginal Peoples would have been employed as stewards and cooks. This parallels the trend found in the civilian world were Aboriginal Peoples are being denied opportunities or are relegated to low paying jobs (Brescia & Fortune, 1989).

Other Considerations about the Canadian Forces Aptitude Test

The reliability of the CFAT for the Focal and Reference groups combined meets the requirement set by Nunnally and Bernstein (1994), which suggest that the CFAT is a

good test to use for selection. However, overall, the reliabilities found in this study were much lower than then those previously reported. This is true of both the Aboriginal Peoples groups and the Reference group. The CFAT VS and SA scales, in particular, bear further investigation.

There was a considerable difference in reliabilities between the Aboriginal Peoples groups and the Reference group on the VS scale. The poor reliability found with the Aboriginal Peoples indicates that the VS scale is not an appropriate measure to use with this group. With regards to Aboriginal Peoples, we cannot be confident that the score obtained on the VS scale is a true measure of their abilities.

The low reliability of the CFAT SA scale was surprising. However, the internal consistency coefficient was consistent for all groups. Nonetheless, given the low reliability coefficient, the items on the SA scale need to be re-examined.

Finding a Suitable Replacement for the Canadian Forces Aptitude Test

The second goal of this study was to determine if another well established verbal or nonverbal measure of cognitive ability could be used in lieu of the CFAT. To achieve this objective, three approaches were used. First, the means for the WPT, SPM and MHV were examined to determine if there were significant group differences. Logistical regression was also used to determine if any items of the WPT, SPM, and MHV displayed DIF. Finally, correlation analysis was performed to examine the relationships between each measure.

Wonderlic Personnel Test

The WPT is a reliable and valid measure of general cognitive ability that has been used extensively in industrial and organizational psychology (Dodrill, 1983). However, it is a verbal measure of cognitive ability and as such there are some concerns regarding its cultural fairness with Aboriginal Peoples. Nevertheless, its high level of validity and reliability make it an excellent tool to measure the construct validity of the SPM and MHV. In order for the WPT to be an unbiased and reliable measure of cognitive ability for Aboriginal Peoples, two conditions had to be met.

The first condition that needed to be met was that there be no significant group differences on test means between Aboriginal Peoples and the Reference group. This condition was not met. The results indicate that there was a significant difference in test means between Aboriginal Peoples and on the WPT. Aboriginal Peoples scored significantly lower on the WPT than the Reference group. This was true for both the Focal and Eligible groups.

The second condition required that none of the WPT items display DIF. This condition was met. Although there is controversy regarding the cultural fairness of the WPT, there was no evidence that any of the items were biased against Aboriginal Peoples.

Raven's Standard Progressive Matrices

The SPM is a nonverbal test designed to assess inductive or analytical reasoning (Bors & Stokes, 1998). It is considered one of the best available measures of general

intelligence and complex reasoning (Marshalek, Lohman, & Snow, 1983). The SPM meets all the criteria for what is generally considered to be a culture fair test (Albert, 1998a). As a nonverbal measure of cognitive ability, it is expected that Aboriginal Peoples should perform similarly, if not better, than the Reference group (McShane & Plas, 1984). However, in order for the SPM to be considered as an unbiased and reliable measure of cognitive ability for Aboriginal Peoples, three conditions had to be met.

The first condition that had to be met was that there should not be significant group differences on test means between Aboriginal Peoples and the Reference group on the SPM. After equating the Aboriginal Peoples with the Reference group on the basis of education and minimum CFAT score, this condition was met. The fact that there were significant differences between the Focal and Reference groups should not be a concern. As was demonstrated by the univariate and multivariate analyses, the differences in test performance were confounded by the group differences in education and the fact that the Reference group had been pre-selected on the basis of CFAT scores.

The second condition required that none of the SPM items display DIF. This condition was not met. Two items displayed non-uniform DIF. The DIF for both items are non-uniform because for those individuals who scored at or below the mean the Focal group is favoured whereas for those scoring above the mean the Reference group is favoured. In other words, the probabilities for getting the item correct are different for the two groups and the differences are not stable across score levels (Robie et al., 2001). Content experts should review these two items in an effort to determine if they are biased.

The final condition stated that the SPM should correlate positively and highly with both the WPT and CFAT. Overall the SPM had a weak to moderate positive

correlation with the CFAT (Reference: $r = .47$; Focal: $r = .61$; Eligible: $r = .68$). The SPM also correlated moderately and positively with the WPT (Reference: $r = .42$; Focal: $r = .46$; Eligible: $r = .61$). These results suggest that SPM is related to the CFAT and WPT and offers evidence of construct validity. However, the moderate correlation suggests that the SPM does not measure the same facets of cognitive ability as the CFAT and WPT.

With the exception of two items, the SPM is an appropriate measure to use with Aboriginal Peoples. Although the test may not measure the same facets of cognitive ability as the CFAT, the SPM is one of the best available measures of inductive or analytical reasoning (Marshalek, Lohman, & Snow, 1983).

Mill Hill Vocabulary Scale

To get a more complete picture of an individual cognitive ability, the SPM is often administered with the MHV. There is no evidence that the MHV is a culture fair or unbiased test. However, the intention of this study was to explore the fairness of the MHV with a sample of Aboriginal Peoples. In order for the MHV to be an unbiased and reliable measure of cognitive ability for Aboriginal Peoples, the following three conditions had to be met.

The first condition which required that there be no significant group differences on test means between Aboriginal Peoples and the Reference group on the MHV was not met. Aboriginal Peoples scored significantly lower than the Reference group on the MHV. This was true for both the Focal and Eligible groups. These results are consistent with previous research findings (McShane & Plas, 1984).

According to the second condition, none of the items of the MHV should display DIF. This condition was not met. Five items from the MHV displayed DIF. Four of the items displayed uniform DIF and one displayed non-uniform DIF. One of the four uniform DIF items and the sole uniform DIF may be biased towards to Aboriginal Peoples. Once again, content experts should review all of the DIF items to determine if they are in fact biased.

To display construct validity, the third condition stated that the MHV should correlate positively with the WPT and CFAT. As expected the MHV had moderate to strong correlation with the CFAT VS (Combined: $r = .73$; Focal: $r = .65$; Eligible: $r = .65$; Reference: $r = .65$) and the WPT (Reference: $r = .43$; Focal: $r = .76$; Eligible: $r = .74$). However the MHV had stronger reliabilities across all groups than the CFAT VS. Unlike the CFAT VS, the internal constancy reliability for the MHV was acceptable for all groups.

Due to the differences in group test means, the MHV does not appear to be a culture free measure of verbal ability. However, the reliabilities suggest that the MHV may be a more appropriate measure of verbal skills than the CFAT VS for both Aboriginal Peoples and the Reference group.

English as a Second Language

For many of the Aboriginal Peoples who participated in this study, English was a second language. Language related factors often confound attempts to accurately measure the cognitive ability of people from various cultures (Sattler, 1982). As Bradford (1960)

points out, participants with inadequate education and low reading levels may score low on verbal tests irrespective of their actual intelligence.

Consequently, separate univariate analyses were performed for each test to determine if the difference in test means between the Eligible and Reference groups were significant after controlling for language. It was expected that language should influence performance on the verbal measures of cognitive ability (CFAT VS and PS, WPT, MHV) but not on the nonverbal measures (CFAT SA, SPM).

As expected, language did not influence performance on the nonverbal measures of cognitive ability. These results support the findings of earlier studies (Sattler, 1982). The purpose of nonverbal tests is to measure the ability of individuals with inadequate education and low reading levels. Consequently language should not affect test scores.

Language did influence performance on the WPT. This was evident by the lack of significant difference in test means after controlling for language. These results support the hypothesis that testing in one's secondary language may confound the results of verbal measures (Krywanuik & Das, 1976; Zarske & Moore, 1982).

The presence of significant differences between the Eligible and Reference groups on the CFAT VS, CFAT PS, CFAT Total and MHV means after controlling for language indicate that there are one or more latent factors influencing performance that have not been controlled for. It is unclear exactly what these factors are. These differences may be affected by some underlying cultural differences that have not been identified. Content experts should carefully review these measures in an effort to determine what factors are affecting performance.

Limitations

There are several potential limitations in this study that may have had an impact on the findings. Firstly, there may be some criticism with regards to the methodology used to create the Eligible group of Aboriginal Peoples from the Focal group. Univariate analysis of the tests means for the Focal and Reference groups produced different results than the analysis for the Eligible and Reference groups. The former analysis revealed significant differences on all test means, however, there were no significant differences in CFAT SA and SPM means in the latter analysis. However, after treating the education and CFAT variables as covariates, the results of univariate analysis indicated that there was a significant difference on performance on the SPM. This finding justifies the creation of the Eligible group of Aboriginal Peoples.

Secondly, there may be some doubts about generalizing the findings of this study to other Aboriginal Peoples in Canada. The Aboriginal Peoples who participated in this study may not be representative of all Aboriginal Peoples in Canada. There are three recognized groups of Aboriginal Peoples in Canada, First Nations, Metis and Inuit. The participants in this study were all First Nations. Although the three groups are often referred to as Aboriginal Peoples, each group is culturally distinct.

When considering the appropriateness of generalizing these results, geographical differences between the various Aboriginal Peoples communities also need to be considered. For instance, the Aboriginal participants were drawn from communities that were designated as special access or remote by the Department of Indian Affairs and Northern Development (DIAND; INAC, 2002a) According to INAC only 20.9% of all registered Aboriginal Peoples live in special access and remote communities (2002a).

Aboriginal Peoples who live in special access and remote communities do not have the same level of access to supplies, financial institutions, educational, health and community services as Aboriginal Peoples living in urban communities (communities located within 50 km from nearest service centre; INAC, 2002a). Consequently, it may not be appropriate to generalize these results to Aboriginal Peoples living in urban communities.

In addition to geographical differences among Aboriginal Peoples, there are also cultural differences. The participants in this study shared a common First Nation language. However there are 53 different First Nation languages and numerous local dialects spoken in Canada (INAC, 2002a).

Thirdly, range restriction may have confounded the results. Although the Eligible group was created in an attempt to equate the Aboriginal Peoples with the Reference group, the results of this study may be underestimated because of range restriction. The Eligible group was created by dropping Aboriginal participants who did not meet the minimum educational requirement and who did not attain the minimum CFAT score. On the other hand, the Reference group was composed entirely of recruits who had successfully met all of the selection requirements, which included a semi-structured interview. The recruits were also subjected to a comprehensive realistic job preview that may have resulted in a more homogeneous group of participants due to self-selection. As a result of the range restriction, correlations between Aboriginal Peoples and Reference group may be smaller than the correlations that would be obtained if the Focal group was compared with the normal population.

Finally, there are some limitations to the findings of the analysis of adverse impact. The four-fifth rule ignores the concepts of chance and statistical significance

(Organization and Management Solutions & Myklebust, 2000). Consequently, due to the small sample size, there is a slight chance that the CFAT was found to be discriminating against Aboriginal Peoples, when in reality no discrimination exists.

Implications for Further Research

Many organizations, including the Canadian government are trying increase the representation of Aboriginal Peoples in their workforce. Traditionally, these organizations have been using verbal measures of cognitive ability to select employees. However, as this study has shown, there are problems with using verbal measures of cognitive ability to select Aboriginal Peoples. Verbal measures of cognitive ability consistently underestimate the ability of Aboriginal Peoples. Consequently, Aboriginal Peoples may be denied opportunities or may be relegated to low paying jobs (Brescia & Fortune, 1989). The results of this study suggest that it may be more appropriate to administer a nonverbal measure instead.

The results of this and other studies raises some interesting questions that deserve further investigation. First, why are verbal measures of cognitive ability underestimating the ability of Aboriginal Peoples? It is clear that the English as a second language factor explains some of the variances. However, the results of this study indicate that, something else is also affecting the results. After analysing the performance of Aboriginal Peoples on the WPT and CFAT PS, it is clear that they had difficulty with word problems involving mathematical equations, especially those involving fractions and/or decimal points. This is curious because their performance on the CFAT SA indicate that they have the potential to learn mathematics.

One possible explanation for the poor performance of Aboriginal Peoples on verbal cognitive ability measures may be related to the level of the education that Aboriginal Peoples receive. In special access, remote and rural (communities that are located between 50 and 350 km from the nearest service centre and having year-round road access) communities, the band councils usually manage the elementary and secondary schools in their jurisdiction (INAC, 2002b). The schools are funded by DIAND and must follow the provincial curriculum. However, band councils are given a lot of leeway to incorporate unique Aboriginal orientated material such as culture and language training (INAC, 2002b). According to INAC 65% of Aboriginal students attend band-operated schools (INAC, 2002b). The difference in performance on verbal cognitive ability test between Aboriginal and the Reference Group may be related to the differences in education. Unfortunately, it appears that little or no research has been conducted in this area.

Another possible explanation for the poor performance of Aboriginal Peoples may be related to the notion of stereotype threat. Stereotype threat is the risk of confirming a negative stereotype about one's group as a self-characteristic (Steele, 1997). This threat affects the performance of women and African American on standardized math and verbal tests (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995). For example, the commonly held assumption that women are less skilled in mathematics than men has been shown to affect the performance of women on standardized math tests. When female participants were primed beforehand of this negative stereotype, scores were significantly lower than if the women were led to believe the tests did not reflect these stereotypes (Spencer, Steele, & Quinn, 1999). It is possible that the performance of

Aboriginal Peoples on verbal measures may be affected by their knowledge of negative stereotypes held about them. Researchers should look at how stereotype threat affects the performance of Aboriginal Peoples on standardized math and verbal tests.

As discussed earlier, there is some concern about the appropriateness of generalizing these results to all Aboriginal Peoples. There may be a difference between First Nations, Metis and Inuit, and between Aboriginal Peoples from urban communities and those from special access and remote communities. Based on previous research it safe to assume that there will be no significant differences in performance on nonverbal measures between First Nations, Metis and Inuit, and Aboriginal Peoples living in urban and special access/remote communities. However, the same cannot be said for verbal measures of cognitive ability. Students from urban communities often go to elementary and secondary schools controlled by the provincial governments and do not receive Aboriginal culture and language training (INAC, 2002b). If there is a difference in test performance between Aboriginal Peoples living in special access/remote and urban communities, it may be linked to the differences in education. Clearly more research is needed before any conclusions can be made.

Another issue that deserves further investigation is whether or not you can have test bias when no items display DIF. According to DIF theory, if there is no DIF than there is no test bias (Zumbo, 1999). However, according to Gestalt theory, the whole is greater than the sum of its parts. Although only one CFAT item was discovered to be potentially biased towards Aboriginal Peoples, it is possible that there is an interaction among the different items and scales of the CFAT that causes the test as a whole to be biased.

This may be the first time that a DIF analysis was conducted on the WPT, SPM and MHV. Considering the extensive use of the WPT in industrial and organization settings, the absence of DIF is a significant finding. However, the samples size in this study were small and there is a risk that there was not enough power to detect DIF items. A DIF analysis of the WPT with a much larger study should be conducted.

The presence of DIF items in the SPM and MHV was troublesome. However, the presence of DIF is not sufficient to conclude that the items are biased. The items need to be examined by content experts in an effort to determine if there are indeed biased. A DIF analysis of the SPM and MHV should also be conducted with a larger sample.

Implications of Findings for the CF

Based on the results of this and other studies, Aboriginal Peoples who live in special access and remote communities should be selected for enrolment in the CF based on performance on CFAT SA and SPM. Both measures should be used because they provide valuable information on two different aspects of cognitive ability. The CF SA scale measures a candidate's ability to generate, retain, and transform a variety of complex three-dimensional figures, which has long been recognized as a factor contributing to success in mathematics, natural sciences, engineering, architecture, and other fields of study (Miller & Bertoline, 1991; Rhoades, 1981). The SPM measure the ability to reason and solve problems involving new information, without relying on a base of knowledge derived from previous experience or schooling (Carpenter, Just, & Shell, 1990).

A major disadvantage of using only nonverbal measures of cognitive ability is that the CF will not be able to assess an individual's ability to speak, read and comprehend English or French, which is essential because all training in the CF is conducted in either English or French. It is unlikely that individuals who do not possess an adequate ability in either language will be able to succeed in training. This leads to another dilemma. Is it practical for the CF to select Aboriginal Peoples who may have low ability to read and comprehend English or French? The CF already has the infrastructure in place to train current CF members to function in their second language (English or French). It should be feasible to send Aboriginal Peoples who have the potential to learn, as demonstrated by their performance on the CFAT SA and SPM, to one of the CF's various language schools. There they could be trained to function in either English or French.

Another disadvantage of using nonverbal measures is that the CF will not be able to assess math knowledge. Many of the technical and operator occupations in the CF require a solid math foundation. However, the CFAT SA does measure the potential to learn math. With some preparation, the CF could also use the language schools to review or teach the mathematical knowledge required to those Aboriginal Peoples who display the potential to learn math.

If the CF decides to use the SPM there are a couple of issues that need to be considered. At present there are no norms available for Aboriginal Peoples, nor are there norms for Canadian adults. Secondly, caution should also be used when interpreting the norms that are provided. Gudjonson (1995a) notes major problems related to the 1992 standardization of SPM norms. The most fundamental flaw is that participants were left

unsupervised to complete the test in their own home. This raises the risk that participants may not have completed the test on their own. Secondly, participants were given approximately one week to complete the test. In a normal test administration most people take between 40 minutes and one hour to complete the SPM (Lezak, 1983). This extended period of time gave the participants the opportunity to take breaks lasting hours or even days. These two flaws may have, in all likelihood, inflated the test scores. Consequently, the 1992 standardization sample appears to be a poor normative group and should be used cautiously when comparing with participants who were tested under normal supervised conditions (Gudjonson, 1995b). Consequently, if the CF does decide to use the SPM it would need to develop its own norms.

Recommendations

The CF should consider using the CFAT SA and SPM to select Aboriginal Peoples living in special access and remote communities. This will require the CF to establish norms and to determine cut-off scores. The CF will also need to determine if there are differences between First Nations, Metis and Inuit, and Aboriginal Peoples living in special access/remote communities and those living in urban communities. It may be possible that existing selection procedures may be appropriate for Aboriginal Peoples living in urban and rural communities.

The CF should also examine the cost benefits of using existing facilities to train Aboriginals Peoples to function efficiently in either English or French and to teach mathematics. The CF should also consult with the Assembly of First Nations, Métis

National Council and Inuit Tapiriit Kanatami to determine if there are more cost effective solutions to upgrading the language and math ability of Aboriginal Peoples.

The reliabilities of the CFAT VS and SA scales were quite low for measures that are being used for selection. The CF should undertake a much larger study to determine if these reliabilities are accurate. If so the CF should investigate the causes of the low reliabilities and try to rectify the problem.

Due to current legislation, the onus is on the CF to demonstrate that its selection procedures and tests are not biased against any members of a designated minority group. If there any doubts, it is the CF's duty to take action to remedy the situation and establish selection practices that are equitable to all Canadians.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29(1)*, 67-91.

Albert, J. A. (1998a). *Short measures of "g": A review of the British Army Recruit Battery (BARB), the Wonderlic Personnel Test and Raven's Progressive Matrices* (Technical Note 98-9). Ottawa, ON: Personnel Research Team.

Albert, J. A. (1998b). *Cognitive measures: Comparison of the Canadian Forces Aptitude Test (CFAT), Raven's Progressive Matrices, and the Wonderlic Personnel Test* (Technical Note 98-5). Ottawa, ON: Personnel Research Team.

Allen, G. L., Kirasic, K. C., Dobson, S. H., Long, R. G., & Beck, S. (1996). Predicting environmental learning from spatial ability: An indirect route. *Intelligence, 22,* 327-355.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standard for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Black, M. S., (1999). *The efficacy of personality and interest measures as a supplement to cognitive measures in the prediction of military training performance*. Unpublished master's thesis, Saint Mary's University, Halifax, Nova Scotia, Canada.

Boardman, A. E. (1979) Another analysis of the EEOC 'four-fifths' rule. *Management Science, 25,* 770-776.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58(3),* 382-398.

Bradford, E. J. (1960). Shipley-Institute of Living Scale for measuring intellectual impairment. In O. K. Buros (Ed.), *The Third Mental Measurements Yearbook.* Highland Park, NJ: Gryphon Press.

Brescia, W. & Fortune, J. C. (1989). Standardized testing of American Indian students. *College Student Journal, 23,* 98-104.

Brown, D. L., & Wheatley, G. H. (1989). Relationship between spatial knowledge. In C. A. Maher, G. A. Goldin, & R. B. Davis (Eds.), *Proceedings of the 11th Annual Meeting, North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 143-148). New Brunswick, NJ: Rutgers University.

Burke, H. R. (1985). Raven's Progressive Matrices (1938): More on norms, reliability, and validity. *Journal of Clinical Psychology, 41(2),* 231-234.

Bussiere, M. T. (1997). *The detection of bias due to gender, visible minority group membership, and Aboriginal status on the Canadian Forces Aptitude Test (CFAT).* Ottawa, ON: Personnel Research Team.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Campbell, S. K. (2001). *Investigating the use of alternative predictors of training performance in the Canadian Forces operator occupations.* Unpublished master's thesis, Saint Mary's University, Halifax, Nova Scotia, Canada.

Campbell, C. A., & Cotton, A. J. (1994). *Guidelines for the development and use of selection tests in the Canadian Forces: The ethical, legal and practical issues associated with bias in tests* (Technical Note 94-1). Willowdale, ON: Canadian Forces Personnel Applied Research Unit.

Carpenter, P.A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 98*, 404-431.

Catano, V. M., Cronshaw, S. F., Wiesner, W. H. Hackett, R. D., & Methot, L. L. (2001). *Recruitment and Selection.* Scarborough, ON: Nelson Thomson Learning.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive tests. *Journal of Applied Psychology, 82*, 311-320.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronshaw, S. F. (1991). *A conceptual and operational model of predictive test bias for the Canadian Forces Selection Tests* (Research Report 91-2). Willowdale, ON: Canadian Forces Personnel Applied Research Unit.

Dahmen, W., Hartje, W., Büssing, A., & Sturm, W. (1982). Disorders of calculations in aphasic patients – spatial and verbal components. *Neuropsychologia, 20(2),* 145-153.

Darou, W. G. (1992). Native Canadians and intelligence testing. *Canadian Journal of Counseling, 26(2)*, 96-99.

Deary, I., J. (1995) Auditory inspection time and intelligence: What is the direction of causation? *Developmental Psychology, 31*, 237-250.

Director Recruiting Education and Training (1998). *Instructions for the Canadian Forces Aptitude Test (CFAT)*. Ottawa, ON: Department of National Defence.

Dodrill, C. B. (1983). Long-term reliability of the Wonderlic Personnel Test. *Journal of Consulting and Clinical Psychology,51*, 316-317.

Dodrill, C. B. (1981). An economical method for the evaluation of general intelligence in adults. *Journal of Consulting and Clinical Psychology, 49*, 688-673.

Dodrill, C. B., & Warner, M. H. (1988). Further studies of the Wonderlic Personnel Test as a brief measure of intelligence. *Journal of Consulting and Clinical Psychology, 56 (3)*, 145-147.

Environics Research Group Limited. (1997). *A survey of visible minorities, aboriginals and women to assess their level of interest in joining the Canadian Forces.* Ottawa, ON: Department of National Defence.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290-38315.

Ewins, J. E. M. (1997). *Canadian Forces Diversity Survey: A population analysis* (Technical Note 1-97). Ottawa, ON: Canadian Forces Personnel Applied Research Unit.

Fairweather, G. (1986, June). *Employment testing: How to make it valid, reliable*. Paper presented at the 47th Annual Conference of the Canadian Psychological Association, Toronto, Ontario.

Frisch, M. B., & Jessop, N. S. (1989). Improving WAIS-R estimates with the Shipley-Fartford and Wonderlic Personnel Tests: Need to control for reading ability. *Psychological Reports, 65*, 923-928.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and substantive reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research, 45*, 353-376.

Gottfredson, L. D. (1986). Societal consequences of the g factor in employment. *Journal of Vocational Behavior, 29*, 379-411.

Gottfredson, L. D. (2002). Where and why g matters: Not a mystery. *Human Performance, 15 (1/2)*, 25-46.

Greenberg, I. (1979). An analysis of the EEOC "four-fifths rule." *Management Science 25*, 726-769.

Guelph Centre for Occupational Research. (1997). *Analysis of general cognitive scores and enrolment outcomes relative to gender and ethnicity of military applicants*. Ottawa, ON: Personnel Research Team.

Gudjonson, G. H. (1995a). The Standard Progressive Matrices: Methodological problems associated with the administration of the 1992 adult standardisation sample. *Personality and Individual Differences, 18(3)*, 441-442.

Gudjonson, G. H. (1995b). Raven's norms on the SPM revisited: A reply to Raven. *Personality and Individual Differences, 18(3)*, 447.

Hambleton, R. A., & Rodgers, H. J. (1994). *Developing an item bias review form.* College Park, Maryland: ERIC Clearinghouse on Assessment and Evaluation.

Hambleton, R. A., & Rodgers, H. J. (1995). Item bias review. *Practical Assessment, Research & Evaluation, 4(6).*

Hambleton, R. A., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests.* (CSE Technical Report 483). Los Angeles, CA: Center for the Study of Evaluation.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Warner & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340-362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law, 2*, 447-472.

Ibel, R. H., & Cotton, C. A. (1994). *Validation of experimental scales of non-commissioned member applicant selection and classification* (Technical Note 17-94). Willowdale, ON: Canadian Forces Personnel Applied Research Unit.

Indian and Northern Affairs Canada. (2002a). *Basic Departmental Data - 2001.* Ottawa, ON: Department of Indian Affairs and Northern Development.

Indian and Northern Affairs Canada. (2002b). *Backgrounder First Nation Elementary/Secondary Education.* Ottawa, ON: Department of Indian Affairs and Northern Development.

Jensen, A. R. (1986). g: Artifact or reality? *Journal of Vocational Behavior, 29*, 301-331.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating power and Type I error rates using an effect size with Logistical Regression procedures for DIF. *Applied Measurement in Education, 14(4)*, 329-349.

Jodoin, M. G., & Huff, K. L. (2001, April). *Examining Type I error and power rates when ability distributions are unequal with the logistical regression procedure for DIF detection.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Kleinfield, J. S. (1973). Intellectual strengths in culturally different groups: An Eskimo illustration. *Review of Educational Research, 43(3)*, 341-359.

Kleinfield, J. & Nelson, P. (1991). Adapting instructions to Native Americans' learning styles. *Journal of Cross-Cultural Psychology, 22*, 273-282.

Krywanuik, L. W., & Das, J. P. (1976). Cognitive strategies in native children: Analysis and intervention. *Alberta journal of Educational Research, 22*, 271-280.

Lezak, M. D. (1983). *Neuropsychological assessment.* Oxford: Oxford University Press.

Li, H. Z. (1999) Information communication in conversation: A cross-cultural comparison. *International Journal of Intercultural Relations, 23,* 387-409.

Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647-677.

MacLennan, R. N. (1997). *Validity generalization across military occupational families* Technical Note 00-97). Ottawa, ON: Personnel Research Team.

Marshalek, B. Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radix and hierarchical models of intelligence. *Intelligence, 7,* 107-127.

Matthews, D. J., (1988). Raven's matrices in the identification of giftedness. *Roeper Review, 10,* 159-162.

McArthur, R. S. (1973). Some ability patterns: Central Eskimos and Nsenga Africans. *International Journal of Psychology, 8(4),* 239-247.

McKelvie, S. J. (1989). The Wonderlic Personnel Test: Reliability and validity in an academic setting. *Psychological Reports, 65,* 161-162.

McShane, D. A., & Berry, J. W. (1988). Native North Americans: Indian and Inuit abilities. In S. H. Irvine and J. W. Berry (Eds.), *Human abilities in cultural context.* Cambridge: Cambridge University Press.

McShane, D. A., & Plas, J. M. (1984). The cognitive functioning of American Indian children: Moving from WISC to the WISC-R. *School Psychology Review, 13,* 61-73.

Miller, C. L., & Bertoline, G. R. (1991). Spatial visualization research and theories: Their importance in the development of an engineering and technical design graphics curriculum model. *Engineering Design graphics Journal, 55(3),* 377-385.

Morris, S. B., & Lobsenz, R. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*, 89-111.

Murphy, K. R. (1984). The Wonderlic Personnel Test. In D. J. Keyser & R. C. Sweetlands (Eds), *Test Critiques. Vol 1*. Kansas City, MO: Test Corp. of America.

Murray, J. B. (1999). *Final Report Aboriginal Recruitment and Retention in the Canadian Forces*. Ottawa, ON: John B. Murray Consultants Ltd.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszeal and Simultaneous Item Bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Nkaya, H. N., Huteau, M., & Bonnet, J. P. (1994). Retest effect on cognitive performance on the Raven-38 Matrices in France and in the Congo. *Perceptual and Motor Skills, 78*, 503-510.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd Edition). New York: McGraw-Hill, Inc.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. *Journal of Applied Psychology, 79*, 845-851.

Organization and Management Solutions & Myklebust, K. (2000). *Analysis of assessment scores and enrolment outcomes relative to employment equity status of military applicants* (Sponsor Research Report 00-12). Ottawa, ON: Director Human Resources Research and Evaluation.

Osborne, B. (1985). Research into Native North Americans' cognition: 1973-1982. *Journal of American Indian Education, 24 (3)*, 9-25.

Outtz, J. L. (2002). The role of cognitive ability tests in employment selection. *Human Performance, 15 (1/2)*, 161-171.

Parmar, R. S. (1989). Cross-cultural transfer of nonverbal intelligence test: An (in)validation study. *British Journal of Educational Psychology, 59*, 379-388.

Raven, J., Raven J. C., & Court, J. H. (2000). *Raven's Manual: Section 3: Standard Progressive Matrices*. Oxford: Oxford Psychologists Press.

Raven, J., Raven J. C., & Court, J. H. (1998a). *Raven's Manual: Section 1: General Overview*. Oxford: Oxford Psychologists Press.

Raven, J., Raven J. C., & Court, J. H. (1998b). *Raven's Manual: Section 5: Mill Hill Vocabulary Scale*. Oxford: Oxford Psychologists Press.

Ree, M. J., & Caretta, T. R. (1997) What makes an aptitude test valid. In Dillon, R. F. (Ed.), *Handbook on testing* (pp. 65-81). London: Greenwood Press.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1*, 86-89.

Reynolds, C. R., & Brown, R. T. (1997). Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R. T. Brown, (Eds.), *Perspective on Bias in Mental Testing*. New York: Plenum Press.

Rhoades, H. M. (1981). Training spatial ability. In E. Klinger (Ed.), *Imagery, Volume 2, Concepts, Results, and Applications* (pp. 247-256). New York: Prenum Press.

Robie, C., Mueller, L. M., & Campion, J. E. (2001). Effects of a motivational inducement on the psychometric properties of a cognitive ability test. *Journal of Business and Psychology, 16 (2)*, 177-189.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszeal procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Saccuzzo, D . P., & Johnson, N. E. (1995). Traditional psychometrics tests and proportionate representation: An intervention and program evaluation study. *Psychological Assessment, 7(2)*, 183-194.

Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities.* Boston: Allyn and Bacon, Inc.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 134(2)*, 262-274.

Schmidt, F. L., & Hunter, J. E. (2000). Select on intelligence. In E. Locke (Ed.), *The* Blackwell Handbook of Principles of Organizational Behavior. Malden, MA: Blackwell Publishing.

Schwartz, S. (1999). *Cognitive test bias and options for trust bias mediation* (Technical Note). Ottawa, ON: Director Human Resources Research and Evaluation.

Sheppard, C., Fiorentino, D., & Collins, L. (1968). Performance errors on Ravens Progressive Matrices by sociopathic and schizotypic personality types. *Psychological Reports, 23(3)*, 1043-1046.

Sidles, C., MacAvoy, J., Bernston, C., & Kuhn, A. (1987). Analysis of Navajo adolescents' performances on the Raven Progressive Matrices. *Journal of American Indian Education, 27(1)*, 1-8.

Smith, D. P. (1995). *The employment equity profile of the Canadian Forces recruitable population* (Technical Note 4-95). Willowdale, ON: Canadian Forces Personnel Applied Research Unit.

Society for Industrial and Organizational Psychology (1987). *Principles for the Validation and Use of Personnel Selection Procedures*, 3rd ed. College Park, MD: Aurthor.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35(1)*, 4-28.

Stark, S., Chernyshenko, S., Chuah, D., Lee W., & Wadlington, P. (2001). [on-line] *Detection of differential item/test functioning (DIF/DTF) using IRT.* http://work.psych.uiuc.edu/irt/dif_main.asp

Steele, C. M. (1997). A threat in the Air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52(2),* 613-629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797-811

Steel R.D, Torrie J.H. Dickey T.A. (1997). *Principles and Practice of Statistics: A Biomedical Approach.* McGraw Hill: New York.

Stough, C., Nettlebeck, T., & Cooper, C. (1993). Raven's Advance Progressive Matrices and increases in intelligence. *Personality and Individual Differences, 15 (1),* 103-104.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Tabachnick, B. & Fidell, L. (2001). *Using Multivariate Statistics* (4th ed.). New York: HarperCollins College Publishers.

Trochim, W. (2000). *The Research Methods Knowledge Base*, 2nd Edition. Atomic Dog Publishing, Cincinnati, OH.

van de Vijver, F. & Tanzer, N. K. (1998). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47 (4)*, 263-279.

Vincent, K. R. (1991). Black/white differences: Does age make the difference? *Journal of Clinical Psychology, 47*, 266-270.

Vining, A. R., McPhillips, D. C., & Boardman, A. E. (1986). Use of statistical evidence in employment discrimination litigation. *The Canadian Bar Review 64*, 660-702.

Watts, K., Baddeley, A. D., & Williams, M. (1982). Automated tailored testing using Raven's Matrices and Mill Hill Vocabulary tests: A comparison with manual administration. *International Journal of Man Machine Studies, 17*, 331-334.

Wheatley, G. H. (1991). Enhancing mathematics learning through imagery. *Arithmetic Teacher, 39(1)*, 34-36.

Whitmore, M. J., & Schumacker, R. E. (1999). A comparison of logistical regression and analysis of variance differential item functioning detection methods. *Educational and psychological Measurement, 59*, 910-927.

Wonderlic, E. F. (1997). *Wonderlic Personnel Test User's Manual*. Libertyville, IL: Wonderlic Personnel Test, Inc.

Woycheshin, D. E. (1999). *Validation of the Canadian Forces Aptitude Test against QL3 course performance* (Technical Note 99-11). Ottawa, ON: Director Human Resources Research and Evaluation.

Zarske, J. A., & Moore, C. (1982). Recategorized WISC-R scores for non-handicapped, learning disabled, educational disadvantage and regular classroom Navajo children. *School Psychology Review, 11*, 319-323.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistical Regression Modelling as a Unitary Framework for Binary and Likert Type (Ordinal) Item Scores* (Technical Note 99-6). Ottawa, ON: Director Human Resources Research and Evaluation.

Zumbo, B. D., & Hubley, A. M. (1998a). *A Psychometric Study of the Canadian Forces Aptitude Test (CFAT)* (Technical Note 98-10). Ottawa, ON: Personnel Research Team.

Zumbo, B. D., & Hubley, A. M. (1998b). *Differential item functioning (DIF) analysis of a synthetic CFAT* (Technical Note 98-4). Ottawa, ON: Personnel Research Team.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia, Prince George, BC, Canada.

Appendix A:

How to Conduct Binary DIF Analysis using SPSS Logistic Regression[1]

In logistical regression, a model comparison is performed in which a binary item response (0 = incorrect, 1 = correct) is predicted by the scale score under investigation, group membership (0 = reference, 1 = focal), and the interaction between scale score and group membership (Robie et al., 2001). This procedure will provide a test of DIF on the relationship between item response and scale score, examining the effect of group membership for uniform DIF, and the interaction of scale score and group membership to assess non-uniform DIF (Zumbo, 1999). These variables are entered hierarchically in the following order:

Step 1: Scale score (TOT),

Step 2: Group membership (GROUP), and

Step 3: Interaction between scale score and group membership (TOT * GROUP).

The equation used for logistic regression is:

$$Y = b_0 + b_1TOT + b_2GROUP + b_3TOT*GROUP,$$

Where Y is a natural log of the odds ratio. Thus the equation can be rewritten as:

$$\ln\left[\frac{p_i}{(1-p_i)}\right] = b_0 + b_1TOT + b_2GROUP + b_3(TOT*GROUP),$$

where p is the proportion of individuals that endorse the item in the direction of the latent variable. With this equation, one can then perform the Chi-Square test (2-degree of freedom) for both uniform and non-uniform DIF.

---

[1] Logistic Regression can be used to conduct DIF analysis of ordinal data. For a complete discussion on how to conduct ordinal logistic regression please see Zumbo (1999).

The Chi-squared (χ^2) test for logistical regression is computed by subtracting the χ^2-value obtained in Step 1 from the χ^2-value in Step 3. It should be noted that the χ^2-value for logistical regression has 2 degrees of freedom. The 2 degrees of freedom is the difference between the χ^2 at Step 3 (3 degrees of freedom) and the χ^2 at Step 1 (1 degree of freedom).

For an item to be classified as displaying DIF, two criteria must be met. The first criterion is that the DIF must be statistically significant. To be considered significant, the p-value for the two-degree of freedom χ^2 in logistical regression must be $\leq .01$ (Robie et al., 2001: Zumbo, 1999).

The second criterion that must be met is that the magnitude of the significance (effect size) must be substantial and meaningful. To meet this criterion, the Zumbo-Thomas (1997) effect size must be > 0.130. The Zumbo-Thomas effect size can be obtained by subtracting the R^2 obtained in Step 1 from the R^2 in Step 3. In addition, the R^2 in Step 2 can be compared to the R^2 in Step 3 to determine how much adding the non-uniform DIF component contributes to the model.

Example of Binary DIF Analysis

In order to conduct Binary DIF analysis using Logistic Regression your data must be coded using the following format:

Item Score: 0 = incorrect, 1 = correct

Group Membership: 0 = reference, 1 = focal

Run logistical regression using the SPSS Syntax presented in Appendix B. SPSS will provide an out-put similar to the one presented in Appendix C. The results of the logistic regression are presented in Table A1.

Table A1

Example of DIF Analysis using Logistic Regression

Item	R^2 at Each Step in Sequential Hierarchical Regression			χ^2 2-df/ (p)	Zumbo-Thomas R^2	DIF?
	R^2_1	R^2_2	R^2_3			
1	.220	.225	.226	1.506 (.471)	.006	No
2	.208	.697	.700	69.32 (.000)	.402	Yes

From the SPSS output presented in Appendix C and Table A, you can see that the χ^2-values for Item 1 and Item 2 are 1.506, p = .4710 and 69.32, p = .000, respectively. With a p-value of .000, only Item 2 meets the statistically significant requirement. The Zumbo-Thomas effect size for Item 2 is .402 (.700-.208), which is substantial and meaningful. Based on the two criteria, Item 2 is identified as displaying DIF.

Using the R^2 calculated in the logistic regression one can also determine if DIF was uniform or non-uniform by subtracting the R^2 obtained in Step 2 from the R^2 in Step 3. In this case the difference between R^2 in Step 3 and R^2 in Step 2 for Item 2 is .003 (.700 - .697). The difference in R^2 is quite small, suggesting that the DIF was predominately uniform.

Appendix B:

SPSS Syntax for DIF with Logistic Regression[1]

```
* SPSS SYNTAX written by:  .
* Bruno D. Zumbo, PhD  .
* Professor of Psychology and Mathematics,  .
* University of Northern British Columbia  .
* e-mail: zumbob@unbc.ca  .

* Instructions .
* Change the filename, currently 'binary.sav' to your file name .
* Change 'item', 'total', and 'grp', to the corresponding variables in your file.
* Run this entire syntax command file.

compute item= item1.
compute total= scale.
compute grp= group.

*   Aggregation.
*   Working with the Centered data.

* Hierarchical regressions approach with the following order of steps:.
*    1. total.
*    2. total + group.
*    3. total + group + interac.
* This also, of course, allows one to compute the relative Pratt Indices.


* Saves the standardized versions of group and total with the.
* eventual goal of centering before computing the cross-product term.

DESCRIPTIVES
  VARIABLES=group total  /SAVE
  /FORMAT=LABELS NOINDEX
  /STATISTICS=MEAN STDDEV MIN MAX
  /SORT=MEAN (A).

* Allows for both uniform and non-uniform DIF.
* Provides the 2df Chi-square test for DIF.
```

---

[1] From *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistical Regression Modelling as a Unitary Framework for Binary and Likert Type (Ordinal) Item Scores*, by B. D. Zumbo, 1999, Ottawa, ON: Director Human Resources Research and Evaluation.

```
LOGISTIC REGRESSION item
  /METHOD=ENTER ztotal /method=enter zgroup ztotal*zgroup
  /SAVE PRED(pre1).
execute.

*  The following command is required to deal with the repeaters in.
*  the data.  The WLS regression will be conducted on the aggregate.
*  file entitled "AGGR.SAV".

AGGREGATE
  /OUTFILE='aggr.sav'
  /BREAK=zgroup ztotal
  /item = SUM(item) /pre1 = MEAN(pre1)
  /Ni=N.

GET
  FILE='aggr.sav'.
EXECUTE .

compute interact=zgroup*ztotal.
execute.

COMPUTE v1 = Ni*pre1 *(1 - pre1) .
EXECUTE .

COMPUTE z1 = LN(pre1/(1-pre1))+ (item-Ni*pre1)/Ni/pre1/(1-pre1) .
EXECUTE .

FORMATS v1, z1 (F8.4).
execute.

* Overall logistic regression.
* Both Uniform and Non-uniform DIF.
REGRESSION
  /MISSING LISTWISE
  /REGWGT=v1
  /descriptives=corr
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHA
  /NOORIGIN
  /DEPENDENT z1
  /METHOD=ENTER ztotal / method=enter zgroup / method= enter interact .
execute.
```

Appendix C:

SPSS Output for Logistic Regression

Item 1

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
GROUP	400	1.00	2.00	1.5000	.50063
TOTAL	400	.00	20.00	10.3050	3.98581
Valid N (listwise)	400				

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	400	100.0
	Missing Cases	0	.0
	Total	400	100.0
Unselected Cases		0	.0
Total		400	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
.00	0
1.00	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted	
			ITEM	
	Observed		.00	1.00
Step 0	ITEM	.00	356	0
		1.00	44	0
	Overall Percentage			

Classification Table[a,b]

Observed			Predicted
			Percentage Correct
Step 0	ITEM	.00	100.0
		1.00	.0
	Overall Percentage		89.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df
Step 0	Constant	-2.091	.160	171.176	1

Variables in the Equation

		Sig.	Exp(B)
Step 0	Constant	.000	.124

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	ZTOTAL	79.833	1	.000
	Overall Statistics		79.833	1	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	94.193	1	.000
	Block	94.193	1	.000
	Model	94.193	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	183.020	.210	.420

Classification Table[a]

			Predicted	
			ITEM	
	Observed		.00	1.00
Step 1	ITEM	.00	347	9
		1.00	28	16
	Overall Percentage			

Classification Table[a]

Observed			Predicted
			Percentage Correct
Step 1	ITEM	.00	97.5
		1.00	36.4
	Overall Percentage		90.8

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df
Step 1[a]	ZTOTAL	2.057	.278	54.694	1
	Constant	-3.329	.330	102.070	1

Variables in the Equation

		Sig.	Exp(B)
Step 1[a]	ZTOTAL	.000	7.820
	Constant	.000	.036

a. Variable(s) entered on step 1: ZTOTAL.

Block 2: Method = Enter

Staistical Test of Significance for DIF

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1.506	2	.471
	Block	***1.506***	**2**	***.471***
	Model	95.698	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	181.514	.213	.426

Classification Table[a]

Observed			Predicted	
			ITEM	
			.00	1.00
Step 1	ITEM	.00	350	6
		1.00	28	16
	Overall Percentage			

Classification Table[a]

Observed			Predicted: Percentage Correct
Step 1	ITEM	.00	98.3
		1.00	36.4
	Overall Percentage		91.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df
Step 1[a]	ZTOTAL	2.163	.302	51.371	1
	ZGROUP	.054	.336	.026	1
	ZTOTAL by ZGROUP	.183	.302	.368	1
	Constant	-3.346	.336	99.450	1

Variables in the Equation

		Sig.	Exp(B)
Step 1[a]	ZTOTAL	.000	8.701
	ZGROUP	.872	1.056
	ZTOTAL by ZGROUP	.544	1.201
	Constant	.000	.035

a. Variable(s) entered on step 1: ZGROUP, ZTOTAL * ZGROUP .

Regression

Correlations[a]

		Z1	Zscore(TOTAL)
Pearson Correlation	Z1	1.000	.469
	Zscore(TOTAL)	.469	1.000
	Zscore(GROUP)	-.042	-.231
	INTERACT	-.088	-.344

Correlations[a]

		Zscore(GROUP)	INTERACT
Pearson Correlation	Z1	-.042	-.088
	Zscore(TOTAL)	-.231	-.344
	Zscore(GROUP)	1.000	.804
	INTERACT	.804	1.000

a. Weighted Least Squares Regression - Weighted by V1

Variables Entered/Removed[b,c]

Model	Variables Entered	Variables Removed	Method
1	Zscore(TOTAL)[a]	.	Enter
2	Zscore(GROUP)[a]	.	Enter
3	INTERACT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Z1

c. Weighted Least Squares Regression - Weighted by V1

Effect Size

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.469[a]	***.220***	.198	2.2686963
2	.474[b]	***.225***	.181	2.2939664
3	.476[c]	***.226***	.158	2.3251280

Model Summary

Model	Change Statistics				
	R Square Change	F Change	df1	df2	Sig. F Change
1	.220	10.162	1	36	.003
2	.005	.211	1	35	.649
3	.002	.068	1	34	.796

a. Predictors: (Constant), Zscore(TOTAL)

b. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP), INTERACT

ANOVA[d,e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	52.302	1	52.302	10.162	.003[a]
	Residual	185.291	36	5.147		
	Total	237.593	37			
2	Regression	53.413	2	26.707	5.075	.012[b]
	Residual	184.180	35	5.262		
	Total	237.593	37			
3	Regression	53.781	3	17.927	3.316	.031[c]
	Residual	183.811	34	5.406		
	Total	237.593	37			

a. Predictors: (Constant), Zscore(TOTAL)

b. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP), INTERACT

d. Dependent Variable: Z1

e. Weighted Least Squares Regression - Weighted by V1

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	t
1	(Constant)	-3.313	.758		-4.371
	Zscore(TOTAL)	2.043	.641	.469	3.188
2	(Constant)	-3.330	.767		-4.340
	Zscore(TOTAL)	2.113	.666	.485	3.174
	Zscore(GROUP)	.216	.469	.070	.460
3	(Constant)	-3.347	.780		-4.289
	Zscore(TOTAL)	2.163	.702	.497	3.083
	Zscore(GROUP)	5.403E-02	.781	.018	.069
	INTERACT	.183	.703	.069	.261

Coefficients[a,b]

Model		Sig.	Collinearity Statistics	
			Tolerance	VIF
1	(Constant)	.000		
	Zscore(TOTAL)	.003	1.000	1.000
2	(Constant)	.000		
	Zscore(TOTAL)	.003	.947	1.056
	Zscore(GROUP)	.649	.947	1.056
3	(Constant)	.000		
	Zscore(TOTAL)	.004	.876	1.142
	Zscore(GROUP)	.945	.351	2.847
	INTERACT	.796	.327	3.058

a. Dependent Variable: Z1

b. Weighted Least Squares Regression - Weighted by V1

Excluded Variables[c,d]

Model		Beta In	t	Sig.	Partial Correlation
1	Zscore(GROUP)	.070[a]	.460	.649	.077
	INTERACT	.083[a]	.526	.602	.089
2	INTERACT	.069[b]	.261	.796	.045

Excluded Variables[c,d]

Model		Collinearity Statistics		
		Tolerance	VIF	Minimum Tolerance
1	Zscore(GROUP)	.947	1.056	.947
	INTERACT	.881	1.134	.881
2	INTERACT	.327	3.058	.327

a. Predictors in the Model: (Constant), Zscore(TOTAL)

b. Predictors in the Model: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Dependent Variable: Z1

d. Weighted Least Squares Regression - Weighted by V1

Collinearity Diagnostics[a,b]

Model	Dimension	Eigenvalue	Condition Index
1	1	1.861	1.000
	2	.139	3.653
2	1	2.054	1.000
	2	.810	1.593
	3	.136	3.890
3	1	2.599	1.000
	2	1.098	1.539
	3	.179	3.807
	4	.124	4.587

Collinearity Diagnostics[a,b]

Model	Dimension	Variance Proportions			
		(Constant)	Zscore(TOTAL)	Zscore(GROUP)	INTERACT
1	1	.07	.07		
	2	.93	.93		
2	1	.05	.05	.07	
	2	.03	.02	.90	
	3	.91	.93	.03	
3	1	.02	.02	.03	.03
	2	.07	.04	.09	.05
	3	.17	.08	.73	.65
	4	.74	.86	.15	.27

a. Dependent Variable: Z1

b. Weighted Least Squares Regression - Weighted by V1

Item 2

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
GROUP	400	1.00	2.00	1.5000	.50063
TOTAL	400	.00	20.00	10.3050	3.98581
Valid N (listwise)	400				

Logistic Regression

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	400	100.0
	Missing Cases	0	.0
	Total	400	100.0
Unselected Cases		0	.0
Total		400	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
.00	0
1.00	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted	
			ITEM	
	Observed		.00	1.00
Step 0	ITEM	.00	212	0
		1.00	188	0
	Overall Percentage			

Classification Table[a,b]

Observed			Predicted
			Percentage Correct
Step 0	ITEM	.00	100.0
		1.00	.0
	Overall Percentage		53.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df
Step 0	Constant	-.120	.100	1.438	1

Variables in the Equation

		Sig.	Exp(B)
Step 0	Constant	.230	.887

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	ZTOTAL	47.429	1	.000
	Overall Statistics		47.429	1	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	50.363	1	.000
	Block	50.363	1	.000
	Model	50.363	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	502.714	.118	.158

Classification Table[a]

			Predicted	
			ITEM	
	Observed		.00	1.00
Step 1	ITEM	.00	159	53
		1.00	88	100
	Overall Percentage			

Classification Table[a]

Observed			Predicted
			Percentage Correct
Step 1	ITEM	.00	75.0
		1.00	53.2
	Overall Percentage		64.8

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df
Step 1[a]	ZTOTAL	.781	.120	42.211	1
	Constant	-.137	.107	1.654	1

Variables in the Equation

		Sig.	Exp(B)
Step 1[a]	ZTOTAL	.000	2.183
	Constant	.198	.872

a. Variable(s) entered on step 1: ZTOTAL.

Block 2: Method = Enter

Staistical Test of Significance for DIF

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	69.324	2	.000
	Block	***69.324***	2	***.000***
	Model	119.687	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	433.389	.259	.345

Classification Table[a]

Observed			Predicted	
			ITEM	
			.00	1.00
Step 1	ITEM	.00	161	51
		1.00	65	123
	Overall Percentage			

Classification Table[a]

Observed			Predicted: Percentage Correct
Step 1	ITEM	.00	75.9
		1.00	65.4
	Overall Percentage		71.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df
Step 1[a]	ZTOTAL	1.251	.154	66.174	1
	ZGROUP	1.049	.140	55.903	1
	ZTOTAL by ZGROUP	-.075	.154	.234	1
	Constant	-.199	.140	2.020	1

Variables in the Equation

		Sig.	Exp(B)
Step 1[a]	ZTOTAL	.000	3.493
	ZGROUP	.000	2.855
	ZTOTAL by ZGROUP	.628	.928
	Constant	.155	.819

a. Variable(s) entered on step 1: ZGROUP, ZTOTAL * ZGROUP .

Regression

Correlations[a]

		Z1	Zscore(TOTAL)
Pearson Correlation	Z1	1.000	.456
	Zscore(TOTAL)	.456	1.000
	Zscore(GROUP)	.343	-.540
	INTERACT	.055	.046

Correlations[a]

		Zscore(GROUP)	INTERACT
Pearson Correlation	Z1	.343	.055
	Zscore(TOTAL)	-.540	.046
	Zscore(GROUP)	1.000	.070
	INTERACT	.070	1.000

a. Weighted Least Squares Regression - Weighted by V1

Variables Entered/Removed[b,c]

Model	Variables Entered	Variables Removed	Method
1	Zscore(TOTAL)[a]	.	Enter
2	Zscore(GROUP)[a]	.	Enter
3	INTERACT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Z1

c. Weighted Least Squares Regression - Weighted by V1

Effect Size

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.456[a]	***.208***	.186	1.5834674
2	.835[b]	***.697***	.680	.9923113
3	.836[c]	***.700***	.673	1.0033710

Model Summary

Model	Change Statistics				
	R Square Change	F Change	df1	df2	Sig. F Change
1	.208	9.435	1	36	.004
2	.490	56.670	1	35	.000
3	.002	.233	1	34	.633

a. Predictors: (Constant), Zscore(TOTAL)

b. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP), INTERACT

ANOVA[d,e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.656	1	23.656	9.435	.004[a]
	Residual	90.265	36	2.507		
	Total	113.921	37			
2	Regression	79.457	2	39.729	40.347	.000[b]
	Residual	34.464	35	.985		
	Total	113.921	37			
3	Regression	79.692	3	26.564	26.386	.000[c]
	Residual	34.230	34	1.007		
	Total	113.921	37			

a. Predictors: (Constant), Zscore(TOTAL)

b. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Predictors: (Constant), Zscore(TOTAL), Zscore(GROUP), INTERACT

d. Dependent Variable: Z1

e. Weighted Least Squares Regression - Weighted by V1

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	t
1	(Constant)	-9.545E-02	.185		-.515
	Zscore(TOTAL)	.626	.204	.456	3.072
2	(Constant)	-.162	.117		-1.391
	Zscore(TOTAL)	1.243	.152	.905	8.190
	Zscore(GROUP)	1.041	.138	.831	7.528
3	(Constant)	-.199	.141		-1.416
	Zscore(TOTAL)	1.251	.154	.910	8.107
	Zscore(GROUP)	1.049	.141	.838	7.452
	INTERACT	-7.451E-02	.154	-.046	-.482

Coefficients[a,b]

Model		Sig.	Collinearity Statistics	
			Tolerance	VIF
1	(Constant)	.610		
	Zscore(TOTAL)	.004	1.000	1.000
2	(Constant)	.173		
	Zscore(TOTAL)	.000	.709	1.411
	Zscore(GROUP)	.000	.709	1.411
3	(Constant)	.166		
	Zscore(TOTAL)	.000	.701	1.426
	Zscore(GROUP)	.000	.699	1.430
	INTERACT	.633	.985	1.015

a. Dependent Variable: Z1

b. Weighted Least Squares Regression - Weighted by V1

Excluded Variables[c,d]

Model		Beta In	t	Sig.	Partial Correlation
1	Zscore(GROUP)	.831[a]	7.528	.000	.786
	INTERACT	.034[a]	.226	.822	.038
2	INTERACT	-.046[b]	-.482	.633	-.082

Excluded Variables[c,d]

Model		Collinearity Statistics		
		Tolerance	VIF	Minimum Tolerance
1	Zscore(GROUP)	.709	1.411	.709
	INTERACT	.998	1.002	.998
2	INTERACT	.985	1.015	.699

a. Predictors in the Model: (Constant), Zscore(TOTAL)

b. Predictors in the Model: (Constant), Zscore(TOTAL), Zscore(GROUP)

c. Dependent Variable: Z1

d. Weighted Least Squares Regression - Weighted by V1

Collinearity Diagnostics[a,b]

Model	Dimension	Eigenvalue	Condition Index
1	1	1.025	1.000
	2	.975	1.025
2	1	1.539	1.000
	2	1.003	1.239
	3	.458	1.832
3	1	1.539	1.000
	2	1.415	1.043
	3	.616	1.581
	4	.430	1.892

Collinearity Diagnostics[a,b]

Model	Dimension	Variance Proportions			
		(Constant)	Zscore(TOTAL)	Zscore(GROUP)	INTERACT
1	1	.49	.49		
	2	.51	.51		
2	1	.00	.23	.23	
	2	.99	.00	.00	
	3	.01	.77	.77	
3	1	.00	.23	.23	.00
	2	.29	.00	.00	.29
	3	.56	.10	.09	.54
	4	.15	.67	.68	.17

a. Dependent Variable: Z1

b. Weighted Least Squares Regression - Weighted by V1

Vanderpool Lt(N) MA@ADM(HR-Mil) DHREE@Ottawa-Hull

From:
Sent:
To:
Subject:

This email authorizes Michael Vanderpool to reprint Figures 1, 2 and 3 (from *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistical Regression Modelling as a Unitary Framework for Binary and Likert Type (Ordinal) Item Scores*, by B. D. Zumbo, 1999, Ottawa, ON: Director Human Resources Research and Evaluation) for use in his Master's Thesis.

R.A. Boswell LCol
DHRRE 2

Intellegere