NOTE TO USERS

COMPARABILITY OF TEST SCORES FOR NON-ABORIGINAL AND ABORIGINAL STUDENTS

by

BONITA MARIE DAVIDSON

B.A., Laurentian University, 1991
B.Ed., Laurentian University, 1992
D.Edu., The University of British Columbia, 1997
M.A., The University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Department of Educational and Counselling Psychology, and Special Education;
Measurement, Evaluation, and Research Methodology Program)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

July 2004

ABSTRACT

The purpose of this study was to examine the comparability of the BC Ministry of Education's Grades 4 and 7 Reading and Numeracy Foundation Skills Assessment (FSA) scores for aboriginal and non-aboriginal students. It was found that the compositions of the constructs being measured had many similarities across the aboriginal and non-aboriginal populations and were congruent for the reading assessments but not for the numeracy assessments. The reliability estimates of the scores for each population were high and very similar. The Grade 7 Numeracy assessment provided more measurement accuracy for the aboriginal group than the non-aboriginal group, while the Grade 4 Numeracy assessment and the Grades 4 and 7 Reading assessments provided less measurement accuracy for the aboriginal group than the non-aboriginal group. For all assessments, items were ordered similarly in terms of their difficulty level and their degree of discrimination, and were ordered moderately similar in their inherent possibility of being answered correctly based on chance. For all assessments there was a low level of differential item functioning.

Overall, the results indicated that for this study, there was a high degree of comparability across the aboriginal and non-aboriginal populations for the Reading FSA scores because all four analyses for both grades showed them to be highly comparable. There was a moderately high degree of comparability across the two populations for the Grade 4 Numeracy FSA scores because three out of the four analyses showed them to be highly comparable. There was a moderate degree of comparability across the two populations for the Grade 7 Numeracy FSA scores because two out of the four analyses showed them to be highly comparable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Thanks to my husband Paul for his love, support, and friendship. Thanks to my new son Sam for being so wonderfully perfect & adorable. Thanks to my mom and dad for believing in me and loving me & Sam so much. Thanks to my niece Alex for her love and support and for being a great influence in my life.

Thanks to Tanya Steele for being an incredibly fun, brilliant, thoughtful, helpful, and motivating friend…the best sister EVER!

Thanks to Rina Bonano for her kindness & friendship…for sharing so much of herself & for thinking that I am special.

Thanks to Kadriye Ercikan for being very smart, patient, and supportive.

Thanks to David Robitaille for being the most perfect roll-model.

Thanks to Charles Ungerleider for being insightful and inspiring.

Finally, thanks to Ralph Hakstian for his knowledge, ideas, and strength.

CHAPTER ONE: INTRODUCTION

A few years ago, I was a teacher in a beautiful coastal community in British Columbia (BC). I was a recent graduate of a teacher education program and eager to motivate the students to learn the skills and concepts included in the BC Ministry of Education curriculum. I was responsible for a split-grade class that was composed mainly of Grade 5 students. About half of the students in my class were aboriginal children who lived on the nearby reserve, and about half were non-aboriginal children who lived on the nearby islands. As I worked and developed a relationship with my students, I began to believe that the cognitive abilities of the two groups were very similar, but their learning styles and types of motivating factors were quite different. I did my best to tailor classroom activities to meet the different learning styles and different types of motivating factors. By Christmas I felt that all the students were finding success as learners.

In the spring, my school principal presented me with a table of my Grade 5 students' results from a large-scale standardized test in which they had taken or written during the previous school year. The results were presented in graphical form, and it was clear that there were two distinct groups of scores: a high-performing group (above the 60^{th} percentile), and a low-performing group (below the 40^{th} percentile). As I examined the results and read the names that belonged to each of the two groups, I was surprised to see that the high-performing group consisted of all my Grade 5 non-aboriginal students and that the low-performing group consisted of all my Grade 5 aboriginal students. I was surprised at these results because my perception of the students' academic abilities did not match the test results; the dramatic difference between the two groups did not, in my view, accurately reflect the students' performance in my class. For years afterwards, I

wondered if that large-scale assessment was a fair and well-designed measurement tool for both groups.

Upon a recent examination of the BC Ministry of Education's Foundation Skills Assessment (FSA) 2000-2001 results, I saw that, at the provincial level, aboriginal students consistently scored lower than non-aboriginal students in almost all content areas measured for Grades 4 and 7. I found myself wondering the same thing about this assessment as I had for the one described above: Was FSA a fair and well-designed measurement tool for both aboriginal and non-aboriginal students? Were the items included in the FSA developed in such a way that neither group was at an unfair advantage or disadvantage by being given an assessment that had content, context, or language that was unfamiliar to them. I decided to perform a construct comparability study of the FSA scores to see if the test was measuring the same thing for both the aboriginal and non-aboriginal groups in an effort to see if biases existed. Gould (1995) says that test scores may be biased either culturally or statistically: "... culturally biased when one group (typically a minority population) performs consistently lower than some reference population ..." and "... statistically biased if two individuals (e.g., one African American, one White) who get the same test score nevertheless perform differently on some criterion external to the test, such as school grades" (p. 2). For the present study, the focus will be on the degree of comparability of test scores for the aboriginal and non-aboriginal students in British Columbia, Canada.

Overview of the Study

Large-scale assessments of students' academic achievement have been widely adopted by states and provinces in North America in an effort to measure learning outcomes and the effectiveness of schooling. Establishing the validity and comparability of scores from large-scale assessments across gender, cultural, racial, or ethnic subgroups is critical to interpreting assessment results accurately. As Messick (1995) has stated, "The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question" (p. 741). The present study examines the validity and comparability of aboriginal and non-aboriginal students' scores from a large-scale assessment, specifically, the 2000-2001 BC Ministry of Education's FSA English (as opposed to French), Numeracy, and Reading scores for Grades 4 and 7.

Problem

The problem addressed in the proposed study stems from the consistently poor academic achievement results of Canadian aboriginal students. A high school dropout rate of nearly 75% for students of aboriginal heritage makes this population a special concern, especially when research findings show that poor academic achievement is a strong factor in a student's decision to quit school (Alexander, Entwisle, & Kabbani, 2001). At a national level, the educational attainment of the Canadian aboriginal population is well below that of the non-aboriginal population; based on 1996 Canadian census data, 42% of the aboriginal population aged 20-29 did not graduate from high school, as opposed to 17% of the comparable non-aboriginal population (Council of Ministers of Education, 2002).

In an effort to understand the causes and sources of the general disparity between the aboriginal and non-aboriginal populations when it comes to academic testing results, one must first examine the test score results to see if the measurement tools being used are equally appropriate for both non-aboriginal and aboriginal populations. Recent FSA results indicate that the aboriginal student population consistently performed at lower levels than non-aboriginal students in the content area of mathematics for Grades 4 and 7. These interpretations are based on the findings for the 2000-2001 academic year (British Columbia Ministry of Education, 2002b).

This paper will determine whether or not FSAs administered to aboriginal and non-aboriginal students actually measure the same constructs for both groups of students by examining evidence of construct validity and comparability.

Purpose of Study & Research Questions

The purpose of the present study is to explore the statistical nature of large-scale assessment scores in an effort to establish the comparability (or lack thereof) of the interpretation of the scores for aboriginal and non-aboriginal students. Two research questions that guide this study are as follows: (1) *Are scores from the Foundation Skills Assessment comparable across aboriginal and non-aboriginal populations*? and (2) *Should score interpretations be the same for both populations*?

The Comparable Nature of Scores

In a recent publication by the United States National Research Council, it was stated that, " . . . a school whose students have higher test scores is not necessarily better than one whose students have lower test scores…the quality of inputs, such as the entry characteristics of students or educational resources available, must be considered"

(Committee on Foundations of Assessments, 2001, p. 36). Further, in an effort to encourage an accurate and respectful understanding of the complexities surrounding the educational performance and attainment of the aboriginal population, the Canadian Council of Ministers of Education (CMEC) made a statement about the context in which this performance and attainment should be viewed: (a) The first language of many aboriginal children is neither English nor French; hence attending a school taught in a language different from their first language offers undue and often un-addressed challenges to the learners; (b) Cultural differences typically exist between aboriginal children and their teachers, and aboriginal children and their non-aboriginal classmates; (c) Negative stereotyping of aboriginal children and their families currently exists; (d) There are relatively few aboriginal people who have found success in postsecondary education who can act as role models for educational attainment in the aboriginal community; and (e) The geographically remote nature of many aboriginal communities makes it difficult to attract and retain well-qualified teachers for the respective schools (Council of Ministers of Education, 2002).

An individual's performance on an assessment can be influenced by many cognitive and non-cognitive factors other than his or her ability. Scores may vary "for reasons unrelated to achievement, such as the specific content being assessed, the particular format of the assessment items, the timing and conditions for administering the assessments..." (Committee on Foundations of Assessments, 2001, p. 37). Further, level of ability with such skills as reading and writing will certainly have an impact on the individual's performance on a typical mathematical assessment if the items are presented in a word problem format. Also, familiarity with the context of the language used in

word problems, as well as allotted time to complete, will have an effect on the individual's performance on a test. Thus, to ensure the validity of test-score interpretations, one must ask, "To what degree—if at all—on the basis of evidence and rationales, should the test scores be interpreted and used in the manner proposed?" (Messick, 1989a, p. 5).

Conceptually, if the members of the aboriginal population interpret the test items in a different manner than the members of the non-aboriginal population, then a unique interpretation for the aboriginal students is necessary in order to be accurate. Item-interpretation differences between the aboriginal and non-aboriginal populations would imply that direct construct comparability does not exist between the two.

Whether one is exploring the comparable nature of measured constructs across two populations with one test, or the valid nature of test score interpretations for a single population, one should judge the validity of the test scores in terms of whether a test accomplishes the mission it was developed to achieve (Messick, 1989b). According to Messick, this judgment requires an evaluation of the intended and unintended social consequences of test interpretation and use. FSA scores are intended to measure the foundation skills of reading, writing, and numeracy. According to the BC Ministry of Education, the purposes of the FSA are as follows:

> The main purpose of this assessment is to help the province, school districts, schools, and school planning councils evaluate how well foundation skills are being addressed and make plans to improve student achievement. A secondary purpose is to provide teachers, students, and parents or guardians with external information about student performance. The information provided by

> FSA can facilitate discussion at the provincial, district, and school levels. (British Columbia Ministry of Education, 2002a)

In terms of *intended* score use, the BC Ministry of Education stated: "As with all assessment data, it is important to place FSA results in context, carefully considering the characteristics of the assessment instrument and various factors that might influence the results" (British Columbia Ministry of Education, 2002a). The Ministry highlighted such influencing factors as participation rate on the assessment, local policy, and instructional strategies (British Columbia Ministry of Education). They also suggested certain approaches to interpreting the FSA results, such as, "in comparison to local expectations, in relation to past performance, and against external references" (British Columbia Ministry of Education). For the sake of clarification, I would like to highlight that the Ministry did not refer to the aboriginal identity of students as a factor that may influence the appropriate FSA score interpretations, or as a factor by which score interpretations should be referenced.

In terms of *unintended* score use, The Fraser Institute published a *Report Card on British Columbia's Elementary Schools: 2003 Edition* in June 2003 that used FSA scores as its sole source of information on which to rate BC elementary schools' overall academic performance. Once rated, the 812 schools were ranked in descending order from best to worst. The Fraser Institute's reports have not been sanctioned by the Ministry, but they have received a great amount of media attention in British Columbia. I would consider this influence to be in the category of *unintended* score interpretations, and they should be identified as part of the construct validation study.

Methods of Construct Comparability

For the proposed study, the aboriginal population is defined as all aboriginal students who were in Grades 4 or 7 during the 2000-2001 school year, and who attended publicly and independently funded BC schools, including independent schools. The comparability of test scores for the aboriginal and non-aboriginal populations will be evaluated by an examination of the degree of congruence of the resulting factor structures, the degree of equivalence of the reliability estimates of the scores, the relative efficiency of the scores in terms of item-information functions, and the existence of items found to have significant Differential Item Functioning (DIF).

Factor analysis involves the study of order and structure in multivariate data; its objective is to summarize the empirical relationships among a given set of data (Gorsuch, 1983). The aim of this factor analysis will be to summarize the interrelationships among the measured variables (items) accurately and succinctly. This allows us to investigate whether the test data have similar structures for both groups; and, in this case, it also allows us to examine whether items are related to the overall test score in the same way. The aim of the reliability estimates will be to indicate the degree to which individuals' scores would remain relatively consistent over repeated administration of the same test or alternate test forms (Crocker & Algina, 1986). Reliability estimates are indicators of accuracy and are at the core of examining the degree to which test scores are accurate. The aim of calculating the relative efficiency of the scores in terms of item information is to display any disparity in the contribution that each item makes in estimating ability along the ability continuum. Finally, the aim of the DIF analysis will be to determine if

groups that are expected to perform similarly differ in their mean performances on specific items (Hambleton, Swaminathan, & Rogers, 1991, p. 109).

Importance of the Study

The principal contribution of the present study will be the determination of score validity and comparability across both the aboriginal and the non-aboriginal groups. If it can be shown that test scores are not valid or comparable across the groups, then a better-suited assessment tool, or at least a better-suited set of score interpretations, could be designed to replace what is currently in use so that each group's scores could be deemed valid.

Preview of Chapter Two

Chapter two reviews and summarizes research literature regarding the educational measurement notion of validity as well as the academic testing performance of the aboriginal population in Canada.

CHAPTER TWO: LITERATURE REVIEW

This chapter reviews and summarizes research literature regarding the educational measurement notion of *validity* as well as the *academic testing performance* of the aboriginal population in Canada. Establishing the appropriateness of educational testing for the aboriginal population in Canada has remained in a strained state since the inception of the Indian Act's policy mandating the formal education of all aboriginal children in Canada (Kirkness, 1999). Early studies of the academic success level of the aboriginal population in Canada show a strong similarity to recent studies of the same topic. Studies from both the early 1960s and the 2000s offer differing hypotheses about factors related to the relatively poor academic performance of the aboriginal population, but one vein that runs through over 40 years of educational research about aboriginal students is that as a whole, they consistently perform at lower levels than their non-aboriginal counterparts.

An identification of the related factors to relatively poor academic performance by the aboriginal population does not provide the information that is needed to ensure, or make changes in the direction of fair and equitable testing. Nor does it provide the information needed to establish if a test is actually measuring the same thing for the aboriginal students as it is for the non-aboriginal students. One very powerful way to accomplish the goals of detecting fair and equitable testing across the aboriginal and non-aboriginal groups is to perform a construct validation study with scores from tests that are administered to both aboriginal and non-aboriginal students.

This study will examine the psychometric properties of test scores in an effort to gain insight into such possible issues as differing factor structures, differing degrees of

internal consistency, differing item information and efficiency functions, and test-item bias. These types of information will give way to the development of a set of test-development or test-modification recommendations that may be actualized in an effort to ensure that test score interpretations are valid for both the aboriginal and non-aboriginal populations.

Scope and Organization of the Review

Validity

For this review, the notion of validity is examined from two perspectives; one is the current professionally-upheld practical rules for the development of a test that ensures the validity of the planned inferences of its scores, and the other is the currently upheld theory underpinning a unitary notion of validity. The practical perspective on validity is meant to inform the development of the instrument from all angles; the theoretical perspective is also meant to inform the development of the instrument, but much of what the validity theory extols is focused on the test scores, implying a post-development perspective of the test. Simply, the practical perspective outlines how valid test score inferences should be obtained, while the theoretical perspective outlines how valid test score inferences should be obtained, how to check if this is so, *and* what to do about it if there is a problem. Both the practical and the theoretical perspective have recent seminal works that will be highlighted in this review. These recent seminal papers on validity have augmented and corrected flaws in past theory, and thus the conclusions from the past papers are suspect, and will, consequently, not be included in the present review.

Aboriginal Educational Performance

For this review, the notion of the educational performance of the aboriginal population is examined from an evolutionary perspective, meaning that early research about the educational performance of aboriginal students is paired with current research in an effort to identify historically-stable or newly-emerging themes. One historically-stable theme that was stated previously is that 40 years of educational research about aboriginal students shows that they consistently perform at lower levels than their non-aboriginal counterparts. This theme will be the main focus of the aboriginal education section of the present review.

Literature Review

Professionally-Upheld Practical Rules for The Development of a Test

The Joint Committee on Testing Practices was established in 1985 by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). According to its bylaws, this committee provides "a means by which professional organizations and test publishers can work together to improve the use of tests in assessment and appraisal" (American Psychological Association, 2002). In 1999, the committee produced their second edition of *The Standards for Educational and Psychological Testing* (from here forward referred to as "the *Standards* document"). The *Standards* document was written as a practical guide that offers a theoretical base for its claims, and is considered to be a major seminal paper on the topic of test/assessment development. In the following paragraphs, I will summarize and critique the *Standards*

document with regard to issues of *validity* and *fairness in testing* for identifiable sub-populations.

Generally speaking, the authors of the *Standards* document (AERA, APA, & NCME, 1999) referred to validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9) and deemed *validity* as the most fundamental consideration in developing and evaluating tests (p. 9). These same authors identified validity as a "unitary concept" (p. 11) that identifies the "degree to which all the accumulated evidence supports the intended interpretation of the test scores for the proposed purpose" (p. 11). Based on the *Standards* document, for either planning or post-hoc score-interpretation validation, the process of identifying the degree of support mentioned above begins with the identification of the construct or concepts the test is intended to measure and is followed with an explicit description of what the test scores will be used for. Based on these two factors, appropriate evidence of validation can be identified and sought. Of noted importance in the search of evidence is the degree of both construct under representation (insufficient attention or focus on the construct or concepts the test is intended to measure) and construct-irrelevant variance (test-response influence(s) that are not relevant to the construct or concepts the test is intended to measure).

In terms of examining evidence that is based on the consequences of testing, the *Standards* document (AERA et al., 1999) clarified that when different populations reveal scores that are of different distributions, this is only evidence of invalidity if the test actually measured constructs or concepts unrelated to what was proposed to be measured,

or if group differences "were due to the test's sensitivity to some examinee characteristic not intended to be part of the test construct" (p. 16).

To assist in accomplishing all that is professed in the *Standards* document (AERA et al., 1999) to be professionally responsible with regard to creating valid score inferences, the document provides a list of 24 Standards that are intended to define the criteria that should be upheld, when applicable, in an effort to ensure that the interpretation of test scores is valid; and 12 Standards that are intended to define the criteria that should be upheld, when applicable, in an effort to ensure fairness-of-testing to every examinee, or sub-population of examinees. Of the 36 Standards mentioned above that are related to the present study; Standards 7.1, 7.2, and 7.3, highlight the critical link between the notions of validity with the criteria of fairness in testing and test use in a manner that informed and justified most of the methodology used in the present study. See below for the three highlighted Standards:

Standard 7.1

> When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant sub-group. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in subsequent test revisions. (p. 80)

Standard 7.2

> When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance on some part of the test, the test should be used if at all only for those subgroups for which evidence indicates that valid inferences can be drawn from test scores. (p. 81)

Standard 7.3

> When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups. (p. 81)

Theory Underpinning a Unitary Notion of Validity

Messick (1995) critiqued traditional notions of validity, claiming that a compartmentalized vision of validity that sees *content validity, criterion validity,* and *construct validity* as separate aspects of a test, was incomplete to a fault because it failed to "take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use" (p. 741). Messick presented a non-compartmentalized vision of validity that relates all aspects of validity while encompassing the value implications of score meaning and the social consequences of score use. His comprehensive view of validity "integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility" (p. 742).

Appearing to be of fundamental importance to Messick (1995) with regard to the true meaning of validity in psychological/educational testing, a claim that he had made in the previous decade was repeated: "Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (p. 741). Messick insisted that the concept of validity must encompass the interpretation of test scores, not the test scores themselves, and that the fairness of a specific test-score use could only be evaluated by examining the particular interpretation of test scores at hand. From there, Messick re-introduced a new notion of *construct validity* that "is based on an integration of any evidence that bears on the interpretation or meaning of the test scores—including content- and criterion-related evidence—which are thus subsumed as part of construct validity" (p. 742).

For the practical application of Messick's (1995) new vision of *construct validity*, he presented a list of six potential sources of evidence, or aspects, of validity for researchers to use or investigate in an effort to make their overall evaluative judgment about the validity of the interpretation of the test scores at hand. These six sources are: (1) content relevance and representativeness; (2) substantive theories, process models, and process engagement; (3) scoring models as a reflective task and domain structure; (4) generalizability and the boundaries of score meaning; (5) convergent and discriminant correlations with external variables; and (6) consequences as validity evidence. Messick defended the appropriateness and completeness of his selection of the six aspects of validity by stating that the "six aspects are highlighted because most score-based interpretations and action inferences, as well as the elaborated rationales or arguments

that attempt to legitimize them either invoke these properties or assume them, explicitly or tacitly" (p. 747).

In his conclusion, Messick (1995) put strong emphasis on the importance of examining the social consequences, "both potential and actual" (p. 748) of the interpretation of test scores. In this specific context, Messick claimed "it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity, such as construct underrepresentation or construct-irrelevant variance" (p. 748). In his closing words, Messick stated that "Thus, validity and values are one imperative, not two, and test validation implicates both the science and the ethics of assessment, which is why validity has force as a social value (p. 749).

Messick's notion of validity makes clear that in order for interpretations to be meaningful and justified, all components of validity need to be studied, and that this is part of ethics of assessment. This unitary notion of validity implies, to some degree, that the content aspects of the test need to be examined as well the statistical aspects of the test.

Early History of Aboriginal Education with Regard to Academic Success

Before European contact in North America, aboriginal people had a form of education " . . . in which the community was the classroom, its members were the teachers, and each adult was responsible to ensure that each child learned how to live a good life" (National Indian Brotherhood, 1973, as cited in Kirkness, 1999). Once contact was made between the Europeans and the Aboriginals, and cultures began to interact with one another, new skills became important for each group. Skills relating to hunting,

fishing, trapping, boating, navigating, and medicine, to name a few, were taught to the Europeans by the aboriginals. Skills relating to commerce and the English language were taught to the aboriginals by the Europeans. By the 1600s, Europeans began to establish day schools in Canada and the United States designed to *civilize* the aboriginals. By the 1800s, the day schools were being replaced by residential schools. At its peak, there were approximately 80 residential schools in Canada in the 1930s. In the 1950s, day schools were brought back to replace the residential schools, and by 1970s, most residential schools in Canada were closed. Near this time, a national policy on integration was brought forth, and aboriginal children began to attend public schools (Kirkness). This integration brought children together in terms of proximity, but it did not ensure an integration of cultures and educational beliefs. Problems related to aboriginal student success in these integrated schools were such that by 1972, it was reported that 96% of aboriginal children did not finish high school (Council of Ministers of Education, 2002).

The integration of aboriginal children into public schools was followed by reports of extremely poor levels of school success. The 1960s saw a research surge to explore the problems of education provided to the aboriginal populations in both Canada and the United States. The resulting research was accompanied by many recommendations for change and reform. The degree to which these recommendations were accepted and acted upon will not be addressed in the present study; rather, the focus is on the similarities between the 1960's research findings and more recent turn-of-the-century research findings.

In 1961, Clarence Wesley of Arizona, an Apache Tribe Chairman, spoke of the "uneasiness and deep concern" he had about the education being provided to aboriginal

children in United States both on and off reservations (Wesley, 1961). His paper was not a formal research paper, but rather an opinion piece from the chairman of a Tribe. This opinion paper echoes problems about aboriginal education that are still being heard today in both the United States and Canada. Wesley highlighted problems such as dropout rates, delinquency, and the weak skills that their students were being told were sufficient to graduate from Grade 12. He claimed that once the Indian students graduated and went to college or university, they were left "unable to compete with the non-Indian students" (Wesley). Wesley refused to attribute the lack of success of aboriginal students to the schools alone, but he did state that the schools must take partial responsibility because it is the fundamental function of the school "to prepare Indians to become responsible participants in the American way of life".

Wesley (1961) identified the causes of the lack of success to things such as the native students' weak English language skills and the dichotomy of SES between the native and non-native families. Wesley claimed that the above two influencing factors made the adoption of non-native language and culture very challenging for native children, but still saw this as part of the solution. He stated that the main advantage of having native children attend the same schools as non-native English-speaking children, "is the fact that here these youngsters are forced to use the English language on the playground because that is the only way they can make themselves understood by their non-Indian playmates." Wesley claimed that when the above scenario is not fully achieved, the school curriculum, which is prepared for students who are expected to have a certain level of proficiency with the English language, there is an extraordinary challenge for the native students to succeed and this scenario is hard to achieve when the

non-native and native groups come from such diverse cultural and SES backgrounds. In the end, Wesley's recommendation for the solution of the problems of native education was to focus on the use of the English language from a very early age while, simultaneously, firmly holding on to the Native culture.

Wesley's (1961) apparent perception of the problems involved with Native education, although written 40 years ago, seem to hold true to some degree today. His ideals were commendable and respectful of both native and non-native cultures; he simply affirmed that aspects of the cultures differed in ways that could have led to culturally-based obstacles for native children in their pursuit of success in the non-native school system.

David Lloyd (1961), also of Arizona, wrote a paper that attempted to examine whether native students performed better on standardized tests if they spend more time in a public school system whose population was made up of over 97% non-native students. He held a hypothesis that Indian students in the public school system were "doing as well, both intellectually and academically, as the non-Indian student." Lloyd did not substantiate his hypothesis with previously published research findings, so it was no surprise to me when, at the end of his paper, he concluded that his hypothesis was not supported by his data. Lloyd's statistical methodology was very weak, thus making his claim of an unsubstantiated hypothesis questionable. He examined both standardized intellectual ability scores and standardized academic achievement scores of a group of students from various grades. He made claims of a certain difference between the Indian and non-Indian students' scores based on his visual inspection of mean scores and graphic stanine scores without taking into account that less than 3% of the student scores

being examined were from the Indian students. To his paper's detriment, he did not perform any inferential statistical tests to see if there was a statistically significant difference between the scores of the non-Indian students and the scores of the Indian students. Lloyd made an additional attempt to compare the Grade 6, 8, and 10 Indian students who had attended the public school system for almost all of their public school years (a one year exception was allowed) to Indian students who have only attended public school for a couple of years (2 years or less for the Grade 6 and 8 students, 3 years or less for the Grade 8 students). Again, Lloyd performed some visual inspection of the scores and made a claim as follows:

> . . . there seems to be evidence that those Indians who have spent their entire educational life in the Mesa Public Schools tend to have a higher mean intelligence quotient for language, non-language and total mental as measured by the California Test of Mental Maturity than those who have been in the system a relatively short time.

Although Lloyd's (1961) conclusions that intellectual and academic achievement scores differ between the Indian and non-Indian students are founded on weak and faulty methodology, his closing remarks are worth noting. He concluded that the tests for intellectual ability do seem to favour students educated in the public school system, and that although some Indian students may have spent the majority of their educational years in public school, they still live in a "segregated situation where the socio-economic standard is much lower and where many of the enriching experiences are lacking." Lloyd wisely suggested that the standardized tests themselves may be biased instruments that favour the non-Indian student; a bias that could have resulted by assuming the tests would

measure the same thing for both groups of students even though the Indian students' come from drastically differing cultural and socio-economic situations.

In his work with the Bureau of Indian Services, Witherspoon (1962) deduced that some of the aspects of the assessment that led to the failures of Indian children were controllable in the assessment design. He saw the following as controllable aspects that could be altered to reduce the assessment failures of the Indian children:

> 1. The predominantly verbal content of the tests; 2. The necessity for speed; 3. The observed difficulty Indian children had with separate answer sheets; and 4. An apparent lack of motivation, which was thought to come at least in part from the difficulties listed above. (Witherspoon)

Witherspoon revised parts of previous-existing measures and combined them in what he thought, was the most meaningful way for the Indian students. In his examination of the students' test scores, both Indian and non-Indian, he found that there remained some consistencies with past research findings regarding Indian students and assessment. He found that the disparity between Indian and non-Indian children grew as the grade level increased, and the most significant progress made by the Indian students in the core subjects was made in the first six to eight years in public school. Through further examination of his results, Witherspoon claimed that Indian children, in general, do not "begin with the same preparation, nor do they, as is sometimes claimed, keep up with their non-Indian peers through the first three, four, or six grades." Although Witherspoon claimed that his research findings did support previous research findings about Indian-student performance with regard to assessments, his paper did not address to what degree

his initial claim about a method of creating better-suited assessment tools for Indian children was affirmed.

In 1961, Keeler, Chairman of the Task Force on Indian Education, gave a speech at the Indian Education Conference at Arizona State University (Keeler, 1961). In this speech, Keeler brought forth what he thought was an erroneous assumption about the simplicity of the problem of Indian education: " . . . there is widespread public assumption that all you have to do to make the Indian child a facsimile of other American youngsters in education and habits of mind is just put them in the public schools with the white children, and the job is done." Keeler stated that for educators to ignore the intellectual habits of the Indian child, habits that are valued in their homes, is "a bad oversight." He spoke about the intellectual habits of the Indian such as " . . . craftsmanship habits that the Indian has, his painstaking ability in craftwork, his attention to detail, his patience. Also, the Indian child, without ever knowing that he's getting it, observes in nature things that the white man still hasn't seen". Keeler cited the words of Dr. Reifel (cite unknown) stating that " . . . the Indian's idea of time was different from the average American and white man because the American was future minded." He claimed that the Indian does not seem to care too much about the duration of time. Keeler ended his speech with a declaration of his belief that "the Indian is an extremely intelligent individual" whose motivation is the factor that educators need to seek and find if they want to find success in educating the Indian child.

Although Keeler's (1961) comments were stereotypical, his perspective was one of respect and admiration of the Indian people. He attributed their lack of academic success to the lack of Indian content and perspective taking in the development of used

curriculum. This is a point that was made in nearly every paper written on Indian school performance at the time.

Evvard and Weaver (1966) examined the score results from the Wechsler Intelligence Scale for Children (WISC) (performance tests only) and the Bender Gestalt Motor Test, attained from native (Navajo) students enrolled in Grades three through seven. From the results, Evvard and Weaver attempted to derive " . . . implications valuable to counsellors and teachers" in an effort to inform counselling and educational practices. Their primary finding was that, for the native children, lack of achievement was not the result of a lack of innate intelligence. Although Evvard and Weaver wrote of specific problems with regard to the testing of the native children, they did not link these problems to a lack of intelligence. Evvard and Weaver described what happened when they attempted to administer the verbal section of their measures to the native children: "They began to wring their hands, to tap their feet, and to show other signs of emotional distress. They became unable to respond in any way that made it possible to score their performance at all."

In speaking of the performance of native children on intelligence tests, Evvard and Weaver (1966) concluded that if the "goal is to enable the Indian students to achieve the same success with these materials as does the Anglo child, then we must deal with the differences in language, environment, and value systems of the Indian student." Alternatively, they offered that if the goal is not to imitate the non-native child's achievements, then an instrument that does not depend on the English language and that is built around the Indian environment is more appropriate. Evvard and Weaver stated

that if either of the above two courses of action were taken, the Indian students could perform equally well and have equal levels of achievement as non-Indian students.

The findings of Evvard and Weaver (1966) would have been better served if they had presented some of the data in their paper, or particular reasons for lower achievement levels for aboriginal children.

Recent History of Aboriginal Education with Regard to Academic Success

Demmert (2001) summarized the current poor state of aboriginal education in a manner that is echoed throughout Canada and United States:

> Except for the tribal schools, responsibility for the education of Native children and youth has been transferred from the tribes to state agencies, mostly to administrators and other individuals outside the communities or tribes. With this transfer of responsibility, Native students began experiencing high levels of educational failure and a growing ambivalence toward learning traditional tribal knowledge and skills. They often exhibited indifference to formal Western academic learning as well. (p. 2)

In response to the poor performance of aboriginal students, recent research has revealed some new causes and influences, but has, not surprisingly, included most of the ones deduced from the 1960s research (presented above) as well. As highlighted by Demmert, academic performance of aboriginal students is still generally lower than non-aboriginal students, yet more recent research has shown that the following efforts have led to a lessening of the gap between the two: (a) a nurturing early childhood environment (Swisher & Deyhle, 1989); (b) inclusion of native language and cultural programs in the school (Ayoungman, 1991; Barnhardt, 1990, 1999; deMarrais, 1992; James, Chavez,

Beauvais, Edwards, & Oetting, 1995; Lipka & McCarty, 1994; Rubie, 1999; Slaughter & Lai, 1994; McLaughlin, 1992; Watahomigie & McCarty, 1994); and (c) enhanced community and parental influences on academic performance (Leveque, 1994; McInerney, McInerny, Ardington, & De Rachewiltz, 1997).

Regardless of all the work being done to enhance the academic performance of aboriginal children, these children continue to face unique problems and continue to remain at risk. The state of aboriginal child welfare is in need of improvement. Recently, members of five Canadian national aboriginal organizations worked to develop an improvement-based document titled "A National Children's Agenda: Developing a Shared Vision." In this document, the aboriginal leaders stated,

> Aboriginal children face far greater risk than most non-aboriginal children since among may things they are: twice as likely to be born prematurely, underweight, or die within their first year of life; three or four times more likely to suffer Sudden Infant Death Syndrome; fifteen to thirty-eight times more likely to suffer from the effects of Fetal Alcohol Syndrome; three times more likely to be physically disabled; six times more likely to die by injury, poisoning or violence; and five times more likely to take their own life. (Indian and Northern Affairs Canada, 2002)

In terms of the future success of aboriginal people in Canada, this group of five aboriginal organizations concluded that hope for the future is vested in their children. They stated, "Aboriginal people firmly believe that children represent the primary means through which cultures can preserve their traditions, heritage and languages. In this sense,

children are considered the hope of the future" (Indian and Northern Affairs Canada, 2002).

With this *hope of the future* vested in the children, attention needs to be put on specific aspects of aboriginal children's impediments to success. The one impediment for success that the present paper explored is the aboriginal children's comparatively poor performance on large-scale assessments. The specific case that was explored in this study was the British Columbia Ministry of Education provincial assessments on mathematics and reading. For the 2000-2001 academic year, students' scores from Grades 4 and 7 provincial assessments revealed that aboriginal students performed at a lower level than any other identified group, including ESL students (British Columbia Ministry of Education, 2002b).

The present study investigated student performance in an effort to find if the assessments were measuring the same constructs, with the same degree of accuracy, for both the aboriginal and non-aboriginal students. If the assessment scores proved to be comparable for the two groups, then the interpretation of the scores for the two groups should be the same. Conversely, if the assessment scores did not prove to be comparable for the two groups, then the interpretation of the scores for the two groups should not be the same. I believe that the findings of this study will assist both the aboriginal organizations as well as federal and provincial organizations to more accurately measure and interpret the performance of the aboriginal students in a manner that will contribute to their improvement.

CHAPTER THREE: METHOD

This chapter provides a detailed description of the methodology used in the study. The selection of the instrument and the data for use in this study are discussed, followed by a description of the methodological procedures used to examine the data.

Instrument

The results from the 2001 British Columbia Ministry of Education's FSA numeracy and reading assessments for Grades 4 and 7 students were used in the present study. For both grades, the numeracy test booklets contained 32 multiple-choice (MC) items, and 4 open-ended (OE) items. For Grade 4, the reading examination booklet contained 35 MC items and 4 OE items. For Grade 7, the reading examination booklet contained 42 MC items and 4 OE items. For each of the four assessments, all students wrote the same items presented in the same order. For the present study, both item types were included in all analyses except for factor analysis as the specific method selected could not accommodate mixed item formats.

For both grades, the numeracy items can be categorized into four sub-content areas: (1) *number*; (2) *patterns and relationships*; (3) *shape and space*; and (4) *statistics and probability*. The *number* sub-content area included such topics as ratio, height estimations, and rounding numbers. The *patterns and relationships* sub-content area included such topics as identification of number patterns and identification of shape patterns. The *shape and space* sub-content area included such topics as estimating height, converting from different metric units, and telling time. The *statistics and probability* sub-content area included such topics as reading bar graphs and calculating probabilities. Summaries of items by sub-content area, context, and item type for Grades 4 and 7

Numeracy items are provided in Tables 1 and 2 below. For example, from Table 1, the *number* sub-content area had 15 items, the context of field trip had 18 items, and the MC item type had 32 items. See Appendixes A and B for full details of each item with regards to sub-content area, context, item type, and number-of-words per item for Grades 4 and 7 Numeracy items.

Table 1

Grade 4 Numeracy Item Summary

Sub-Content Area	Total # of Items	Context			
		Field Trip		Activity Day	
		MC	OE	MC	OE
Number	15	6	0	7	2
Patterns & Relationships	7	3	1	3	0
Shape & Space	8	5	0	3	0
Statistics & Probability	6	2	1	3	0

Table 2

Grade 7 Numeracy Item Summary

Sub-Content Area	Total # of Items	Context			
		Ski Trip		School Fun Fair	
		MC	OE	MC	OE
Number	15	7	1	7	0
Patterns & Relationships	7	2	1	3	1
Shape & Space	9	4	0	4	1
Statistics & Probability	5	3	0	2	0

For both grades, the reading items can be categorized into three sub-content areas: 1) *critical analysis*; 2) *identify and interpret key concept and main idea*; and 3) *locate, interpret, organize details*. The *critical analysis* sub-content area items included such aspects as requiring students to read between the lines to infer correct meaning, and students' recognizing the author's purpose for presenting specific information in an article. The *identify and interpret key concept and main idea* sub-content area items included such aspects as distinguishing events in the story from the main idea of the story

and distinguishing trivial points of the story from the main idea of the story. The *locate, interpret, organize details* sub-content area items included such aspects as separating one's own ideas from those portrayed in the story and using context clues to infer the meaning of a word. Summaries of items by sub-content area, context, and item type for Grades 4 and 7 Reading items are provided in Tables 3 and 4. For example, from Table 3, the *critical analysis* sub-content area had 6 items, the context of crime solving had 18 items, and the MC item type had 35 items. See Appendixes C and D for the full details of each item with regards to sub-content area, context, item type, and number-of-words per item for Grades 4 and 7 Reading items.

Table 3

Grade 4 Reading Item Summary

Context	Total # of Items	Sub-Content Area					
		Critical Analysis		Identify & Organize Key Concept & Main Idea		Locate, Interpret, & Organize Details	
		MC	OE	MC	OE	MC	OE
Crime Solving	4	0	0	1	0	3	0
Rain	5	0	0	1	0	3	1
House Pets	5	1	0	0	0	4	0
Polar Bears	5	0	0	1	0	3	1
Frogs & Toads	5	1	0	0	0	4	0
Rabbits	5	1	1	0	0	3	0
Memory	5	1	0	1	0	3	0
Tree Growth	5	1	0	1	0	2	1

Table 4

Grade 7 Reading Item Summary

Context	Total # of Items	Sub-Content Area					
		Critical Analysis		Identify & Organize Key Concept & Main Idea		Locate, Interpret, & Organize Details	
		MC	OE	MC	OE	MC	OE
Egypt	6	1	0	0	0	4	1
Baseball Game	6	2	0	0	1	3	0
Whales	4	1	0	0	0	3	0
Ponies	6	2	0	1	0	2	1
Frogs & Toads	5	1	0	0	0	4	0
Goldfish	6	4	0	0	0	2	0
Willow Tree	7	1	0	0	0	5	1
Snakes	6	1	0	1	0	4	0

Participants

All students who completed the English version (as opposed to French) of the FSA assessment in May of 2001 were included in the study (for both grades, more than 99% of the students wrote the English version of the FSA assessment). Students were divided into two populations: the aboriginal population and the non-aboriginal population. The aboriginal student population, who made up approximately 8% of the population for both assessments, included all students who identified themselves, or were identified by their parents or guardians, as being of aboriginal ancestry. The non-aboriginal student population, who made up approximately 92% of the population for both assessments, included students who were not identified as aboriginal.

Procedures

The validity and comparability of test scores for the aboriginal and non-aboriginal populations was evaluated by examining the resulting factor structures, reliability estimates, item information functions, item response theory based parameters, and DIF items.

Factor Analyses

Exploratory factor analysis methods were used to identify the factor structures of the scores for each population in each grade. Factor analysis helped in understanding the structure of the correlation matrix of the items from the FSA assessments. In factor analysis studies, the purpose is to encapsulate the relationships among the variables "in a concise but accurate manner as an aid in conceptualization … by including the maximum amount of information from the original variables in as few derived variables, or factors, as possible to keep the solution understandable" (Gorsuch, 1983, p. 2).

Exploratory factor analysis models are of two types: the full component model and the common factor model, both being variants of the multivariate linear factor model. Each of these factor analysis models can be further divided depending on whether the researcher assumes the factors to be correlated or uncorrelated. The full component model "is based on perfect calculation of the variables from the components" (Gorsuch, 1983, p. 14) with the assumption of no other sources of variance; while the common factor model "includes sources of variance not attributable to the common factors" (Gorsuch, p. 14) that are unique to each variable. Selection of a model should be based on existing substantive theory in the area of research as well as cost and time efficiency issues (Gorsuch).

For this study, I selected the common factor model. It is the more complex of the two, but I believe that it is more meaningful to assume that, for the data at hand in the present study, sources of variance that are not attributable to the common factors do exist for each variable. Further, the model allows for the factors to be correlated; if they prove to be uncorrelated, the model will allow for that, and no error would be made.

There are two main steps in applying a common factor model that assumes correlated factors to actual data: (1) prepare the relevant correlation matrix and extract the initial factors; and (2) determine the optimal rotation. When the same number of factors are extracted for the two groups being compared, these steps are followed by the calculation of congruence coefficients which determine the degree of comparability of relevant factors for the different populations.

For all subsequent analysis, unweighted least squares (ULS) common factor analysis was used because this approach ignores the diagonal of the correlation matrix, thus maximally accounting for the variance in the off diagonal elements, resulting in minimized off-diagonal residuals.

Number of Factors to Extract

The first step in factor analysis was to find the number of factors that could adequately explain the correlations among the observed variables or items. The main concern in this step is finding out if a smaller number of factors can account for the covariation among a much larger number of variables. In order to perform an extraction of factors, the researcher must provide either the number of common factors to be extracted, or the criteria by which such a number can be determined.

I used three criteria by which such a number could be determined: (1) the scree test, (2) the Kaiser-Guttman criterion of retaining only those components with associated eigenvalues greater than one, and (3) the Likelihood Ratio Test of fit (Hakstian, Rogers, & Cattell, 1982). These procedures are described below in order of priority. The first criterion was the scree test. For this test, the relevant eigenvalues were plotted in descending order. The pattern that was revealed from the plot of eigenvalues was

examined for a sharp decent and a point of levelling off. The number of eigenvalues existing in the sharp descent (in the part of the plot before the eigenvalues started to level off) will correspond to the number of factors that will be extracted based on the scree test.

> To be precise about the scree procedure, it consists of entering unities in the diagonal of the given correlation matrix and extracting successive latent roots by a principal axis program down to n roots, when n is the number of variables (the last root may be zero)... When the size of these roots (their variances) are successively plotted, one characteristically gets a falling curved section followed by a straight line (or two or even three), at a much lesser angle to the horizontal, extending over perhaps ten successive eigenvalues each. The resemblance of this "debris" to the scree of rock debris funning straight, at an angle of "boulder stability" at the foot of a mountain gave the term "scree test". (Cattell & Vogalmann, 1977, p. 292)

The second criterion was the Kaiser-Guttman criterion of retaining only those components with associated eigenvalues greater than one (the eigenvalue-greater-than-one criteria). The third criterion was the Likelihood Ratio Test of fit (a Chi-Squared test using the maximum likelihood factor analysis procedure), which was used to perform statistical inference testing on the number of factors proposed to explain the data. For the present study an alpha level of 0.05 was used.

Not all criteria necessarily lead to a convergence on the same number of factors to be extracted. The three criteria were presented in order of priority, with the scree plot test being the best criterion because, unlike the other two, it is not subject to arbitrary boundaries of eigenvalues and alpha levels; it lets the data "talk" without pre-set

boundaries of importance. Hence, when the three criteria revealed different numbers of factors to be extracted, I used the scree test exclusively (Hakstian, Rogers, & Cattell, 1982). When there was agreement between two criteria, I used the number of factors for extraction that they converged upon, regardless of which two.

Optimal Rotation

For common factor analyses, "it is generally assumed that the initial factors will be rotated so that the factors meet criteria that make them more relevant to the purpose of the study" (Gorsuch, 1983, p. 176). The purpose of rotation, not unlike a linear transformation, is to shift the factors into the most parsimonious position without altering their relative values. In factor analyses, the most parsimonious position is referred to as the simple structure. For this study, the criteria I chose to use to determine simple structure were when a particular rotation (transformation) revealed (a) a maximum number of salient loadings, (b) a minimum number of complex loadings, and (c) a maximum number of hyperplanar coefficients. Gorsuch (1983) highlighted a general approach to determining salient loadings in that "anything that would be of interest for interpretation would be significant" (p. 208) if sample sizes were large. Typically, maximum salient loadings are determined when a variable's factor loading is larger than 0.30, but because of the very large sample sizes of the present study, this critical value may resolve to be different than 0.30 once the factor loadings are examined. Complex loadings occur when a variable has salient loadings with more than one factor, and hyperplanar coefficients occur when a variable has a factor loading of less than 0.10.

In search of the optimal rotation for each of the subgroups, I used oblique transformations because they allow for the correlation of factors. Further, the oblique

transformation procedure was selected based on Hakstian (1971) who showed that, in comparison to other currently prominent methods, the oblique procedure "produces solutions that best exemplify simple structure" (p. 175).

A common factor pattern rotated to simple structure was obtained for each subgroup via an Oblimin method, which is a method for oblique (nonorthogonal) rotation. This oblique technique rotates to simple structure through a variance or covariance function of the factor loadings. The Oblimin function allows for the researcher to test various values for a parameter that determines the degree of obliquity. For each subgroup, four values of the *delta*-parameter (the *delta*-parameter roughly determines the degree of obliquity) were used: 0.00, 0.25, 0.50 and 0.80. The corresponding results were examined in search of the *delta* value that led to the most exemplary simple structures. These values of *delta* were then used, and the resulting set of rotated factors was considered to have simple structure.

Congruence Coefficients

When there were equal numbers of factors extracted for each of the aboriginal and non-aboriginal populations for a given assessment, congruence coefficients were calculated for each factor in an effort to examine the degree of factor structure comparability, and are denoted by Φ_{lm}, where

$$\Phi_{lm} = \frac{\sum_{j=1}^{p} (a_{1jl})(a_{2jm})}{\sqrt{\sum_{j=1}^{p} a_{1jl}^{2} \sum_{j=1}^{p} a_{2jm}^{2}}} \quad . \qquad (1)$$

In this definition, a_{1jl} is a pattern element for the first sample, jth variable, lth factor, and p is the number of variables (Harman, 1976). There are no firm rules for interpreting congruence coefficients, and no procedures for testing significance. They range from +1.0 to –1.0, with +1.0 being perfect agreement and –1.0 being perfect inverse agreement (Harman). In general, 0.95 might be interpreted to be very comparable, 0.90 quite comparable, and 0.70 only somewhat comparable (Harman).

Reliability

Reliability estimates were used to indicate the degree to which individuals' scores would remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). The reliability estimates for each population were compared in an effort to identify differences in the internal consistency of the scores for the aboriginal and non-aboriginal populations. Considering that census-style data are being used, a difference in the reliability of test scores was deemed to exist if coefficients differed for the aboriginal and non-aboriginal populations in their respective grades and subject areas. In the absence of an appropriate external criterion for determining a meaningful difference between coefficients, I used my subjective judgement that a difference of more than a 0.10 would be considered a meaningful difference.

In an effort to determine the most appropriate reliability coefficient to use for this study, the best fitting model for reliability had to be determined. Qualls (1995) found that the two most common internal consistency estimation techniques, the split-halves and the Cronbach's alpha coefficient, "would generally be inappropriate for multidimensional instruments" (p. 111) as well as instruments with multiple item formats. Hence, when the test

parts have different *functional lengths*, then a *congeneric model* should be employed (Qualls). In identifying that the congeneric model fits the data best, it was recognised that for both Numeracy and Reading FSAs, there are multiple item formats as well as subcategories of items based on differing content, implying differences in the *functional lengths* of the test parts or items. Differences in functional length between two test parts can arise when different item types (multiple-choice and open-ended) and different scoring methods (i.e., dichotomously or polytomously) are employed in the same test. When such differences in functional length are expected, the use of the Feldt-Raju formula for estimation of the reliability of congeneric measures is deemed most appropriate, and denoted $F - R\rho xx'$, where

$$F - R\rho xx' = \frac{\sigma_x^2 - \sum \sigma^2 y_j}{(1 - \sum \hat{\lambda}_j^2)\sigma_x^2} \quad , \qquad (2)$$

$\sigma_{y_j}^2$ equals the observed part-score variances, $\hat{\lambda}_j$ represents the functional length of each test part (or item), and σ_X^2 is the total test variance (Qualls). For the present study, the Feldt-Raju reliability estimate in Equation 2 was used to calculate the reliabilities of the FSAs.

Item Information Functions

Item Response Theory (IRT) provides a method of describing and examining the measurement accuracy provided by test items. For each item, IRT can be used to estimate an *item information function*, denoted by $I_i(\theta)$, where

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad (i=1, 2, 3, \ldots, n). \tag{3}$$

$I_i(\theta)$ is the total information provided by item i at ability θ, $P_i(\theta)$ is the probability that a randomly selected examinee with ability θ will answer item i correctly, $P_i'(\theta)$ is the first derivative of $P_i(\theta)$, and $Q_i(\theta)$=[1- $P_i(\theta)$]. Item-information functions present the "contribution items can make to ability estimation at any point along the ability continuum" (Hambleton, 1989, p.162). In other words, the item-information functions reveal three things: 1) at which point along the ability continuum each item functions maximally, 2) measurement accuracy provided by the item at each ability level, and 3) the degree of discriminating power a particular item has at a particular ability level.

The item information functions were used to compare the aboriginal and non-aboriginal populations in two ways: 1) to determine the difference in the area under the item information functions for the two populations for each item of the four FSAs; and 2) to determine the degree of similarity of the standard error of measurement (SEM) values at various ability levels for each set of items that comprise the FSAs for the two populations. In the absence of appropriate external criteria for determining a meaningful difference between the two aspects of the item information functions that were analyzed, I used my subjective judgement to determine that a difference of more than a 20% would indicate a meaningful difference in area under the item information functions and a visual inspection of the similarities of the graphically represented SEM values across various ability levels will be discussed in an effort to indicate similarity of measurement accuracy across the populations.

IRT Based Parameters

In an effort to determine the degree of similarity of the IRT-based parameters, I examined the respective correlations for each FSA for the two populations. A correlation coefficient equal to or greater than 0.70 would be taken to indicate a meaningfully strong relationship between the IRT parameters.

Differential Item Functioning (DIF)

It is commonly agreed upon by psychometricians that an item is seen as showing DIF "if individuals having the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, et al., 1991, p. 110). In other words, an item shows DIF if the response functions across different subgroups are not the same (Hambleton et al.). Some of the sources of DIF are understood to be the use of language, content, or context that is not universally understood the same way amongst the subgroups. There are a number of different DIF detection methods. These methods tend to provide consistent DIF identification, yet many identify items as DIF when items are not DIF. To verify the DIF status of items, this study used two statistical DIF methods for identifying DIF items: the Linn-Harnisch and the Logistic Regressions DIF detection methods. Consensus of these two methods of DIF detection was used to ensure the DIF status of items. These methods of DIF detection were conducted in an effort to determine if aboriginal and non-aboriginal students, who have similar abilities, have differing probabilities of answering the items correctly.

Linn and Harnisch Method

The IRT-based LH method (Linn & Harnisch, 1981) of DIF detection compares the item characteristic curves (ICC) for groups. This method of DIF detection is

implemented by PARDUX software (Burket, 1998). The LH method assesses the fit of the model for the minority group using item ability parameter estimates.

The LH method obtains estimates of the parameters for the combined group, assuming a 3-parameter logistic model (3PL) for the MC items (Lord, 1980) and a two-parameter partial credit model (2PPC) model (Yen, 1993) for the OE items. For the 3-parameter logistic model, the parameter estimates and the probability P_{ij} (the estimated probability that person j would answer item i correctly) could be estimated where,

$$P_{ij} = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta_j - b_i)]}, \tag{4}$$

and a_i, b_i, c_i and θ are all parameter estimates. The probabilities for the reference group (the non-aboriginal group in this case) are based on item parameter estimates for the combined sample, and those for the target group (the aboriginal group in this case) are based on the aboriginal group sample. These probabilities are then compared for the target group and reference group using the observed proportion correct statistics. The proportion of people in a target group that are expected to correctly answer an item is,

$$P_{ig} = \frac{1}{n_g} \sum_{j \in g} P_{ij}, \tag{5}$$

where j is a member of a target group g, i is the item, and n_g is the number of people in the target group (Linn & Harnisch, 1981). And the proportion of people in the combined group is,

$$P_i = \frac{\sum_g n_g P_{ig}}{\sum_g n_g} . \qquad (6)$$

From this, the observed proportion correct on item *i* for target group *g,* can be calculated by dividing the number of people in target group *g* who answered item *i* correctly by the number of persons in the target group *g*. For the complete target group (including all ability levels), the observed proportion correct, O_{ig}, for an item *i*, is

$$O_i = \frac{\sum_g n_g O_{ig}}{\sum_g n_g} . \qquad (7)$$

From the above equations, an index of the degree to which members of a target group perform better or worse than the complete group, called the overall difference, can be easily calculated:

$$D_i = O_i - P_i . \qquad (8)$$

Further, such differences calculated for a number of different levels of ability can be used to identify if differences are not uniform over all ability levels. This can happen, for example, when an item with a small overall difference has a large positive difference for one ability value and a large negative difference for another ability value.

The item parameters for the OE items were obtained using the two-parameter partial credit model (2PPC) (Yen, 1993). The 2PPC model is a special case of Bock's (1972) nominal model and is equivalent to Muraki's (1992) generalized partial credit model. Similar to the generalized partial credit model, in 2PPC, items can vary in their discriminations and each item has location parameters, one less than the number of score levels. The nominal model states that the probability of an examinee with ability θ having a score at the k-th level of the j-th item is

$$P_{jk}(\theta) = P(X_j = k-1|\theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, \; k=1\ldots m_j, \tag{9}$$

where

$$Z_{jk} = A_{jk}\theta + C_{jk}. \tag{10}$$

For the special case of the 2PPC model used here, the following constraints were used:

$$A_{jk} = \alpha_j (k-1), \tag{11}$$

and

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}, \tag{12}$$

where $\gamma_{jo} = 0$, and β_j and γ_{ji} are the free parameters to be estimated from the data. The first constraint implies that items can vary in their discriminations and that higher item scores reflect higher ability levels. In the 2PPC model, for each item there are $m_j - 1$ independent γ_{ji} difficulty parameters and one α_j discrimination parameter; a total of m_j independent item parameters are estimated. To summarize, the LH based PARDUX software:

> computes the observed and expected mean response (expected and observed *p*-values) and the difference between them (observed minus predicted, p_{diff}) for each item by deciles of the specified group. The expected values are computed using the parameter estimates obtained from the entire sample, and the theta estimates (ability estimates) for the members of the specified subgroup. Based on the difference between expected and observed *p*-values, a *Z*-statistic is calculated for each decile and an average *Z*-statistic for the item is computed for identifying degree of DIF. (Ercikan, Gierl, McCreith, Puham, & Koh, in press, p. 8)

The LH method uses both statistical significance and effect size in categorizing DIF items. The Level 1 degree of DIF (free of DIF) includes items with |Z|<2.58; Level 2 degree of DIF (moderate differences) includes items where the absolute value of the expected mean difference is <0.10, and |Z|>=2.58; and Level 3 degree of DIF (relatively large differences) includes items where the absolute value of the expected mean difference is >=0.10, and |Z|>=2.58 (Ercikan & McCreith, 2002).

For items having Levels 2 or 3 DIF, the interpretation is that the parameters for these items are not invariant across the two groups and the model obtained for the total group will not fit the target/minority group. A negative difference implies the item

favours the reference group, in this case, the non-aboriginals. Conversely, a positive difference implies the item favours the target/minority group, in this case, the aboriginals.

Logistic Regression Method

The Logistic Regression (LR) method (Swaminathan & Rogers, 1990) of DIF detection is based on a model for predicting the probability of a correct response to an item using the standard logistic regression model for predicting a dichotomous dependent variable from given independent variables,

$$P(u=1 \mid \theta) = \frac{e^{(\beta_0+\beta_1\theta)}}{[1+e^{(\beta_0+\beta_1\theta)}]} \tag{13}$$

where u is the response to the item, θ is the observed ability of an individual, β_0 is the intercept parameter, and β_1 is the slope parameter.

For the LR method, separate logistic regression curves are calculated for the *target group* and for the *population minus the target group*. DIF is considered to be present only if the logistic regression curves for the two groups are not the same. If the regression curves differ in either a uniform or non-uniform manner, DIF will be considered present in the respective manner.

The regression model formula can be reformulated to include group membership for DIF testing:

$$P(u=1) = \frac{e^{z}}{[1+e^{z}]}, \tag{14}$$

where

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g). \quad (15)$$

In this formulation, g represents group membership, τ_2 corresponds to the group difference and τ_3 corresponds to the interaction between group and ability. If τ_2 is nonzero while τ_3 is zero, nonuniform DIF is inferred. If τ_3 is nonzero, whether or not τ_2 is zero, we can infer nonuniform DIF. The null hypotheses are $\tau_{2=0}$ and $\tau_{3=0}$ against the hypothesis that $\tau_2 \neq 0$ and $\tau_3 \neq 0$.

The Chi-square model fit statistic is the test of significance associated with the LR method that actualizes differently for *uniform* and *non-uniform* DIF. When testing for the presence of *uniform* DIF, the Chi-square statistic is examined for a statistical difference when a term for group membership is added. When testing for the presence of *non-uniform* DIF, the Chi-square statistic is examined for a statistical difference when a group-by-ability interaction term is added (as cited in Ercikan, 2003). It follows that,

$$\chi^2 = \hat{\tau}' C' (C \sum C')^{-1} C \hat{\tau}' \quad (16)$$

which has the χ^2 distribution with 2 degrees of freedom. When the calculated value of χ^2 exceeds $\chi^2_{\alpha:2}$, then DIF is found to exist; when the calculated value of χ^2 does not exceed $\chi^2_{\alpha:2}$, then no DIF is identified.

The results from both DIF detection methods were used for deciding on the DIF status of the items. Those items that were identified as DIF by both methods were considered to be truly functioning differently for the two groups.

CHAPTER FOUR: RESULTS

The results presented in this chapter are directly related to the research questions and the analysis procedures introduced in the earlier chapters. This chapter provides a detailed description of the participants, their scores on the FSA, and the results of the data analysis used to answer the research questions. In terms of a research question, the unitary concept of construct validation was applied to the data in an effort to explore the degree to which the FSA scores are comparable across aboriginal and non-aboriginal populations. Implied in the above exploration is finding out if the FSA measured the same constructs for both populations and whether or not score interpretations should be the same for both populations.

Participants

All students who completed the English version of the FSA assessment in May of 2001 were included in the study. Students were divided into two populations: the aboriginal population and the non-aboriginal population. Once the population was split into aboriginal and non-aboriginal populations, the percent of males and females for each population for each assessment was near a 50:50 split in all cases. Table 5 shows these percentages for each case. For example, for the aboriginal students who wrote the Grade 4 Numeracy assessment, 50 percent were female and 50 percent were male.

Table 5

Gender Percentages for each Assessment and Population

Assessment	Population			
	Aboriginal		Non-Aboriginal	
	% Females	% Males	% Females	% Males
Grade 4 Numeracy FSA	50	50	49	51
Grade 7 Numeracy FSA	51	49	49	51
Grade 4 Reading FSA	50	50	49	51
Grade 7 Reading FSA	51	49	50	50

Table 6 shows the number of students, mean scores, and standard deviations based on FSA scores for aboriginal and non-aboriginal students. For example, 3,339 aboriginal Grade 4 students wrote the Numeracy FSA; they had a mean score of 15.21 (out of 32) and a standard deviation of 5.99 on the MC items. The aboriginal students' mean scores for Grade 4 numeracy MC items are approximately 3.90 points lower than the corresponding non-aboriginal students (based on 42,547 students), with a difference in the population-based standard deviation of 0.14.

Table 6

Mean Scores and Standard Deviations for the FSA Scores

	Population					
	Aboriginal			Non-Aboriginal		
Assessment	N	MC *M(SD)*	OE *M(SD)*	N	MC *M(SD)*	OE *M(SD)*
Numeracy 4	3339	15.21 (5.99)	7.28 (3.79)	42547	19.11 (5.85)	9.30 (3.52)
Numeracy 7	3072	13.76 (5.54)	5.44 (3.32)	44043	18.16 (6.40)	7.81 (3.53)
Reading 4	3329	22.50 (7.19)	6.70 (4.17)	42579	26.98 (5.75)	9.27 (3.86)
Reading 7	3106	23.57 (6.92)	6.00 (4.26)	44126	28.00 (6.17)	8.82 (4.06)

Note. All MC numeracy mean scores are based on scores that have a maximum value of 32. All MC reading mean scores are based on scores that have a maximum value of 35 for Grade 4 and 42 for Grade 7. All OE mean scores are based on scores that have a maximum value of 16.

The data sets did not contain any missing values when they were delivered to me; rather, a score of zero had been entered for a student's non-response. For the cases where students had a zero entered for every item, I made the assumption that those students did not respond to the test, and in an effort to clean the data, those cases were removed from the data set. Table 7 shows the counts of removed cases. For example, for the Grade 4 Numeracy assessment, 241 aboriginal students were removed. This accounted for 6.7% of the total number of cases.

Table 7

Cases Removed Based on Non-Response

Assessment	Population			
	Aboriginal		Non-Aboriginal	
	Removed	% Removed	Removed	% Removed
Numeracy 4	241	6.7	1104	2.5
Numeracy 7	217	6.9	1099	2.4
Reading 4	246	6.6	1053	2.7
Reading 7	181	5.5	1093	2.4

Factor Analysis

The purpose of the factor analysis procedure was to examine the relationships among the test items making up the assessments. When factor analyses are conducted separately for groups, differences in these relationships provide information regarding the degree to which similar constructs are being assessed for the groups. For this study, these relationships were examined across the aboriginal and non-aboriginal populations for similarities and differences in the number of factors and their respective compositions as shown by factor loadings. This exploratory factor analysis identified the factor structures of the scores for each population in each grade, as well as their degree of congruence across the aboriginal and non-aboriginal populations. The extraction and rotation of factors was used to identify the factor structures of the scores for each population in each grade.

When examining the proceeding factorial results, the reader should keep in mind that the primary purpose of the present study was not to explore the dimensionality of the FSAs but rather to explore the comparability of the aboriginal and non-aboriginal populations with regard to score performance on the FSAs. Further, my interpretations of the factors were subjective to some extent, as is always the case with common factor analysis, and may not match that of all readers. Some of the factors are easily understood,

but other factors presented a greater interpretive challenge (Hakstian, Farrell, & Tweed, 2002).

Selection of the Number of Factors to be Extracted

For this common factor analysis, the number of factors to be extracted for each population was based on three values: the scree-plot count (see Appendixes E-L), the number of eigenvalues greater than one (see Appendixes M-N), and the number of factors that led to a fit via the Maximum Likelihood Ratio Test (see Appendixes O-P). In a principal components analysis (PCA), consideration of percent of variance explained would be one of the criteria for the selection of the number of factors to be extracted because the goal of a PCA "is to extract maximum variance from the data set with each component" (Tabachnick & Fidell, 1996, p.664). But for the common factor analysis, which is the analysis of this study, the goal "is to minimize squared differences between the observed and reproduced correlation matrices" (Tabachnick & Fidell, p.665); hence considering percent of variance explained as a criteria to select the number of factors, would be meaningless here because the purpose of the common factor analysis is to reduce total variance to common variance in an effort to reproduce the correlation matrix in the best possible manner.

The selection of number of factors to extract was based on consensus among the three methods. When consensus did not occur, then the value determined by at least two methods was used. If all three methods converged on different values, then the scree plot was used as it is the least affected by arbitrary limit settings (Cattell, 1966). Scree plots were examined in an effort to identify a major change in the slope of the line that follows the descent of the eigenvalues. I examined the scree plots from left to right, stopping my

count of eigenvalues when I came across a major change in the slope of the line that follows the decent of the eigenvalues. For the purpose of amplifying my ability to visually identify a major change in slope, I expanded the eigenvalue axis.

For the scores examined, Table 8 shows the number of factors that were selected for extraction for each assessment and population. For example, for the aboriginal student scores on the Grade 4 Numeracy assessment, the scree plot analysis suggested three factors, the eigenvalue analysis suggested six factors, and the maximum likelihood analysis suggested nine factors. Three factors were selected for the number of factors to be extracted. The number of factors selected for extraction for each of the matched student populations (e.g., aboriginal Grade 4 Numeracy was matched with non-aboriginal Grade 4 Numeracy) was the same for Grades 4 and 7 Reading and different for Grades 4 and 7 Numeracy. Hence, congruence coefficients were only calculated for each factor of the Grades 4 and 7 Reading scores.

Table 8

Selection of Number of Factors to be Extracted

	Population							
	Aboriginal				Non-Aboriginal			
Assessment	Scree	Eigenvalue	Maximum Likelihood	Selection	Scree	Eigenvalue	Maximum Likelihood	Selection
Num 4	3	6	9	3	4	4	17	4
Num 7	5	7	10	5	2	5	20	2
Read 4	3	5	11	3	3	6	20	3
Read 7	3	9	10	3	3	7	24	3

Note. "Num" represents "Numeracy", and "Read" represents "Reading"

Optimal Rotation

Once the number of factors to be extracted was decided upon, then the optimal rotation was determined. The purpose of rotation was to shift the factors into the most parsimonious position without altering their relative values; this parsimonious position leads to a simple structure. Oblique rotations were used in the present study. These types

of rotations focus on simplifying the factor structure by shifting the factors so that each item loads on the fewest number of factors. For the present study, the criteria of a simple structure was met when a particular rotation revealed a maximum number of salient loadings, a minimum number of complex loadings and a maximum number of hyperplanar coefficients. Because these were oblimin (oblique) rotations, the counts for the salient loadings, complex loadings, and hyperplanar coefficients were calculated for four different values (*delta* values) that represent a spectrum of degrees of obliquity in an effort to select the best fitting rotation. Based on the counts related to each value of *delta* for each subject and grade, and each population, the best value of *delta* was selected as shown in Tables 9-16. For example, in Table 9, the *delta* value of 0.25 for the Grade 4 aboriginal numeracy factor rotation resulted in 22 salient loadings, 0 complex loadings, and a hyperplanar count of 45; this was the best result of the four values of *delta*. The pattern matrix corresponding to each of the best fitting rotations was selected for the examination and composition of factor loadings.

When interpreting factors, a pattern loading is considered salient if it is "sufficiently high to assume that a relationship exists between the variable and the factor" and that "the variable can aid in interpreting the factor and vice versa" (Gorsuch, 1983, p. 208). Gorsuch stated that what may be a salient value for one analysis may not be a salient value for another, and that there is "no exact way to determine salient loadings" (p. 208). Gorsuch highlighted a general approach to determining salient loadings in that "anything that would be of interest for interpretation would be significant" (p. 208) if sample sizes were large, as they are for the present study. Loadings of approximately 0.25 for the rotated factor pattern solutions appeared to define a variable's relationship

with a factor when the other loadings for the other factors were near zero. Hence, for the presentation of salient loadings and the discussion of the composition and meaning of the factors, the loading value of 0.25 was used.

For the Grade 4 Numeracy scores, simple structure was found with a rotation *delta* value of 0.25 for the aboriginal population and 0.50 for the non-aboriginal population.

Table 9

Simple Structure for Grade 4 Aboriginal Numeracy

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	20	0	46	
0.25	22	0	45	Selected
0.50	*	*	*	
0.80	*	*	*	

Note. * indicates that the rotation failed to converge in 1000 iterations.

Table 10

Simple Structure for Grade 4 Non-Aboriginal Numeracy

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	13	0	66	
0.25	14	0	72	
0.50	26	0	54	Selected
0.80	*	*	*	

Note. * indicates that the rotation failed to converge in 1000 iterations.

For the Grade 4 Reading scores, simple structure was found with a rotation *delta* value of 0.50 for both the aboriginal and non-aboriginal populations.

Table 11

Simple Structure for Grade 4 Aboriginal Reading

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	35	2	42	
0.25	36	2	43	
0.50	37	2	42	Selected
0.80	77	32	6	

Note. * indicates that the rotation failed to converge in 1000 iterations.

Table 12

Simple Structure for Grade 4 Non-Aboriginal Reading

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	33	1	46	
0.25	33	1	46	
0.50	35	2	49	Selected
0.80	*	*	*	

Note. * indicates that the rotation failed to converge in 1000 iterations.

For the Grade 7 Numeracy scores, simple structure was found with a rotation *delta* value of 0.50 for the aboriginal population and 0.80 for the non-aboriginal population.

Table 13

Simple Structure for Grade 7 Aboriginal Numeracy

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	20	0	99	
0.25	21	0	99	
0.50	26	3	99	Selected
0.80	*	*	*	

Note. * indicates that the rotation failed to converge in 1000 iterations.

Table 14

Simple Structure for Grade 7 Non-Aboriginal Numeracy

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	30	1	14	
0.25	30	0	14	
0.50	31	0	16	
0.80	32	0	14	Selected

Note. * indicates that the rotation failed to converge in 1000 iterations.

For the Grade 7 Reading scores, simple structure was found with a rotation *delta* value of 0.50 for both the aboriginal and non-aboriginal populations.

Table 15

Simple Structure for Grade 7 Aboriginal Reading

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	30	2	62	
0.25	30	2	62	
0.50	32	3	59	Selected
0.80	*	*	*	

Note. * indicates that the rotation failed to converge in 1000 iterations.

Table 16

Simple Structure for Grade 7 Non-Aboriginal Reading

Delta	Salient Loadings	Complex Loadings	Hyperplanar Count	Simple Structure
0.00	25	0	68	
0.25	29	0	69	
0.50	32	1	59	Selected
0.80	97	39	13	

Note. * indicates that the rotation failed to converge in 1000 iterations.

Factor Patterns and Correlations

Once the number of factors to be extracted and the best fitting rotation were determined, the factor analysis procedure was performed on each of the eight assessment categories: the aboriginal and non-aboriginal populations for each of the four assessments. Executing this procedure led to the production of eight factor pattern matrices which are presented in complete form in Appendixes Q-X, and in a summary form in Tables 17-24. The summary tables include only the items with salient loadings which were matched with item details such as number-of-words per item, context, and sub-content area. The pattern loadings taken from the pattern matrices "may be interpreted as measures of the unique contribution each factor makes to the variance of the variables" (Rummel, 1970, p. 397). In other words, these pattern loadings indicate the degree of dependence of the variables on the different factors (Rummel). These loadings allow for the determination of the clusters of variables defined by the resulting oblique factors.

The summary Tables 17-24 present the item number, number-of-words per item, context, sub-content area, and pattern loading for each item that had a salient loading with a factor. Pairs of tables matched across the two populations were examined separately as well as in comparison to one another to identify similarities and differences in the patterns of the salient loadings. Table 17, which was the summary table for the Grade 4 Numeracy scores for the aboriginal population, showed three factors. Two of the factors were dominant and appeared to split by either the context of items, or the order of items. This means that the two factors were either associated with being set in one of two contexts, or with being set in the first half or the second half of the test. The

corresponding (paired) table, Table 18, which was the summary table for the non-aboriginal Grade 4 Numeracy scores, showed four factors. Upon visual examination, there did not appear to be an obvious pattern in these salient loadings, and they did not appear to be very similar to the corresponding aboriginal group's pattern of loadings.

Eight salient loading summary tables (Tables 17-24) are presented next, followed by a discussion of the similarities and differences of salient loading patterns for the paired populations. Further, in cases where there was the same number of factors for each paired population, congruence coefficients were calculated for each factor and are presented in the text. Following the salient loading summary tables are a presentation (Tables 25-32) and discussion of the eight corresponding factor correlation matrices which indicate the strength of the relationships between factors.

Table 17

Salient Loadings and Item Detail for Aboriginal Grade 4 Numeracy Scores

				Factor		
Item #	# of Words	Context	Sub-Content Area	1	2	3
1	21	Field Trip	Number			0.28
2	26	Field Trip	Patterns & Relationships	0.25		
3	31	Field Trip	Number	0.35		
4	32	Field Trip	Shape & Space			0.25
5	11	Field Trip	Shape & Space			0.44
6	7	Field Trip	Statistics & Probability			0.35
7	11	Field Trip	Number			0.29
8	41	Field Trip	Number			0.32
9	25	Field Trip	Shape & Space			0.37
10	26	Field Trip	Shape & Space			0.34
11	25	Field Trip	Number			0.45
12	25	Field Trip	Number			0.38
13	9	Field Trip	Statistics & Probability			0.33
14	37	Field Trip	Patterns & Relationships			0.36
15	20	Field Trip	Patterns & Relationships	0.31		
16	26	Field Trip	Shape & Space	0.28		
19	17	Activity Day	Number	0.12		
20	29	Activity Day	Patterns & Relationships	0.44		
21	18	Activity Day	Shape & Space	0.29		
22	25	Activity Day	Patterns & Relationships	0.50		
23	19	Activity Day	Patterns & Relationships	0.44		
24	26	Activity Day	Shape & Space	0.29		
25	18	Activity Day	Number	0.44		
26	19	Activity Day	Number	0.30		
28	26	Activity Day	Number	0.38		
30	25	Activity Day	Statistics & Probability		-0.43	
31	19	Activity Day	Statistics & Probability		-0.40	
32	20	Activity Day	Statistics & Probability	0.33		
34	26	Activity Day	Number	0.31		

Table 18

Salient Loadings and Item Detail for Non-Aboriginal Grade 4 Numeracy Scores

Item #	# of Words	Context	Sub-Content Area	Factor 1	2	3	4
1	21	Field Trip	Number	0.34			
3	31	Field Trip	Number			0.27	
4	32	Field Trip	Shape & Space			0.34	
5	11	Field Trip	Shape & Space			0.48	
7	11	Field Trip	Number			0.30	
8	41	Field Trip	Number	0.28			
9	25	Field Trip	Shape & Space			0.29	
10	26	Field Trip	Shape & Space			0.34	
11	25	Field Trip	Number			0.40	
12	25	Field Trip	Number	0.75			
13	9	Field Trip	Statistics & Probability			0.37	
14	37	Field Trip	Patterns & Relationships	0.36			
15	20	Field Trip	Patterns & Relationships				-0.26
16	26	Field Trip	Shape & Space	0.25			
19	17	Activity Day	Number				-0.28
20	29	Activity Day	Patterns & Relationships				-0.29
21	18	Activity Day	Shape & Space				-0.44
22	25	Activity Day	Patterns & Relationships				-0.32
23	19	Activity Day	Patterns & Relationships				-0.32
24	26	Activity Day	Shape & Space				-0.56
25	18	Activity Day	Number				-0.69
26	19	Activity Day	Number	0.82			
27	19	Activity Day	Number	0.53			
30	25	Activity Day	Statistics & Probability		0.78		
31	19	Activity Day	Statistics & Probability		0.67		
32	20	Activity Day	Statistics & Probability		0.35		
34	26	Activity Day	Number	0.25			

Table 19

Salient Loadings and Item Detail for Aboriginal Grade 7 Numeracy Scores

Item #	# of Words	Context	Sub-Content Area	Factor 1	2	3	4	5
1	36	Ski Trip	Number	0.38				-0.25
3	37	Ski Trip	Number		0.31		0.39	
4	25	Ski Trip	Number			-0.47		
5	32	Ski Trip	Number			-1.03		
6	19	Ski Trip	Number	0.47				-0.25
7	37	Ski Trip	Number		0.30			
8	28	Ski Trip	Patterns & Relations		0.52			
9	25	Ski Trip	Statistics & Probability		0.40			
10	31	Ski Trip	Patterns & Relations		0.43			
11	20	Ski Trip	Shape & Space		0.38			
12	15	Ski Trip	Shape & Space	0.29				
13	17	Ski Trip	Shape & Space		0.29			
15	44	Ski Trip	Statistics & Probability	0.33				
19	27	School Fun Fair	Number	0.40				
20	35	School Fun Fair	Shape & Space	0.52				
21	27	School Fun Fair	Statistics & Probability	0.59				
24	26	School Fun Fair	Patterns & Relations				0.31	
26	19	School Fun Fair	Patterns & Relations	0.39				
27	30	School Fun Fair	Patterns & Relations	0.50				
30	34	School Fun Fair	Number					0.26
31	23	School Fun Fair	Shape & Space	0.42				
32	25	School Fun Fair	Shape & Space	0.31				
33	28	School Fun Fair	Statistics & Probability					0.28

Table 20

Salient Loadings and Item Detail for Non-Aboriginal Grade 7 Numeracy Scores

Item #	# of Words	Context	Sub-Content Area	Factor 1	Factor 2
1	36	Ski Trip	Number		0.53
2	34	Ski Trip	Statistics & Probability	0.40	0.07
3	37	Ski Trip	Number	0.48	-0.13
4	25	Ski Trip	Number	0.36	-0.02
5	32	Ski Trip	Number	0.43	
6	19	Ski Trip	Number		0.54
7	37	Ski Trip	Number	0.32	
8	28	Ski Trip	Patterns & Relations	0.47	
9	25	Ski Trip	Statistics & Probability		0.28
10	31	Ski Trip	Patterns & Relations	0.47	
11	20	Ski Trip	Shape & Space	0.51	
12	15	Ski Trip	Shape & Space		0.31
13	17	Ski Trip	Shape & Space		0.31
14	16	Ski Trip	Shape & Space	0.31	
15	44	Ski Trip	Statistics & Probability		0.34
16	33	Ski Trip	Number	0.26	
19	27	School Fun Fair	Number		0.38
20	35	School Fun Fair	Shape & Space		0.36
21	27	School Fun Fair	Statistics & Probability		0.52
22	35	School Fun Fair	Number	0.35	
23	17	School Fun Fair	Number		0.28
24	26	School Fun Fair	Patterns & Relations	0.40	
25	32	School Fun Fair	Number	0.25	
26	19	School Fun Fair	Patterns & Relations		0.25
27	30	School Fun Fair	Patterns & Relations		0.43
28	20	School Fun Fair	Shape & Space	0.41	
29	19	School Fun Fair	Number	0.48	
30	34	School Fun Fair	Number	0.48	
31	23	School Fun Fair	Shape & Space		0.29
32	25	School Fun Fair	Shape & Space		0.25
33	28	School Fun Fair	Statistics & Probability	0.44	
34	46	School Fun Fair	Number	0.33	

Table 21

Salient Loadings and Item Detail for Aboriginal Grade 4 Reading Scores

Item #	# of Words	Context	Sub-Content Area	Factor 1	Factor 2	Factor 3
1	6	Crime Solving	Identify		-0.46	
2	7	Crime Solving	Locate		-0.52	
3	3	Crime Solving	Locate		-0.63	
4	27	Crime Solving	Locate		-0.35	
5	8	Rain	Identify		-0.34	
6	6	Rain	Locate		-0.49	
7	7	Rain	Locate		-0.42	
8	12	Rain	Locate		-0.47	
10	9	House Pets	Locate		-0.66	
11	8	House Pets	Locate		-0.65	
12	4	House Pets	Locate		-0.77	-0.25
13	9	House Pets	Locate		-0.58	
14	7	House Pets	Critical		-0.61	
15	8	Polar Bears	Identify		-0.41	
16	8	Polar Bears	Locate		-0.44	
17	11	Polar Bears	Locate		-0.31	
18	16	Polar Bears	Locate		-0.31	
20	5	Frogs & Toads	Locate	0.39		
21	7	Frogs & Toads	Locate	0.39		
22	9	Frogs & Toads	Locate	0.33		
24	10	Frogs & Toads	Critical	0.50		
25	6	Rabbits	Locate	0.33		
26	19	Rabbits	Locate	0.50		
27	11	Rabbits	Locate	0.71		
28	12	Rabbits	Critical	0.54		
30	6	Memory	Identify	0.49		
31	8	Memory	Critical	0.36		
32	6	Memory	Locate	0.42		
33	12	Memory	Locate	0.44		
34	8	Memory	Locate	0.63		
35	8	Tree Growth	Identify	0.63		
36	10	Tree Growth	Locate	0.36		0.34
37	9	Tree Growth	Locate	0.27		
38	11	Tree Growth	Critical	0.42		

Note. The sub-content areas of *identify*, *locate* and *critical* represent *identify and interpret key concepts and main ideas*; *locate, interpret and organize details*; and *critical analysis* respectively.

Table 22

Salient Loadings and Item Detail for Non-Aboriginal Grade 4 Reading Scores

Item #	# of Words	Context	Content	Factor 1	Factor 2	Factor 3
1	6	Crime Solving	Identify	-0.03	-0.49	
2	7	Crime Solving	Locate		-0.40	
3	3	Crime Solving	Locate		-0.40	
4	27	Crime Solving	Locate		-0.39	
5	8	Rain	Identify		-0.36	
6	6	Rain	Locate		-0.56	
7	7	Rain	Locate		-0.59	
8	12	Rain	Locate		-0.51	
10	9	House Pets	Locate		-0.29	0.45
11	8	House Pets	Locate		-0.36	0.33
12	4	House Pets	Locate		-0.33	0.59
13	9	House Pets	Locate		-0.36	
14	7	House Pets	Critical		-0.45	
15	8	Polar Bears	Identify		-0.38	
16	8	Polar Bears	Locate		-0.47	
17	11	Polar Bears	Locate		-0.36	
18	16	Polar Bears	Locate		-0.44	
20	5	Frogs & Toads	Locate	0.26		
21	7	Frogs & Toads	Locate	0.32		
22	9	Frogs & Toads	Locate	0.29		
24	10	Frogs & Toads	Critical	0.45		
25	6	Rabbits	Locate	0.31		
26	19	Rabbits	Locate	0.44		
27	11	Rabbits	Locate	0.55		
28	12	Rabbits	Critical	0.53		
30	6	Memory	Identify	0.35		
31	8	Memory	Critical	0.26		
32	6	Memory	Locate	0.32		
33	12	Memory	Locate	0.38		
34	8	Memory	Locate	0.53		
35	8	Tree Growth	Identify	0.55		
36	10	Tree Growth	Locate		-0.34	
38	11	Tree Growth	Critical	0.37		

Note. The sub-content areas of *identify*, *locate* and *critical* represent *identify and interpret key concepts and main ideas*; *locate, interpret and organize details*; and *critical analysis* respectively.

Table 23

Salient Loadings and Item Detail for Aboriginal Grade 7 Reading Scores

Item #	# of Words	Context	Sub-Content Area	Factor 1	Factor 2	Factor 3
1	15	Baseball Game	Critical			0.27
3	21	Baseball Game	Locate			0.36
4	10	Baseball Game	Critical			0.34
5	10	Baseball Game	Locate			0.25
12	12	Ponies	Critical			0.45
13	15	Ponies	Locate			0.26
14	14	Ponies	Critical			0.31
17	5	Frogs & Toads	Locate		-0.68	
18	7	Frogs & Toads	Locate		-0.84	
19	9	Frogs & Toads	Locate		-0.55	
20	12	Frogs & Toads	Locate		-0.29	
21	10	Frogs & Toads	Critical		-0.82	
22	8	Goldfish	Locate	0.54		
23	8	Goldfish	Locate	0.59		
24	11	Goldfish	Critical	0.25		
25	10	Goldfish	Critical	0.49		
27	9	Goldfish	Critical	0.32		
28	9	Willow Tree	Locate	0.27		0.35
30	17	Willow Tree	Locate			0.28
31	13	Willow Tree	Locate	0.28		0.28
33	12	Willow Tree	Critical	0.40		
35	5	Snakes	Identify	0.50		
36	8	Snakes	Locate	0.28		
37	5	Snakes	Critical	0.66		
38	9	Snakes	Locate	0.33		
39	6	Snakes	Locate	0.73		
40	6	Snakes	Locate	0.47		-0.26
41	9	Egypt	Locate	0.34		
42	16	Egypt	Critical	0.27		
43	13	Egypt	Locate			0.28

Note. The sub-content areas of *identify*, *locate* and *critical* represent *identify and interpret key concepts and main ideas*; *locate, interpret and organize details*; and *critical analysis* respectively.

Table 24

Salient Loadings and Item Detail for Non-Aboriginal Grade 7 Reading Scores

Item #	# of Words	Context	Sub-Content Area	Factor 1	Factor 2	Factor 3
1	15	Baseball Game	Critical			0.33
3	21	Baseball Game	Locate			0.51
4	10	Baseball Game	Critical			0.45
8	13	Whales	Critical			0.37
9	15	Whales	Locate			0.36
12	12	Ponies	Critical			0.50
13	15	Ponies	Locate			0.36
14	14	Ponies	Critical			0.38
15	9	Ponies	Locate			0.27
17	5	Frogs & Toads	Locate		-0.65	
18	7	Frogs & Toads	Locate		-0.92	
19	9	Frogs & Toads	Locate		-0.63	
20	12	Frogs & Toads	Locate		-0.31	
21	10	Frogs & Toads	Critical		-0.9	
22	8	Goldfish	Locate	0.49		
23	8	Goldfish	Locate	0.57		
25	10	Goldfish	Critical	0.41		
27	9	Goldfish	Critical	0.29		
28	9	Willow Tree	Locate	0.26		0.39
29	11	Willow Tree	Locate			0.27
30	17	Willow Tree	Locate			0.39
31	13	Willow Tree	Locate			0.41
33	12	Willow Tree	Critical	0.38		
35	5	Snakes	Identify	0.45		
37	5	Snakes	Critical	0.58		
38	9	Snakes	Locate	0.34		
39	6	Snakes	Locate	0.70		
40	6	Snakes	Locate	0.42		
43	13	Egypt	Locate			0.41
44	6	Egypt	Locate			0.36
45	6	Egypt	Locate			0.27

Note. The sub-content areas of *identify*, *locate* and *critical* represent *identify and interpret key concepts and main ideas*; *locate, interpret and organize details*; and *critical analysis* respectively.

Factor Loading Summary

Grade 4 Numeracy

The salient factor loadings for Grade 4 Numeracy are presented in Tables 17 and 18, for aboriginal and non-aboriginal groups, respectively. There were three factors for the aboriginal population and four factors for the non-aboriginal population. For the aboriginal population, the first and third factors were dominant with 15 and 12 salient loadings respectively. These two factors appeared to split by either the context of items, or the order of items in terms of the order of their presentation to the examinees. This means that the two factors were either associated with being set in one of two contexts, or with being set in the first half or the second half of the test. For the non-aboriginal population, visual inspection did not reveal any consistent patterns in these salient loadings except for the second factor that appeared to be related to the sub-content area of *statistics and probability* with three salient loadings. Further, the loadings of the two populations did not appear to be very similar to one another. Because there were a different number of factors for each population's results, no congruence coefficients could be calculated and interpreted.

As will be shown later in this study, there is a low degree of local dependence amongst the items which means that responses to the test questions are conditionally (locally) independent given the examinee's ability. This gives evidence that item order in itself was not the source of the difference in factors.

Grade 7 Numeracy

Table 19 represented the salient loadings for the aboriginal Grade 7 Numeracy factors, and Table 20 represented the salient loadings for the non-aboriginal Grade 7 Numeracy factors. For the aboriginal population, there were five factors; for the non-

aboriginal population there were two factors. For the aboriginal population, the first two factors were dominant, with 11 and 7 salient loadings respectively, and appeared to split by the order of items in terms of the order of their presentation to the examinees. For the non-aboriginal population, both factors appeared dominant with 18 and 17 salient loadings respectively. Visual inspection did not reveal any consistent patterns in these salient loadings. Further, the loadings of the two populations did not appear to be very similar to one another. Because there were a different number of factors for each population's results, no congruence coefficients could be calculated.

Grade 4 Reading

Table 21 represented the salient loadings for the aboriginal Grade 4 Reading factors, and Table 22 represented the salient loadings for the non-aboriginal Grade 4 Reading factors. For both populations, there were three factors. For both populations, there appeared to be two dominant factors and one minor one. For the aboriginal population, there were 17 salient loadings on each of the first two factors; for the non-aboriginal population, there were 16 and 18 salient loadings on the first two factors respectively. For both populations, the dominant factors appeared to be related to the order of items. There appeared to be an anomaly in that Item 36 loaded on the third factor for the aboriginal population and on the second factor for the non-aboriginal population. For the non-aboriginal population, this item was the only one from the second half of the test that loaded on the second factor. In this study, Item 36 was not found to be DIF, and there was little difference in the measurement accuracy provided by this item for the two populations. Examining sub-content area, context, and number-of-words per item of Item

36 did not reveal any information about why this item had a different factor loading pattern.

Congruence coefficients gave more evidence about the similarity of the three factors across the populations. The congruence coefficient for the first factor was 0.94, meaning that there was a very high degree of comparability for this factor across the two populations. The congruence coefficient for the second factor was 0.89, meaning that there was a high degree of comparability for this factor across the two populations. The congruence coefficient for the third factor was 0.71, meaning that there is a relatively low degree of comparability for this factor across the two populations. Factors 1 and 2 equalled or exceeded 0.85, which was stated to be a criterion for factorial equivalence by Harpur, Hakstian, and Hare (1988).

Grade 7 Reading

Table 23 represented the salient loadings for the aboriginal Grade 7 Reading factors, and Table 24 represented the salient loadings for the non-aboriginal Grade 7 Reading factors. For both populations, there were three factors. For both populations, the first factor appeared to represent the second half of the items in terms of the order of their presentation to the examinees, with 16 and 11 salient loadings respectively; this could be related to test fatigue. For both populations, the second factor appeared to be related to one specific context with five salient loadings matched with the five items of a *frogs and toads* context for both populations. For the aboriginal population, there was no consistent pattern in the 12 salient loadings of the third factor. But for the non-aboriginal population, the third factor appeared to be consistent with sets of items based on context. Of the 16 salient loadings on this factor, all are accounted for when aligned with specific

contexts. For example, there are three salient loadings that match with three items of a *baseball game* context, and four salient loadings that match with four items of a *ponies* context. As for the aboriginal population, the second factor appeared to be related to one specific context with 5 salient loadings matched with the five items of a *frogs and toads* context. Again, like the aboriginal population, there was no consistent pattern in the 11 salient loadings of the third factor. Congruence coefficients gave more evidence about the similarity of the three factors across the populations. The congruence coefficient for the first factor was 0.72, meaning that there is a relatively low degree of comparability for this factor for the two populations. The congruence coefficient for the second factor was 0.98, meaning that there is a very high degree of comparability for this factor for the two populations. The congruence coefficient for the third factor was 0.97, meaning that there is a very high degree of comparability for this factor for the two populations. Factors 2 and 3 equalled or exceeded 0.85, which was stated to be a criterion for factorial equivalence.

Factor Correlation Matrices

Correlations amongst factors are found through oblique rotation. Oblique rotation does not require that correlations exist, but it does allow for it. There is a resulting factor correlation matrix for each of the eight factor analyses performed in this study. These correlations indicate how the factors are related to each other for each test and population, and therefore provide evidence regarding the degree to which the factor structures are similar for the two groups. The correlation matrices for each test and population are presented below in Tables 25-32. Table 25 is the factor correlation matrix for the aboriginal Grade 4 Numeracy analysis. From this table, it can be seen that there

are weak relationships between Factors 1 and 2, and 2 and 3; this could be explained by the fact that there were very few salient loadings on Factor 2, and therefore, there were few strong loadings for the loadings of Factors 1 and 3 to correlate with. This explanation follows through to partly explain the high correlation between Factors 1 and 3, because there were many salient loadings on each factor. The only two items that loaded on the second factor were common with one another in that their sub-content area was statistics and probability, but these were not the only items with that sub-content area, so this alone does not explain the factor.

Table 25

Factor Correlation Matrix for Aboriginal Grade 4 Numeracy

Factor	1	2	3
1	1.00	-0.27	0.74
2		1.00	-0.31
3			1.00

Table 26 is the factor correlation matrix for the non-aboriginal Grade 4 Numeracy analysis. From this table, it can be seen that there is a strong relationship between all the factors.

Table 26

Factor Correlation Matrix for Non-Aboriginal Grade 4 Numeracy

Factor	1	2	3	4
1	1.00	0.81	0.85	-0.87
2		1.00	0.79	-0.79
3			1.00	-0.85
4				1.00

Table 27 is the factor correlation matrix for the aboriginal Grade 7 Numeracy analysis. From this table, it can be seen that there is a weak relationship between Factors 1 and 4, 2 and 4, 3 and 4, 1 and 5, 2 and 5, and 3 and 5; there is a weak negative

relationship between factors 4 and 5 and there is a moderately strong relationship between Factors 1 and 2, 1 and 3, and 2 and 3.

Table 27

Factor Correlation Matrix for Aboriginal Grade 7 Numeracy

Factor	1	2	3	4	5
1	1.00	0.61	-0.69	0.10	0.27
2		1.00	-0.65	0.09	0.57
3			1.00	-0.18	-0.36
4				1.00	0.00
5					1.00

Table 28 is the factor correlation matrix for the non-aboriginal Grade 7 Numeracy analysis. From this table, it can be seen that there is a strong relationship between the two factors.

Table 28

Factor Correlation Matrix for Non-Aboriginal Grade 7 Numeracy

Factor	1	2
1	1.00	0.82
2		1.00

Table 29 is the factor correlation matrix for the aboriginal Grade 4 Reading analysis. From this table, it can be seen that there is a weak relationship between Factors 3 and 2, and 1 and 3; and a strong relationship between Factors 1 and 3.

Table 29

Factor Correlation Matrix for Aboriginal Grade 4 Reading

Factor	1	2	3
1	1.00	-0.82	-0.13
2		1.00	0.10
3			1.00

Table 30 is the factor correlation matrix for the non-aboriginal Grade 4 Reading analysis. From this table, it can be seen that there is a strong relationship between Factors 1 and 2; and a weak relationship between Factors 1 and 3, and 2 and 3.

Table 30

Factor Correlation Matrix for Non-Aboriginal Grade 4 Reading

Factor	1	2	3
1	1.00	-0.82	0.35
2		1.00	-0.25
3			1.00

Table 31 is the factor correlation matrix for the aboriginal Grade 7 Reading analysis. From this table, it can be seen that there is a moderately strong relationship between Factors 1 and 2, and a weak relationship between Factors 1 and 3, and 2 and 3.

Table 31

Factor Correlation Matrix for Aboriginal Grade 7 Reading

Factor	1	2	3
1	1.00	-0.61	0.58
2		1.00	-0.53
3			1.00

Finally, Table 32 is the factor correlation matrix for the non-aboriginal Grade 7 Reading analysis. From this table, it can be seen that there is a moderately strong relationship amongst all three factors.

Table 32

Factor Correlation Matrix for Non-Aboriginal Grade 7 Reading

Factor	1	2	3
1	1.00	0.69	0.70
2		1.00	-0.73
3			1.00

The inter-factor correlations presented in Tables 25-32 suggest that there is a range in differences in meaning among these factors across the populations for all four assessments.

Reliability

The Feldt-Raju reliability estimates that are presented in Table 33 were used to indicate the degree to which individuals' scores would remain relatively consistent over repeated administration of the same test or alternate test forms (Crocker & Algina, 1986). These estimates, calculated separately for each population, were compared in an effort to identify differences in the internal consistency of the scores for the aboriginal and non-aboriginal populations. For Grades 4 and 7 Numeracy, the FSA scores were slightly more reliable for the non-aboriginal population than the aboriginal population; for Grades 4 and 7 Reading, the scores were slightly more reliable for the aboriginal population than the non-aboriginal population. When the non-aboriginal reliability coefficient is subtracted from the corresponding aboriginal coefficient, the difference is –0.01. These results indicate few to no differences between the reliability estimates of the Numeracy and Reading tests for the aboriginal and non-aboriginal groups.

Table 33

Feldt-Raju Reliability Estimates

Assessment	Population				Difference in Reliability
	Aboriginal		Non-Aboriginal		
	Reliability	N	Reliability	N	
Numeracy 4	0.81	3339	0.82	42547	-0.01
Numeracy 7	0.78	3072	0.84	44043	-0.06
Reading 4	0.88	3329	0.84	42579	0.04
Reading 7	0.83	3106	0.80	44126	0.03

Note. "Difference" is calculated by subtracting the non-Aboriginal Feldt-Raju reliability estimates from the corresponding Aboriginal ones.

IRT-Based Analyses

IRT-based analyses were conducted to compare the measurement accuracy provided at the item and test level and parameters of items for the two populations. First the appropriateness of the IRT models were examined by evaluating the degree to which IRT model assumptions were met. Evaluations of the IRT model assumptions were conducted by examining fit statistics, local item dependence statistics, and unidimensionality. All of these analyses capture deviations of response data from the model.

Evaluation of IRT Model Assumptions

Model Fit

The fit of item responses to the respective IRT models was evaluated by the Q_1 statistic (Yen, 1981). The Q_1 statistic compares observed and predicted trace lines and is a χ^2 statistic. The power of the fit statistic is affected by sample size, hence, for large sample sizes, small deviations from model predictions can be statistically significant.

When the Q_1 fit statistic was examined for the four sets of test scores, corresponding to the four assessments, model misfit was identified for items whose *Z*-statistic was greater than 4.60 (Ercikan, et al., 1998). For all four assessments, many of the items were identified as having poor fit. This was not surprising considering the influence of a relatively large sample size (over 40 000) on the χ^2 statistic. Upon visual examination of the Item Characteristic Curves (ICC) and the corresponding observed and predicted statistics for each item identified as *misfitting*, I found that the items do fit their respective models. Tables 34-37 present the observed and predicted values for each item that was found to be poorly fitting. A comparison between the observed proportions of a

given response, and the proportions that would be predicted using the estimated thetas and item parameters was made for each item to show that differences were minimal. For example, from Table 34, Item 36 was found to have the poorest fit to the Q_1 fit statistic. This item was assessed for fit to the 2PPC model with 45,856 cases. It produced a χ^2 statistic value of 770.21 with 35 degrees of freedom, and a *Z*-statistic value of 87.87, which is substantially over the threshold of 4.60 for identifying poorly fitting items. For this item, the difference between observed and predicted values of model fit is 0.0026, which is less than 1% different than the predicted value.

A summary of the worst fitting items for each assessment is presented in Table 38. The comparison of differences between the observed and predicted values were nearly zero for the poorest fitting item of each assessment indicating that the data do fit the models well and hence the Q_1 fit statistic is not an accurate indicator of fit with the large sample sizes of the present study.

Table 34

Model Goodness of Fit for Grade 4 Numeracy

Item	Model	χ^2 Statistic	DF	Total N	*Z*-Statistic	Observed	Predicted	Observed-Predicted
1	3PL	25.42	7	45856	4.92	0.7724	0.7690	0.0035
3	3PL	60.96	7	45856	14.42	0.6080	0.6095	-0.0015
5	3PL	51.88	7	45856	11.99	0.8812	0.8754	0.0057
9	3PL	24.57	7	45856	4.69	0.6789	0.6772	0.0016
11	3PL	27.13	7	45856	5.38	0.7974	0.7928	0.0046
12	3PL	41.58	7	45856	9.24	0.5237	0.5270	-0.0033
13	3PL	33.14	7	45856	6.99	0.6048	0.6045	0.0004
15	3PL	37.38	7	45856	8.12	0.4549	0.4605	-0.0056
24	3PL	24.54	7	45856	4.69	0.4834	0.4900	-0.0066
25	3PL	47.06	7	45856	10.71	0.3022	0.3118	-0.0096
26	3PL	39.48	7	45856	8.68	0.5907	0.5918	-0.0011
27	3PL	38.79	7	45856	8.50	0.6695	0.6696	-0.0001
30	3PL	46.47	7	45856	10.55	0.8706	0.8647	0.0059
32	3PL	70.46	7	45856	16.96	0.6850	0.6828	0.0022
33	3PL	39.39	7	45856	8.66	0.4210	0.4256	-0.0046
17	2PPC	606.81	35	45856	68.34	0.6246	0.6229	0.0017
18	2PPC	214.34	35	45856	21.44	0.4447	0.4476	-0.0028
35	2PPC	140.09	35	45856	12.56	0.5836	0.5835	0.0001
36	2PPC	770.21	35	45856	87.87	0.6344	0.6317	0.0026

Table 35

Model Goodness of Fit for Grade 7 Numeracy

Item	Model	χ^2 Statistic	DF	Total N	Z-Statistic	Observed	Predicted	Observed-Predicted
1	3PL	127.61	7	47056	32.23	0.7903	0.7854	0.0049
3	3PL	38.89	7	47056	8.52	0.3383	0.3445	-0.0062
5	3PL	36.77	7	47056	7.96	0.4863	0.4908	-0.0044
6	3PL	90.08	7	47056	22.20	0.9032	0.8953	0.0079
7	3PL	36.62	7	47056	7.92	0.4956	0.4987	-0.0031
8	3PL	63.41	7	47056	15.08	0.4552	0.4608	-0.0055
9	3PL	79.38	7	47056	19.34	0.5778	0.5779	-0.0001
10	3PL	37.45	7	47056	8.14	0.3282	0.3348	-0.0066
11	3PL	158.30	7	47056	40.44	0.4334	0.4401	-0.0068
13	3PL	36.32	7	47056	7.84	0.5825	0.5836	-0.0010
15	3PL	32.68	7	47056	6.86	0.7048	0.7001	0.0047
19	3PL	30.08	7	47056	6.17	0.6270	0.6267	0.0003
20	3PL	24.30	7	47056	4.62	0.7756	0.7686	0.0070
21	3PL	138.91	7	47056	35.26	0.8303	0.8220	0.0083
22	3PL	75.68	7	47056	18.36	0.2997	0.3051	-0.0054
24	3PL	32.08	7	47056	6.70	0.3911	0.3971	-0.0061
26	3PL	36.20	7	47056	7.80	0.6597	0.6594	0.0003
28	3PL	35.65	7	47056	7.66	0.4365	0.4413	-0.0048
29	3PL	69.54	7	47056	16.71	0.4612	0.4666	-0.0055
30	3PL	102.06	7	47056	25.41	0.4233	0.4290	-0.0056
33	3PL	65.45	7	47056	15.62	0.4149	0.4202	-0.0054
34	3PL	56.54	7	47056	13.24	0.5121	0.5163	-0.0042
17	2PPC	831.12	35	47056	95.15	0.5724	0.5699	0.0025
18	2PPC	294.99	35	47056	31.08	0.3881	0.3911	-0.0030
35	2PPC	218.88	35	47056	21.98	0.2703	0.2741	-0.0037
36	2PPC	660.03	35	47056	74.70	0.6800	0.6744	0.0056

Table 36

Model Goodness of Fit for Grade 4 Reading

Item	Model	χ^2 Statistic	DF	Total N	*Z*-Statistic	Observed	Predicted	Observed-Predicted
4	3PL	34.63	7	45774	7.39	0.7076	0.7106	-0.0030
5	3PL	37.11	7	45774	8.05	0.7450	0.7454	-0.0004
7	3PL	34.44	7	45774	7.33	0.7667	0.7665	0.0002
8	3PL	25.41	7	45774	4.92	0.8605	0.8581	0.0024
10	3PL	56.98	7	45774	13.36	0.9003	0.8963	0.0039
11	3PL	27.10	7	45774	5.37	0.8350	0.8320	0.0031
12	3PL	29.40	7	45774	5.99	0.9267	0.9212	0.0055
13	3PL	51.82	7	45774	11.98	0.9179	0.9126	0.0053
18	3PL	48.10	7	45774	10.98	0.7412	0.7419	-0.0007
20	3PL	8.97	7	45774	0.53	0.7597	0.7605	-0.0008
21	3PL	65.92	7	45774	15.75	0.8501	0.8389	0.0112
22	3PL	16.66	7	45774	2.58	0.7248	0.7274	-0.0026
23	3PL	165.64	7	45774	42.40	0.5670	0.5700	-0.0030
27	3PL	79.87	7	45774	19.48	0.8774	0.8745	0.0029
28	3PL	39.13	7	45774	8.59	0.8495	0.8457	0.0038
30	3PL	252.64	7	45774	65.65	0.7570	0.7588	-0.0018
31	3PL	35.34	7	45774	7.57	0.6125	0.6169	-0.0044
33	3PL	36.87	7	45774	7.98	0.6708	0.6735	-0.0027
34	3PL	25.19	7	45774	4.86	0.8524	0.8496	0.0029
35	3PL	27.66	7	45774	5.52	0.8887	0.8844	0.0043
36	3PL	102.68	7	45774	25.57	0.5173	0.5247	-0.0074
37	3PL	91.60	7	45774	22.61	0.3735	0.3857	-0.0122
38	3PL	29.23	7	45774	5.94	0.7197	0.7201	-0.0004
9	2PPC	315.95	35	45774	33.58	0.7645	0.7636	0.0009
19	2PPC	1005.15	35	45774	115.96	0.4148	0.4233	-0.0085
29	2PPC	1488.69	35	45774	173.75	0.4745	0.4813	-0.0068
39	2PPC	226.58	35	45774	22.90	0.6125	0.6146	-0.0022

Table 37

Model Goodness of Fit for Grade 7 Reading

Item	Model	χ^2 Statistic	DF	Total N	*Z*-Statistic	Observed	Predicted	Observed-Predicted
2	3PL	56.55	7	47227	13.24	0.9693	0.9650	0.0043
3	3PL	41.99	7	47227	9.35	0.3703	0.3791	-0.0089
4	3PL	32.89	7	47227	6.92	0.6437	0.6473	-0.0036
7	3PL	26.82	7	47227	5.30	0.8798	0.8767	0.0031
8	3PL	82.94	7	47227	20.30	0.4309	0.4390	-0.0081
11	3PL	131.42	7	47227	33.25	0.5091	0.5119	-0.0028
12	3PL	98.87	7	47227	24.55	0.2974	0.3075	-0.0101
14	3PL	28.88	7	47227	5.85	0.5646	0.5693	-0.0047
15	3PL	42.22	7	47227	9.41	0.6782	0.6807	-0.0024
18	3PL	28.37	7	47227	5.71	0.8670	0.8649	0.0020
19	3PL	33.41	7	47227	7.06	0.7347	0.7354	-0.0006
21	3PL	39.59	7	47227	8.71	0.8580	0.8565	0.0014
22	3PL	58.54	7	47227	13.78	0.8957	0.8891	0.0067
23	3PL	58.01	7	47227	13.63	0.9358	0.9300	0.0058
24	3PL	34.68	7	47227	7.40	0.6751	0.6763	-0.0011
25	3PL	78.35	7	47227	19.07	0.8308	0.8266	0.0042
26	3PL	44.29	7	47227	9.97	0.4868	0.4904	-0.0035
27	3PL	42.97	7	47227	9.61	0.7969	0.7958	0.0010
28	3PL	53.54	7	47227	12.44	0.7022	0.7018	0.0003
30	3PL	33.96	7	47227	7.20	0.5888	0.5919	-0.0031
31	3PL	38.85	7	47227	8.51	0.6196	0.6217	-0.0021
32	3PL	25.37	7	47227	4.91	0.4444	0.4475	-0.0031
33	3PL	52.93	7	47227	12.28	0.8340	0.8284	0.0056
35	3PL	182.36	7	47227	46.87	0.8550	0.8498	0.0051
37	3PL	71.30	7	47227	17.19	0.8618	0.8559	0.0059
39	3PL	175.70	7	47227	45.09	0.9091	0.9013	0.0078
40	3PL	802.33	7	47227	212.56	0.6873	0.6879	-0.0007
41	3PL	48.57	7	47227	11.11	0.7851	0.7842	0.0009
42	3PL	69.58	7	47227	16.72	0.5892	0.5900	-0.0008
43	3PL	73.19	7	47227	17.69	0.5436	0.5482	-0.0047
44	3PL	46.73	7	47227	10.62	0.4420	0.4481	-0.0061
45	3PL	30.53	7	47227	6.29	0.2789	0.2857	-0.0068
6	2PPC	156.84	35	47227	14.56	0.6544	0.6538	0.0007
16	2PPC	561.27	35	47227	62.90	0.5156	0.5184	-0.0029
34	2PPC	397.66	35	47227	43.35	0.6182	0.6182	0.0000
46	2PPC	278.92	35	47227	29.15	0.3698	0.3754	-0.0055

Table 38

Goodness of Fit Information for the Poorest Fitting Item for Each Test

Test	Item	Model	χ^2 Statistic	DF	Total N	*Z*-Statistic	Observed	Predicted	Observed-Predicted
4 Numeracy	36	2PPC	770.21	35	45856	87.87	0.6344	0.6317	0.0026
7 Numeracy	17	2PPC	831.12	35	47056	95.15	0.5724	0.5699	0.0025
4 Reading	29	2PPC	1488.69	35	45774	173.75	0.4745	0.4813	-0.0068
7 Reading	40	3PL	802.33	7	47227	212.56	0.6873	0.6879	-0.0007

Unidimensionality

The IRT models used in the present study assumed that the underlying ability being measured was unidimensional. The assumption of unidimensionality can be satisfied if the test data can be represented by a "*dominant* component or factor" (Hambleton, 1989, p. 150). Using this criterion the factor analytic results indicated that these tests were essentially unidimensional, as represented by a dominant factor for all the solutions (see Appendixes M & N). Specifically, for all four assessments, for both populations, there is one relatively large eigenvalue along with many relatively small ones, indicating that each assessment has one dominant component. Hence, the structure of the underlying ability being measured is found to be essentially unidimensional for each assessment (Reckase, 1979).

Local Item Dependence (LID)

The IRT models used in this study assume that the responses to the test questions are conditionally (locally) independent given the examinee's ability. Local independence requires that any two items be uncorrelated when ability is fixed (Lord, 1980). Local item independence was evaluated by the Q_3 statistic (Yen, 1984). The Q_3 statistic is the correlation between performance on two items, after taking into account overall test performance. When a value of the Q_3 statistic is greater than 0.20 (Ercikan et al., 1998),

then the corresponding pair of items is identified as displaying LID. If the Q_3 statistic values were found to be greater than 0.20 but were still relatively low for a relatively small number of items, the effect of LID on applying the IRT models is expected to be minimal.

For the Grade 4 Numeracy assessment, one item pair out of a possible 648 was found to display LID; for the Grade 7 Numeracy assessment, none of the item pairs out of a possible 648 was found to display LID; for the Grade 4 Reading assessment, two item pairs out of a possible 760 was found to display LID; and for the Grade 7 Reading assessment, five item pairs out of a possible 1058 was found to display LID. Table 39 below presents a summary of item pairs found to display LID for the four FSAs. For example from Table 39, for the Grade 4 Numeracy FSA, there was one item pair identified as displaying LID. The correlation between these items is 0.20 when the abilities influencing test performance are held constant (Hambleton et al., 1991).

Table 39

FSA Item Pairs Displaying Local Item Dependence

Assessment	Item Pair	$\|Q_3\|$ Value
Grade 4 Numeracy	35 & 36	0.200
Grade 7 Numeracy		
Grade 4 Reading	9 & 11	0.253
	10 & 11	0.207
Grade 7 Reading	15 & 16	0.266
	16 & 17	0.224
	15 & 19	0.230
	16 & 19	0.359
	17 & 19	0.244

Because of the small number of items displaying LID and the relatively small values of the corresponding Q_3 statistics, the effect of LID on the four FSA score sets when applying the IRT models is low.

Item Information Functions

Item information function values were computed based on the item parameter estimates using FLUX (Burket, 1993) software. The item information functions indicate the degree of measurement accuracy provided by test items for different ability levels. The maximum value of the item information function (information), and the area under the item information function (area), an indicator of total measurement accuracy provided by the test item, for each item were calculated for each population. In an effort to determine if measurement accuracy provided by the test items were comparable for the aboriginal and non-aboriginal populations, a difference in the area values was calculated for each item. In order to obtain comparable item information functions across the two populations, the scales were linked using the Stocking and Lord (1983) equating procedure. This method solves for the linear transformation that minimizes the squared differences between the test characteristic curves from two separate calibrations for a given ability level. This equating method does not affect the relative value of the item parameters to one another and, therefore, does not affect the definition of the scale or trait being estimated. The linking of the scales for aboriginal groups and non-aboriginal groups were based on a set of 10 test items that were considered to be comparable for the two groups and were DIF free. Table 40 presents the transformation values used in the linking procedure.

Table 40

Stocking and Lord Transformation Values

Test	Multiplicative Constant	Additive Constant
Numeracy 4	1.08	-0.70
Numeracy 7	0.90	-0.75
Reading 4	1.16	-0.73
Reading 7	1.09	-0.70

The transformed (and therefore comparable) information function values are shown in Tables 41-44. Table 41 shows that for Item 1 from the Grade 4 Numeracy assessment, the area under the item information function is 0.0076 for the aboriginal scores, and 0.0073 for the non-aboriginal scores and the height of the item information function at the location of maximum utility is 0.0300 for the aboriginal scores and 0.0270 for the non-aboriginal scores. For this item, the difference between the area across the aboriginal and non-aboriginal populations was 4%.

The sum of the differences in area across all items for the two populations is an indication of the difference in the amount of information provided by the FSA scores of the two populations. It was found that the Grade 4 Numeracy assessment and both reading assessments provided less information for the aboriginal group than the non-aboriginal group. In contrast, it was found that the Grade 7 Numeracy assessment provided more information for the aboriginal group than the non-aboriginal group. Specifically, the Grade 4 Numeracy FSA provided 4% *less* information for the aboriginal population; the Grade 7 Numeracy FSA provided 19% *more* information for the aboriginal population; the Grade 4 Reading FSA provided 3% *less* information for the aboriginal population; and the Grade 7 Reading FSA provided 3% *less* information for the aboriginal population.

Table 41

Grade 4 Numeracy Item Information

	Population					
	Aboriginal		Non-Aboriginal		Difference	
Item	Information	Area	Information	Area	Area	Area %
1	0.0300	0.0076	0.0270	0.0073	0.0003	4
2	0.0270	0.0076	0.0160	0.0069	0.0007	9
3	0.0820	0.0128	0.0910	0.0142	-0.0014	-11
4	0.0310	0.0092	0.0340	0.0105	-0.0013	-14
5	0.0550	0.0107	0.0520	0.0108	-0.0001	-1
6	0.0200	0.0070	0.0380	**0.0085**	-0.0015	-21
7	0.0250	0.0073	0.0340	0.0082	-0.0009	-12
8	0.0230	0.0076	0.0260	0.0088	-0.0012	-16
9	0.0220	0.0074	0.0270	0.0087	-0.0013	-18
10	0.0070	0.0032	0.0080	0.0031	0.0001	3
11	0.0480	0.0104	0.0570	0.0124	-0.0020	-19
12	0.0590	0.0118	0.0700	0.0137	-0.0019	-16
13	0.0300	0.0088	0.0310	0.0100	-0.0012	-14
14	0.0200	0.0058	0.0280	**0.0072**	-0.0014	-24
15	0.0960	0.0156	0.0830	0.0148	0.0008	5
16	0.0580	0.0116	0.0500	0.0110	0.0006	5
19	0.0500	0.0104	0.0460	0.0097	0.0007	7
20	0.0250	0.0085	0.0300	0.0100	-0.0015	-18
21	0.0240	0.0083	0.0300	**0.0100**	-0.0017	-20
22	0.0750	0.0134	0.0650	0.0139	-0.0005	-4
23	0.0700	0.0129	0.0550	0.0118	0.0011	9
24	0.0620	0.0104	0.0520	0.0096	0.0008	8
25	0.1240	0.0191	0.0980	0.0174	0.0017	9
26	0.0730	0.0138	0.0900	0.0155	-0.0017	-12
27	0.0570	0.0097	0.0700	0.0117	-0.0020	-21
28	0.0510	0.0122	0.0590	**0.0147**	-0.0025	-20
29	0.0240	0.0079	0.0270	0.0090	-0.0011	-14
30	0.0500	0.0102	0.0400	0.0086	0.0016	16
31	0.0530	0.0119	0.0500	0.0100	0.0019	16
32	0.0410	0.0107	0.0320	0.0098	0.0009	8
33	0.0060	0.0032	0.0240	**0.0062**	-0.0030	-94
34	0.0390	0.0100	0.0350	0.0101	-0.0001	-1
17	0.0040	0.0020	0.0030	0.0018	0.0002	10
18	0.0070	0.0034	0.0060	0.0030	0.0004	12
35	0.0130	**0.0051**	0.0090	0.0037	0.0014	27
36	0.0070	**0.0040**	0.0040	0.0024	0.0016	40
Total		0.3315		0.3450	-0.0135	

Note. Values in bold text indicate that the difference in area for the two populations was at least 20%. 'Difference' represents Aboriginal-Non-aboriginal

Table 42

Grade 7 Numeracy Item Information

	Population					
	Aboriginal		Non-Aboriginal		Difference	
Item	Information	Area	Information	Area	Area	Area %
1	0.0940	0.0143	0.0510	0.0117	0.0026	18
2	0.0700	0.0121	0.0880	0.0131	-0.0010	-8
3	0.1420	**0.0174**	0.0870	0.0138	0.0036	21
4	0.0420	0.0090	0.0440	0.0092	-0.0002	-2
5	0.1280	0.0182	0.1260	0.0173	0.0009	5
6	0.0770	**0.0115**	0.0440	0.0084	0.0031	27
7	0.0940	0.0138	0.0600	0.0117	0.0021	15
8	0.3100	**0.0286**	0.1740	0.0207	0.0079	28
9	0.1120	0.0150	0.0570	0.0125	0.0025	17
10	0.2330	**0.0241**	0.1080	0.0172	0.0069	29
11	0.2360	0.0194	0.1730	0.0175	0.0019	10
12	0.0370	0.0083	0.0320	0.0077	0.0006	7
13	0.0910	0.0141	0.0790	0.0135	0.0006	4
14	0.0360	0.0097	0.0460	0.0104	-0.0007	-7
15	0.0510	0.0110	0.0410	0.0107	0.0003	3
16	0.0630	0.0114	0.0600	0.0113	0.0001	1
19	0.1010	0.0158	0.0890	0.0146	0.0012	8
20	0.0560	0.0119	0.0390	0.0098	0.0021	18
21	0.0700	**0.0135**	0.0430	0.0096	0.0039	29
22	0.1580	**0.0181**	0.0620	0.0114	0.0067	37
23	0.0520	0.0096	0.0440	0.0092	0.0004	4
24	0.1210	0.0146	0.1020	0.0133	0.0013	9
25	0.0900	0.0128	0.0550	0.0104	0.0024	19
26	0.0430	0.0104	0.0560	0.0098	0.0006	6
27	0.0620	0.0112	0.0440	0.0099	0.0013	12
28	0.1200	0.0175	0.0960	0.0156	0.0019	11
29	0.3110	**0.0234**	0.0940	0.0132	0.0102	44
30	0.3460	**0.0256**	0.0850	0.0133	0.0123	48
31	0.0350	0.0094	0.0310	0.0089	0.0005	5
32	0.0450	0.0103	0.0290	0.0092	0.0011	11
33	0.2300	**0.0233**	0.1000	0.0165	0.0068	29
34	0.2030	**0.0201**	0.1100	0.0153	0.0048	24
17	0.0080	**0.0040**	0.0020	0.0012	0.0028	70
18	0.0140	**0.0053**	0.0070	0.0035	0.0018	34
35	0.0400	**0.0103**	0.0250	0.0073	0.0030	29
36	0.0080	**0.0037**	0.0020	0.0011	0.0026	70
Total		0.5087		0.4098	0.0989	

Note. Values in bold text indicate that the difference in area for the two populations was at least 20%. 'Difference' represents Aboriginal-Non-aboriginal

Table 43

Grade 4 Reading Item Information

	Population					
	Aboriginal		Non-Aboriginal		Difference	
Item	Information	Area	Information	Area	Area	Area %
1	0.0580	0.0121	0.0540	0.0109	0.0012	10
2	0.0400	0.0077	0.0450	0.0083	-0.0006	-8
3	0.0570	0.0103	0.0510	0.0095	0.0008	8
4	0.0130	0.0048	0.0280	**0.0065**	-0.0017	-35
5	0.0270	0.0080	0.0220	0.0066	0.0014	18
6	0.0650	0.0111	0.0660	0.0108	0.0003	3
7	0.0760	0.0119	0.0920	0.0127	-0.0008	-7
8	0.0570	0.0097	0.0790	0.0108	-0.0011	-11
10	0.0710	0.0144	0.0700	0.0137	0.0007	5
11	0.0610	0.0139	0.0670	0.0146	-0.0007	-5
12	0.1110	0.0180	0.1410	**0.0217**	-0.0037	-21
13	0.1030	0.0166	0.1240	0.0194	-0.0028	-17
14	0.0700	0.0158	0.0920	0.0177	-0.0019	-12
15	0.0550	0.0128	0.0500	0.0132	-0.0004	-3
16	0.1040	0.0172	0.1140	0.0192	-0.0020	-12
17	0.0230	0.0080	0.0320	0.0094	-0.0014	-18
18	0.0450	0.0107	0.0570	0.0103	0.0004	4
20	0.0350	0.0077	0.0230	0.0065	0.0012	16
21	0.0400	0.0065	0.0440	**0.0094**	-0.0029	-45
22	0.0590	0.0090	0.0520	0.0087	0.0003	3
23	0.0030	**0.0018**	0.0020	0.0011	0.0007	39
24	0.0510	0.0097	0.0500	0.0092	0.0005	5
25	0.0330	0.0089	0.0280	0.0087	0.0002	2
26	0.0440	0.0091	0.0610	**0.0117**	-0.0026	-29
27	0.0350	0.0083	0.0510	**0.0112**	-0.0029	-35
28	0.1170	0.0168	0.1390	0.0172	-0.0004	-2
30	0.0230	**0.0076**	0.0110	0.0042	0.0034	45
31	0.0460	0.0101	0.0550	0.0110	-0.0009	-9
32	0.0320	0.0088	0.0220	0.0080	0.0008	9
33	0.0440	0.0100	0.0460	0.0098	0.0002	2
34	0.0450	0.0107	0.0500	0.0115	-0.0008	-7
35	0.0720	0.0123	0.0760	**0.0148**	-0.0025	-20
36	0.1290	0.0184	0.1210	0.0176	0.0008	4
37	0.0970	0.0156	0.0910	0.0151	0.0005	3
38	0.0850	0.0144	0.0680	0.0126	0.0018	13
9	0.0010	0.0009	0.0020	0.0010	-0.0001	-11
19	0.0330	0.0111	0.0320	0.0107	0.0004	4
29	0.0380	**0.0120**	0.0250	0.0093	0.0027	23
39	0.0160	0.0071	0.0130	0.0058	0.0013	18
Total		0.4198		0.4304	-0.0106	

Note. Values in bold text indicate that the difference in area for the two populations was at least 20%. 'Difference' represents Aboriginal-Non-aboriginal

Table 44

Grade 7 Reading Item Information

	Population					
	Aboriginal		Non-Aboriginal		Difference	
Item	Information	Area	Information	Area	Area	Area %
1	0.0360	0.0076	0.0290	0.0065	0.0011	14
2	0.0320	0.0066	0.0470	**0.0083**	-0.0017	-26
3	0.0650	0.0128	0.0670	0.0124	0.0004	3
4	0.0910	**0.0123**	0.0540	0.0094	0.0029	24
5	0.0220	**0.0057**	0.0110	0.0043	0.0014	25
7	0.0240	0.0067	0.0210	0.0059	0.0008	12
8	0.0820	0.0117	0.0610	0.0099	0.0018	15
9	0.0440	0.0083	0.0480	0.0080	0.0003	4
10	0.0200	**0.0063**	0.0150	0.0045	0.0018	29
11	0.0120	**0.0046**	0.0070	0.0028	0.0018	39
12	0.1290	0.0177	0.0850	0.0146	0.0031	18
13	0.0400	0.0084	0.0350	0.0076	0.0008	10
14	0.0430	0.0097	0.0400	0.0084	0.0013	13
15	0.0210	0.0062	0.0260	0.0060	0.0002	3
17	0.0610	**0.0099**	0.0250	0.0064	0.0035	35
18	0.0830	**0.0102**	0.0540	0.0075	0.0027	26
19	0.0920	**0.0130**	0.0630	0.0102	0.0028	22
20	0.0050	0.0024	0.0180	**0.0048**	-0.0024	-100
21	0.0620	0.0092	0.0400	0.0090	0.0002	2
22	0.0790	**0.0168**	0.0560	0.0108	0.0060	36
23	0.0950	**0.0181**	0.0790	0.0128	0.0053	29
24	0.0130	0.0051	0.0120	0.0046	0.0005	10
25	0.0470	0.0126	0.0380	0.0105	0.0021	17
26	0.0090	0.0038	0.0100	0.0036	0.0002	5
27	0.0130	0.0050	0.0150	0.0052	-0.0002	-4
28	0.0840	0.0150	0.1000	0.0154	-0.0004	-3
30	0.0500	0.0106	0.0420	0.0096	0.0010	9
31	0.0620	0.0125	0.0560	0.0116	0.0009	7
32	0.0170	0.0057	0.0130	0.0049	0.0008	14
33	0.0510	0.0123	0.0780	**0.0156**	-0.0033	-27
35	0.0560	0.0138	0.0480	0.0121	0.0017	12
36	0.0250	0.0072	0.0210	0.0079	-0.0007	-10
37	0.0780	0.0169	0.0650	0.0148	0.0021	12
38	0.0410	0.0111	0.0430	0.0121	-0.0010	-9
39	0.1070	0.0199	0.1100	0.0200	-0.0001	-1
40	0.0080	**0.0036**	0.0020	0.0011	0.0025	69
41	0.0120	0.0045	0.0130	0.0046	-0.0001	-2
42	0.0230	**0.0089**	0.0190	0.0060	0.0029	33
43	0.0360	0.0103	0.0500	0.0104	-0.0001	-1
44	0.0280	0.0084	0.0410	0.0098	-0.0014	-17
45	0.1310	**0.0168**	0.0380	0.0084	0.0084	50
6	0.0150	0.0065	0.0640	**0.0187**	-0.0122	-188
16	0.0280	0.0100	0.1170	**0.0304**	-0.0204	-204
34	0.0220	0.0087	0.0900	**0.0259**	-0.0172	-198
46	0.0200	0.0082	0.0740	**0.0232**	-0.0150	-183
Total		0.4416		0.4565	-0.0149	

Note. Values in bold text indicate that the difference in area for the two populations was at least 20%. 'Difference' represents Aboriginal-Non-aboriginal

A test information function is obtained by summing the information of the items that contributed to the test score. The standard error of measurement (SEM) of a given ability level is the reciprocal of the square root of the test information at that ability level. In an effort to examine the similarities and differences of the accuracy of each FSA across the two populations, SEM as a function of scaled scores was presented for each assessment in Figures 1-4. These resulting functions visually depict the measurement accuracy provided by each test for the two groups across many ability levels. For the four figures, the vertical axis represents the SEM values, and the horizontal axis represents the scaled score values. When examining the figures, the reader should interpret the functions as such: the lower the values on the vertical axis, the higher the accuracy of the test. So when comparing the functions of the two populations, functions that are lower on the SEM axis for certain levels of the scaled scores, mean that their population's scores are more accurate at those points. It should be noted that the comparison of the two populations' functions could show that the scores for one population are more accurate than the other for some scaled score values, and the other is more accurate for other scaled scores. From Figure 1, the SEM functions for Grade 4 Numeracy revealed that the accuracy of the test is similar for both populations, but the test is more accurate for higher scores for the aboriginal population, and for lower scores for the non-aboriginal population. From Figure 2, the SEM functions for Grade 7 Numeracy revealed that the test is slightly more accurate for the aboriginal population than the non-aboriginal population. From Figure 3, the SEM functions for Grade 4 Reading revealed that the accuracy of the test is similar for both populations, but the test is more accurate for higher

scores for the aboriginal population, and for lower scores for the non-aboriginal population. From Figure 4, the SEM functions for Grade 7 Reading revealed that the accuracy of the test is similar for both populations, but the test is more accurate for higher scores for the aboriginal population, and for lower scores for the non-aboriginal population.

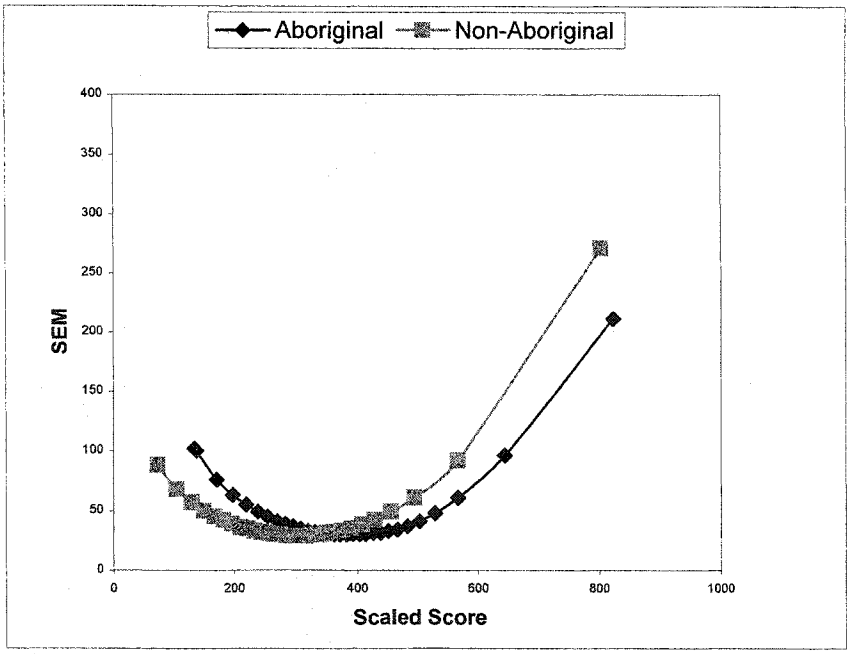


Figure 1. Grade 4 Numeracy: Standard error of measurement as a function of scaled scores for the aboriginal and non-aboriginal populations.

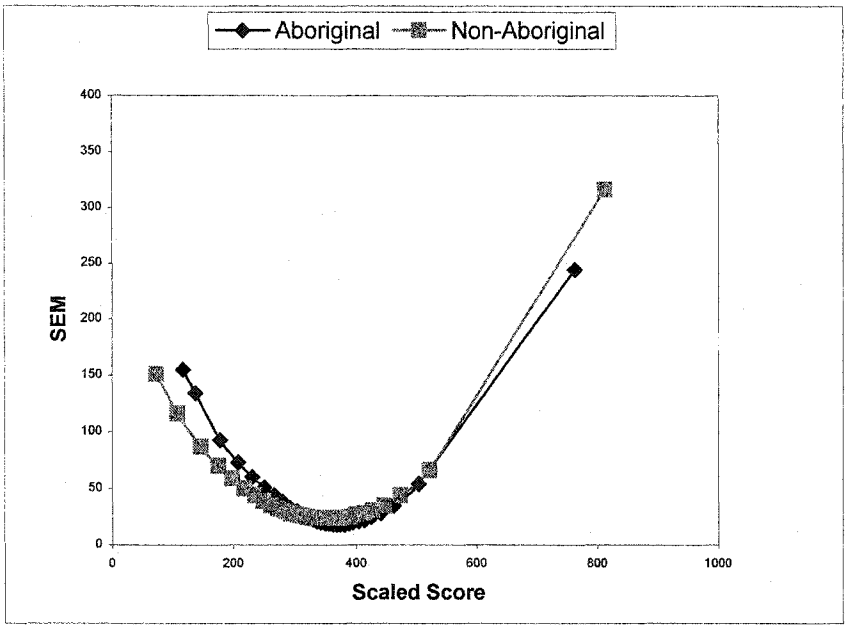


Figure 2. Grade 7 Numeracy: Standard error of measurement as a function of scaled scores for the aboriginal and non-aboriginal populations.

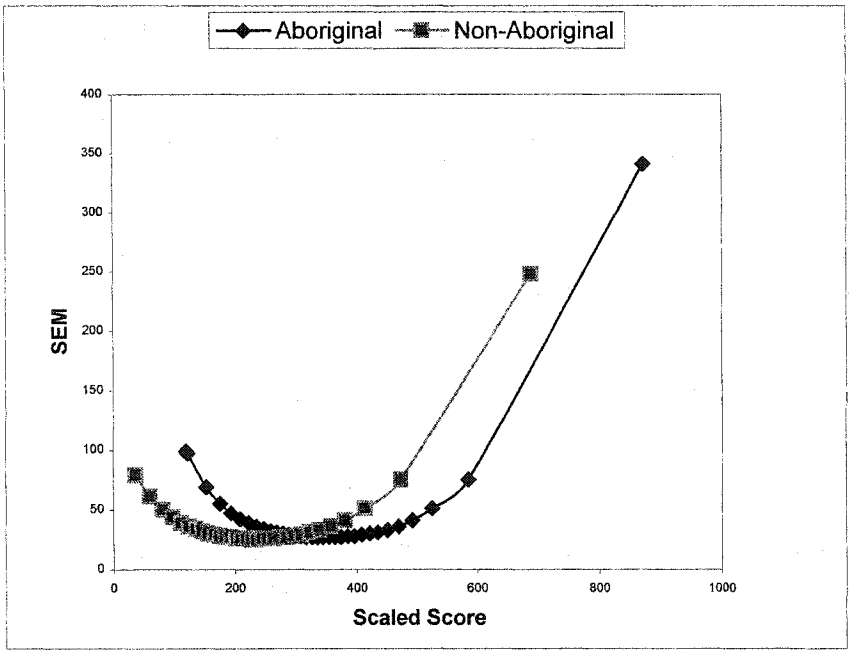


Figure 3. Grade 4 Reading: Standard error of measurement as a function of scaled scores for the aboriginal and non-aboriginal populations.

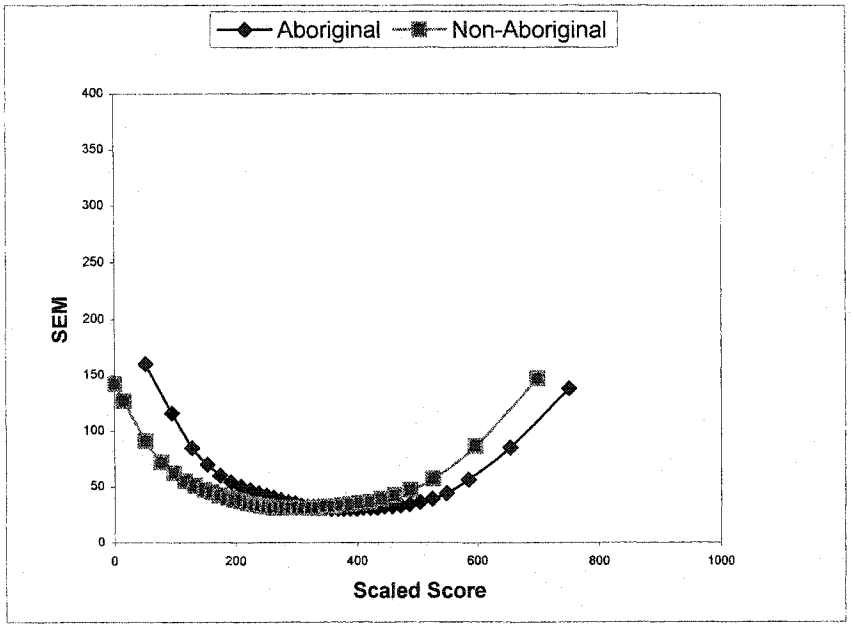


Figure 4. Grade 7 Reading: Standard error of measurement as a function of scaled scores for the aboriginal and non-aboriginal populations.

IRT Item Parameter Correlations

Further item comparability was revealed by examining the correlations between the IRT parameters for the two groups as shown in Table 49. For example, for the Grade 4 Numeracy scores, the correlation of the MC items' discrimination parameter *a* for the aboriginal and the non-aboriginal populations is 0.85; the correlation of the MC items' difficulty parameter *b* is 0.92; the correlation of the MC items' guessing/chance parameter *c* is 0.53; the correlation of the OE items' discrimination parameter α is 0.90; the correlation of the OE items' difficulty parameter β is 0.99; and the correlation of all of the items' difficulty parameter *p*, calculated as the percent of students who answered the item correctly, is 0.98. With the exception of the guessing/chance parameters, which were expected to be low because of the typically poor nature of estimating this parameter, all parameters for all assessments are highly correlated. These high correlations indicate similarity of functioning of test items for the two groups.

Table 45

Correlations between IRT Item Parameters for Aboriginals and Non-Aboriginals

Assessment	*a*	*b*	*c*	α	β	*p*
Numeracy Grade 4	0.85	0.92	**0.53**	0.90	0.99	0.98
Numeracy Grade 7	0.77	0.98	0.77	0.99	1.00	0.99
Reading Grade 4	0.89	0.94	**0.68**	0.98	0.71	0.98
Reading Grade 7	0.84	0.96	0.80	0.99	0.92	0.99

Note. Correlation values that are less than 0.70 are in bold text.

Differential Item Functioning

DIF analyses were conducted using both the Linn-Harnisch (LH) and Logistic Regression (LR) DIF detection methods. DIF status of items was determined when an item was identified as DIF by both methods. Table 46 presents the DIF items identified by both methods, and Table 47 summarizes the number of items found to be DIF by both

methods. For example from Table 45, for the Grade 4 Numeracy assessment, there were seven items (Items 1, 2, 4, 8, 10, 29, & 33) identified as DIF using the LR method that favoured the aboriginal population, two of which were identified by both methods of DIF detection, and nine items (Items 7, 9, 11, 13, 19, 20, 22, 23, & 28) identified as DIF that favoured the non-aboriginal population, two of which were identified by both methods of DIF detection. All items identified by both detection methods were consistent in their indication of which population, the aboriginal or the non-aboriginal students, was favoured by a particular DIF item.

Table 46

Identified DIF Items

Assessment	DIF Detection Method					
	Logistic Regression		Linn-Harnish		Both	
	Pro-Aboriginal	Pro-Non-Aboriginal	Pro-Aboriginal	Pro-Non-Aboriginal	Pro-Aboriginal	Pro-Non-Aboriginal
Grade 4 Numeracy	1, 2, 4, 8, 10, 29, 33	7, 9, 11, 13, 19, 20, 22, 23, 28	1, 14, 27, 29, 36	7, 18, 19, 25, 31, 32	1, 29	7, 19
Grade 7 Numeracy	23, 25	5, 8, 9, 12,19, 26, 28, 31, 32	6, 11, 14, 15, 20	8, 18, 35		8
Grade 4 Reading	2, 3, 5, 10, 11,12,13, 24	15, 20, 25, 34, 35, 38	2, 3, 11, 12, 13, 27, 28	15, 17, 19, 20, 29, 35, 36, 39	2, 3, 11, 12, 13	15, 20, 35
Grade 7 Reading	7, 11, 22, 25, 35, 37, 38, 39, 42	1, 2, 5, 10, 17, 18, 21, 27, 36	12, 22, 23, 25, 37, 39	3, 10, 16, 18, 20, 21, 34, 46	22, 25, 37, 39	10, 18, 21

Table 47

Number of DIF Items Found Using Both the LH and the LR DIF Detection Methods

Assessment	Pro-Aboriginal				Pro-Non-Aboriginal			
	LH-Level 2	LH-Level 3	LR	LH & LR	LH-Level 2	LH-Level 3	LR	LH & LR
Numeracy 4	5	0	7	2	6	0	9	2
Numeracy 7	5	0	2	0	3	0	9	1
Reading 4	7	0	9	5	7	1	6	3
Reading 7	6	0	9	4	8	0	9	3

Items that were determined to be DIF by both detection methods are presented with item details in Tables 48-51. This was done in an effort to identify patterns of DIF items based on specific characteristics of items as well as to present the corresponding estimated amount of DIF-related bias for each assessment. The item details presented in Tables 48-51 describe the item in terms of item type, context, and sub-content area, as well as describe the item in terms of the degree of DIF such as the *Z*-statistic, the number of participants in the focal group (the aboriginal population), the observed and predicted values of the mean responses over deciles, and the difference between the observed and

predicted mean responses over deciles. An example from Table 48, for Grade 4 Numeracy, Item 1 favours the aboriginal population, is a MC item type, is in the *number* sub-content area of Numeracy, and is written in the context of a field trip. Item 1 showed a *Z*-statistic value of 2.703, an observed value of the mean responses over deciles of 0.68, a predicted value of the mean responses over deciles of 0.66, with an observed-predicted value of 0.02. For interpretation purposes, the focus of the item DIF details was on the observed-predicted value and the sum of these values for each population favoured, for each assessment. From the manner in which the observed-predicted values were calculated, a positive result implies that the item favours the aboriginal population, and a negative result implies that the item favours the non-aboriginal population. For each FSA, these values were summed over the DIF items favouring a particular population to provide a crude estimation of the proportion of bias for each population in terms of item functioning.

For the Grade 4 Numeracy FSA, the proportion of bias in favour of the aboriginal population, based on two DIF items, was 0.04; the proportion of bias in favour of the non-aboriginal population, based on two DIF items, was 0.06.

Table 48

Grade 4 Numeracy DIF Item Details

Item	# of Words	Item Type	Sub-Content Area	Context	Z-Statistic	N	O	P	O-P
				Pro-Aboriginal					
1	21	MC	Number	Field trip	2.702	3339	0.68	0.66	0.02
29	23	MC	Number	Activity day	2.855	3339	0.54	0.52	0.02
				Pro-Non-Aboriginal					
7	11	MC	Number	Field trip	-3.984	3339	0.64	0.67	-0.03
19	17	MC	Number	Activity day	-3.943	3339	0.48	0.51	-0.03

Note. 'O' represents 'Observed', and 'P' represents 'Predicted'.

For the Grade 7 Numeracy FSA, the proportion of bias in favour of the aboriginal population, because there were no DIF items, was 0.00; the proportion of bias in favour of the non-aboriginal population, based on one DIF item, was 0.02.

Table 49

Grade 7 Numeracy DIF Item Details

Item	# of Words	Item Type	Sub-Content Area	Context	Z-Statistic	N	O	P	O-P
				Pro-Non-Aboriginal					
8	28	MC	Patterns & Relations	Ski trip	-2.794	3071	0.27	0.29	-0.02

Note. 'O' represents 'Observed', and 'P' represents 'Predicted'.

For the Grade 4 Reading FSA, the proportion of bias in favour of the aboriginal population, based on five DIF items, was 0.13; the proportion of bias in favour of the non-aboriginal population, based on three DIF items, was 0.10.

Table 50

Grade 4 Reading DIF Item Details

Item	# of Words	Item Type	Sub-Content Area	Context	Z-Statistic	N	O	P	O-P
				Pro-Aboriginal					
2	7	MC	Locate	Crime solving	3.130	3328	0.90	0.88	0.02
3	3	MC	Locate	Crime solving	4.042	3328	0.87	0.85	0.02
11	8	MC	Locate	House pets	4.189	3328	0.74	0.71	0.03
12	4	MC	Locate	House pets	4.270	3328	0.85	0.82	0.03
13	9	MC	Locate	House pets	4.763	3328	0.84	0.80	0.03
				Pro-Non-Aboriginal					
15	8	MC	Identify	Polar bears	-5.263	3328	0.49	0.53	-0.04
20	5	MC	Locate	Frogs & Toads	-4.113	3328	0.62	0.66	-0.03
35	8	MC	Identify	Tree growth	-4.330	3328	0.33	0.36	-0.03

Note. 'O' represents 'Observed', and 'P' represents 'Predicted'. The sub-content areas of "identify", "locate" and "critical" represent "identify and interpret key concepts and main ideas", "locate, interpret and organize details", and "critical analysis" respectively.

For the Grade 7 Reading FSA, the proportion of bias in favour of the aboriginal population, based on four DIF items, was 0.13; the proportion of bias in favour of the non-aboriginal population, based on three DIF item, was 0.09.

Table 51

Grade 7 Reading DIF Item Details

Item	# of Words	Item Type	Sub-Content Area	Context	Z-Statistic	N	O	P	O-P
				Pro-Aboriginal					
22	8	MC	Locate,	Goldfish	4.379	3106	0.82	0.78	0.03
25	10	MC	Critical	Goldfish	2.728	3106	0.74	0.72	0.02
37	5	MC	Critical	Snakes	5.022	3106	0.78	0.74	0.04
39	6	MC	Locate	Snakes	4.305	3106	0.83	0.79	0.04
				Pro-Non-Aboriginal					
10	10	MC	Locate	Whales	-3.216	3106	0.64	0.67	-0.03
18	7	MC	Locate	Frogs & Toads	-3.180	3106	0.74	0.76	-0.03
21	10	MC	Critical	Frogs & Toads	-2.931	3106	0.74	0.76	-0.03

Note. 'O' represents 'Observed', and 'P' represents 'Predicted'. The sub-content areas of "identify", "locate" and "critical" represent "identify and interpret key concepts and main ideas", "locate, interpret and organize details", and "critical analysis" respectively.

For the Grade 4 Numeracy scores, neither number-of-words per item, item type, sub-content area, nor context can be used to suggest an explanation of the sources of DIF. The minimal nature of DIF items for the Grade 7 Numeracy scores, only one in total, indicates that in this case, there is no evidence to suggest that a specific item detail is the source of DIF. The item details for the DIF items identified for each population of Grade 4 Reading scores indicated that items of the *locate, interpret, and organize* sub-content area favoured the aboriginal population. Because of the similarity of the item details for the DIF items identified for each population for Grade 7 Reading scores, neither number-of-words per item, item type, sub-content area, nor context can be used to suggest an explanation of the sources of DIF.

Results Summary

In this study, the dimensional structure of the test was found to be very similar for Grades 4 and 7 Reading FSAs, with high levels of comparability based on the congruence coefficients calculated for the factors across the two populations. The dimensional structure of the test was not found to be very similar for Grades 4 and 7 Numeracy FSAs. For all four FSAs, the reliability estimates of the test scores were found to be high and similar for both populations, meaning that the degree to which individuals' scores would remain relatively consistent over repeated administration of the same test or alternate test forms was high. For each of the Grade 4 Numeracy and the Grades 4 and 7 Reading FSAs, there was less than 4% difference in the level of accuracy of scores between the two populations, and for the Grade 7 Numeracy FSA, there was 19% difference in the level of accuracy of scores between the two populations, favouring the aboriginal population. For all four FSAs, relatively low degrees of DIF were found. In terms of the number of DIF items for each FSA, there were two favouring each of the two populations for the Grade 4 Numeracy; there was only one for the Grade 7 Numeracy and it favoured the non-aboriginal population; there were five favouring the aboriginal population and three favouring the non-aboriginal population for Grade 4 Reading; and there were four favouring the aboriginal population and three favouring the non-aboriginal population for Grade 7 Reading. Further, for all four FSAs, there was a minimal difference across the populations in the proportions of bias related to the collective impact of all DIF items.

CHAPTER FIVE: DISCUSSION

This study examined the psychometric properties of scores from a large-scale assessment in an effort to establish the comparability (or lack thereof) of the interpretation of the scores for aboriginal and non-aboriginal students. The comparability of the scores was addressed by the following two research questions: (1) *Are scores from the Foundation Skills Assessment comparable across aboriginal and non-aboriginal populations*?; and (2) *Should score interpretations be the same for both populations*? In answering these research questions, I performed four statistical analyses with the FSA data: factor analysis, comparison of reliability estimates, comparison of item information functions, and an analysis of differential item functioning. Because both research questions are associated with the unitary notion of validity, and the unitary notion of validity can only be assessed with an integrative look at all of the findings, I will present a short summary of the findings before answering the research questions in detail.

Summary of Statistical Findings

Factor Analysis

For each FSA (Grade 4 Numeracy, Grade 4 Reading, Grade 7 Numeracy, and Grade 7 Reading) there were varying degrees of similarities in the factor structures of the scores across the aboriginal and non-aboriginal populations. These similarities in factor structure were strong across the two populations for the Grades 4 and 7 Reading FSAs. For both Reading FSAs, an equal number of factors and a high degree of comparability across the two populations for the two (out of three) dominant factors were found. The similarities in factor structure were not as strong across the two populations for the Grades 4 and 7 Numeracy FSAs. For both Numeracy FSAs, an unequal number of factors

was found for the two populations making it impossible to mathematically estimate the degree of comparability. For both Numeracy FSAs, visual inspection of the composition of the factors did not reveal any common patterns across the two populations. There was limited overlap in the composition of factor solutions with respect to sub-content area, context, and number of words.

Reliability

For each assessment the reliability estimates were high (ranging from 0.78-0.88). Further, there were small differences found in the reliability estimates of the scores for the aboriginal and non-aboriginal populations. The Grade 4 and Grade 7 Numeracy FSAs had scores that were slightly more reliable for the non-aboriginal population, and the Grade 4 and Grade 7 Reading FSAs had scores that were slightly more reliable for the aboriginal population. In general, the reliability of the scores for each population was high, and very similar to one another.

Item Information Functions and Item Parameters

The item information functions indicate the degree of measurement accuracy provided by test items for different ability levels. The area under the item information function is an indicator of total measurement accuracy provided by each test item. For each FSA, the item information functions were summed to get the total test information for each population. Then, the differences between the test information for the two populations were used as an indication of the difference in the amount of information provided by the scores of the two populations. It was found that the Grade 4 Numeracy assessment and both reading assessments provided less information for the aboriginal group than the non-aboriginal group. In contrast, it was found that the Grade 7 Numeracy

assessment provided more information for the aboriginal group than the non-aboriginal group. Specifically, the Grade 4 Numeracy FSA provided 4% *less* information for the aboriginal population; the Grade 7 Numeracy FSA provided 19% *more* information for the aboriginal population; the Grade 4 Reading FSA provided 3% *less* information for the aboriginal population; and the Grade 7 Reading FSA provided 3% *less* information for the aboriginal population.

To further examine the differences in accuracy of the two populations for each FSA, the standard error of measurement (SEM) functions for each scale score point were estimated. The SEM functions provide a graphic display of measurement accuracy provided by the test for each scale score point. An examination of the SEM functions revealed that the Grade 4 Numeracy test and the Grades 4 and 7 Reading tests provided lower measurement accuracy for the non-aboriginal group at the higher end of the scale, and lower measurement accuracy for the aboriginal group at the lower end of the scale. For the Grade 7 Numeracy test, the above was not the case, but rather, except for the very low end of the ability scale, the test provided more measurement accuracy for the aboriginal group.

For each assessment examined, comparisons of IRT based parameters for the aboriginal and non-aboriginal populations showed that the items were ordered very similarly in terms of their difficulty level and their degree of discrimination, and moderately similar in their probability of responding correctly due to chance.

Differential Item Functioning

For the present study, DIF items were identified based on the consensus of two different DIF detection methods. For all four assessments the presence of DIF was

minimal. Overall, there were more DIF items in the two Reading FSAs than in either of the Numeracy FSAs. For the Grade 4 Numeracy FSA, approximately 10% of the items (4 items) were found to be differentially functioning, which were evenly distributed between the two populations in terms of the direction they favored. The proportion of bias in favour of the aboriginal population was 0.04; the proportion of bias in favour of the non-aboriginal population was 0.06. There was no pattern evident with respect to item details (i.e., number-of-words per item, item type, sub-content area, and context) that provided a plausible explanation of the sources of DIF.

For the Grade 7 Numeracy FSA, only one item was identified as DIF. This item favored the non-aboriginal population, and represents a proportion of bias in favour of this group of 0.02.

For the Grade 4 Reading FSA, approximately 21% of the items of were found to be DIF, with 13% of the items (five items) favoring the aboriginal population and 8% of the items (three items) favoring the non-aboriginal population. The proportion of bias in favour of the aboriginal population was 0.13; the proportion of bias in favour of the non-aboriginal population was 0.10. The item details for the DIF items identified for each population indicated that 6 of the DIF items of a total of 8 DIF items were from the *locate, interpret, and organize* sub-content area favoured the aboriginal population.

For the Grade 7 Reading FSA approximately 15% of the items of were found to be DIF, with 9% of the items (four items) favoring the aboriginal population and 6% (three items) of the items favoring the non-aboriginal population. The proportion of bias in favour of the aboriginal population was 0.13; the proportion of bias in favour of the non-aboriginal population was 0.09. Because of the similarity of the item details for the

DIF items identified for each population, neither number-of-words per item, item type, sub-content area, nor context provided a plausible explanation of the sources of DIF.

Research Question 1: Are Scores From the Foundation Skills Assessment Comparable Across Aboriginal and Non-Aboriginal Populations?

The BC Ministry of Education used the FSA scores to compare individuals from the aboriginal and non-aboriginal populations based on a self-declaration of being of aboriginal ancestry by each student. For these comparisons, differentiation with regard to membership in each group (aboriginal or non-aboriginal) can be confounded with such features as language, socio-economic status, environmental exposure, recreational skills, access to computers, teachers' skills and level of dedication, and availability of good text books. Once group distinction is made by the bodies governing the assessments, comparisons of scores amongst the groups can be problematic when all of the related factors such as those listed above cannot be disentangled.

The purpose of this study was to examine four FSAs to provide information about the degree of equivalence and comparability of scores for these two populations: aboriginal and non aboriginal. This project was based on Messick's (1989a; 1989b; 1995) unitary notion of validity in which one must make an overall evaluative judgment of the degree to which evidence supports the adequacy and appropriateness of score inferences. A number of pieces of evidence was gathered in order to make the determination of degree of comparability, each of which is presented below, as well as how each piece of evidence contributes to the evaluation of comparability.

One piece of evidence about the degree of comparability of these assessments for aboriginals and non-aboriginals was the factor analysis. This set of analyses examined

whether the dimensional structure of a test was found to be consistent for both populations. If the structure of the test was found to be the same for both populations then this evidence was deemed consistent with the hypothesis that the test is measuring the same construct for both populations. In this study, the dimensional structure of the test was found to be very similar for Grades 4 and 7 Reading FSAs, with high levels of comparability based on the congruence coefficients calculated for the factors across the two populations. The dimensional structure of the test was not found to be very similar for Grades 4 and 7 Numeracy FSAs. For both Numeracy assessments, the number of factors defining the dimensional structure of the test scores across the two populations was different. Further, the composition of the resulting dimensions for both Numeracy FSAs did not reveal common factors across the populations. This means that how the construct of numeracy is represented and measured is somehow different for the two populations. For example, there may be a difference in the levels of the skills required to solve the Numeracy assessments across the two populations.

Similarity of factor structures for the two populations is a necessary, but not sufficient, component of construct comparability. From this study, the factor structure evidence is supportive of a high degree of comparability across the aboriginal and non-aboriginal populations for the Reading FSA scores, but a low degree of comparability across the aboriginal and non-aboriginal populations for the Numeracy FSA scores.

Investigating the internal consistency of scores, as indicated by the reliability estimates, was another piece of evidence gathered to examine the construct comparability. For this study, for all four assessments, the reliability estimates of the test scores were high and similar for both populations, meaning that the degree to which

individuals' scores would remain relatively consistent over repeated administration of the same test or alternate test forms was high. The reliability analysis gave evidence to a high degree of comparability across the aboriginal and non-aboriginal populations.

Differences in the item information functions were examined for each FSA in an effort to detect differences in the degree of measurement accuracy provided by the test items for the two populations. For each of the Grade 4 Numeracy and the Grades 4 and 7 Reading FSAs, there was less than 4% difference in the level of accuracy of scores between the two populations; minimal by my subjective standard. For the Grade 7 Numeracy FSA, there was a substantial difference, 19%, in the level of accuracy of scores between the two populations, favoring the aboriginal population. The item information function analysis provided consistent evidence that there is a high degree of comparability across the aboriginal and non-aboriginal populations for Grade 4 Numeracy and Grades 4 and 7 Reading scores, and a low degree of comparability across the aboriginal and non-aboriginal populations for Grade 7 Numeracy scores.

The presence of DIF items for the two populations showed that, depending on group membership (aboriginal or non-aboriginal), a student's probability of answering particular items correctly when matched on ability could be different on some items. Because each FSA had fewer than 30% of the items identified as DIF, high degrees of DIF were not found in this study (Zenisky, Hambleton, & Robin 2003). In terms of the number of DIF items for each FSA, there were two items favoring each of the two populations for the Grade 4 Numeracy; there was only one item for the Grade 7 Numeracy and it favored the non-aboriginal population; there were five items favoring the aboriginal population and three items favoring the non-aboriginal population for

Grade 4 Reading; and there were four favoring the aboriginal population and three favoring the non-aboriginal population for Grade 7 Reading.

Further, for each FSA, an estimate was made of the amount of total bias related to the collective impact of DIF items for each population based on the average difference in the probability of obtaining the maximum item score for the two populations, matched on ability. For the Grade 4 Numeracy FSA, the proportions of bias were 4% and 6% in favor of the aboriginal and non-aboriginal populations, respectively. For the Grade 7 Numeracy FSA, the proportions of bias were 0% and 2% in favor of the aboriginal and non-aboriginal populations, respectively. For the Grade 4 Reading FSA, the proportions of bias were 13% and 10% in favor of the aboriginal and non-aboriginal populations, respectively. For the Grade 7 Reading FSA, the proportions of bias were 13% and 9% in favor of the aboriginal and non-aboriginal populations, respectively. For all four FSAs, the relatively low degree of DIF and the minimal difference across the populations in the proportions of bias related to the collective impact of all DIF items, lead me to conclude that the DIF analysis provided evidence in support of a high degree of comparability across the aboriginal and non-aboriginal populations' FSA scores. Overall, DIF analyses demonstrated the presence of DIF for the two populations, but the results did not show consistent bias against one group or the other.

This study resulted in many findings. Table 52 summarizes the determination of the degree of comparability across the aboriginal and non-aboriginal populations for each analysis. For ease of presentation and synthesis, results of analyses were considered to provide evidence for a low or high degree of comparability. From Table 52, it can be seen that for this study, there was a high degree of comparability across the two populations

for the Grades 4 and 7 Reading FSA scores because all four analyses showed them to be highly comparable. There was a moderately high degree of comparability across the two populations for the Grade 4 Numeracy FSA scores because three out of the four analyses showed them to be highly comparable scores. There was a moderate degree of comparability across the two populations for the Grade 7 Numeracy FSA scores because two out of the four analyses showed them to be highly comparable scores.

Table 52

Degree of Comparability Across the Aboriginal and Non-Aboriginal Populations

	Degree of Comparability	
Analysis	High	Low
Common Factor Analysis	Grade 4 Reading Grade 7 Reading	Grade 4 Numeracy Grade 7 Numeracy
Reliability	Grade 4 Numeracy Grade 7 Numeracy Grade 4 Reading Grade 7 Reading	
Item Information Functions	Grade 4 Numeracy Grade 4 Reading Grade 7 Reading	Grade 7 Numeracy
Differential Item Functioning	Grade 4 Numeracy Grade 7 Numeracy Grade 4 Reading Grade 7 Reading	

The differences of the FSA scores across the populations as indicated by the findings above, although minimal in nature, are likely because the items are assessing additional skills/competencies and the distribution of these additional skills/competencies is different for the two populations.

Research Question 2: Should Score Interpretations be the Same for Both Populations?

The phrasing of this research question could imply that there may be separate interpretations for the aboriginal and non-aboriginal FSA scores, but what it really asks is if the scores were adequately attained in preparing an unbiased common scale for the two populations. To re-phrase what was stated above, this study found that the scores were very comparable across the aboriginal and non-aboriginal populations for Grades 4 and 7 Reading FSAs; quite comparable for Grade 4 Numeracy FSAs; and moderately comparable for Grade 7 Numeracy FSAs.

In answering Research Question 2, this study found that: (a) interpretations should be the same for the aboriginal and non-aboriginal populations for the Grades 4 and 7 Reading FSAs scores; (b) interpretations should be made with care and caution when comparing the two populations for the Grade 4 Numeracy FSA scores; and (c) interpretations that imply score comparability across the two populations should not be made for the Grade 7 Numeracy FSA. Making the same interpretations for these two populations based on scores from the Grade 7 Numeracy FSA could lead to faulty conclusions such as comparing the aboriginal and non-aboriginal students based on their performance and reporting the differences.

Findings in Context

As previously stated in the literature review, the *Standards* document (AERA et al., 1999) contends that when different populations have scores that have different distributions, this is evidence of invalidity only if different score distributions "were due to the test's sensitivity to some examinee characteristic not intended to be part of the test

construct" (p.16). One statistical method used to distinguish between whether there are true differences in the abilities of the two groups, or whether or not the differences represent construct irrelevant bias is differential item functioning (DIF). For each assessment there were relatively few DIF items based on the context of the item, and this provides evidence that the group differences were minimally influenced by characteristics that were not intended to be part of what was intended to be measured as laid out by the BC Ministry of Education. As was stated in Chapter 1 of this study, the main purpose of the FSAs was to help the province, school districts, schools, and school planning councils evaluate how well foundation skills are being addressed and make plans to improve student achievement (British Columbia Ministry of Education, 2002a).

Some literature that dealt with the disparity of scores and the difference in the level of academic success between the aboriginal and non-aboriginal populations. Some studies presented a set of hypotheses that consistently blamed the drastic differences in cultural (including language) and socio-economic situations between the groups as the explanation (Wesley, 1961; Lloyd, 1961). Others presented explanations that indicated that it was the lack of aboriginal culture and aboriginal-like perspective taking in the assessments that resulted in the disparity of scores and the difference in the level of academic success between the aboriginal and non-aboriginal populations (Keeler, 1961; Evvard & Weaver, 1966). More recent research has shown that the following efforts have led to a reduction in the disparity of scores between the aboriginal and non-aboriginal populations: (a) a nurturing early childhood environment (Swisher & Deyhle, 1989) (b) inclusion of native language and cultural programs in the school (Ayoungman, 1991; Barnhardt, 1990, 1999; deMarrais, 1992; James, Chavez, Beauvais, Edwards, & Oetting,

1995; Lipka & McCarty, 1994; Rubie, 1999; Slaughter & Lai, 1994; McLaughlin, 1992; Watahomigie & McCarty, 1994), and (c) enhanced community and parental influences on academic performance (Leveque, 1994; and McInerney, McInerny, Ardington, & De Rachewiltz, 1997). It is my belief that these researchers presented plausible explanations for the differences, and suggestions for improvement, but unfortunately, I did not have the data, or resources, to explore their hypotheses in full. A natural extension of this study would be to further explore these plausible explanations by teaming up with other researchers, such as those listed above, as well attaining relevant data that could inform these issues in British Columbia from Statistics Canada, the BC Ministry of Education, the BC Ministry of Children and Family Development, the BC Ministry of Community, Aboriginal, and Women's Services, and the BC Ministry of Health Services.

Implications of Findings

Interpretation Implications

For the Grades 4 and 7 Reading FSAs, the high degree of comparability of scores found in this study led to the conclusion that these tests are measuring the same construct with nearly the same degree of accuracy for the aboriginal and non-aboriginal populations. The implication of how these scores should be interpreted is simple: the scores from the two populations can be interpreted in the same way and can be compared to one another in a psychometrically sound manner.

For the Grade 4 Numeracy FSA, the moderately high degree of comparability of scores found in this study led to the conclusion that this test did not measure exactly the same construct, but what was measured was done with nearly the same degree of accuracy for the aboriginal and non-aboriginal populations. The implication of how these

scores should be interpreted is that the scores from the two populations should not be interpreted as if they were comparable. The reliability analysis, the item information functioning analysis, and the differential item functioning analysis all indicated that there was a high degree of comparability. However, the differences in the factor structures led me to conclude that further investigation into how this construct is represented and measured for these populations is necessary in order to understand, address and possibly reduce the differences across the populations.

For the Grade 7 Numeracy FSA, the moderate degree of comparability of scores found in this study led to the conclusion that this test did not measure exactly the same construct, and what was measured was done with similar degrees of accuracy for the aboriginal and non-aboriginal populations. The implication of how these scores should be interpreted is that the scores from the two populations should not be interpreted as if they were comparable. The fact that the reliability analysis and the differential item functioning analysis both indicated that there was a high degree of comparability led me to believe that further investigation into the factor structure and the item information functions may lead to an understanding of the differences across the populations, which could in turn be addressed and possibly reduced.

In the case where substantial psychometric differences were found to exist in the FSA scores across the aboriginal and non-aboriginal populations, these differences would have had an impact on the FSA total scores. This impact on total scores would have continued its influence as the BC Ministry of Education classified each student into cut-score-based proficiency levels (not yet within expectations, meets expectations, exceeds expectations). The BC Ministry of Education then used individual proficiency levels to

compare aboriginals and non-aboriginals at the provincial level. In conclusion, such comparisons would have misrepresented the true differences between academic performance for these two populations.

Of the psychometric differences found, the differences in the degree of measurement accuracy of scores for the two populations is of serious concern. It was shown that for the Grade 4 Numeracy and the Grades 4 and 7 Reading FSAs, the scores were more accurate for the non-aboriginal population than the aboriginal population at the lower end of the ability scale, and the opposite was true at the higher end of the ability scale. This means that there is less measurement accuracy for the aboriginal students with lower ability levels than the non-aboriginal students with lower abilities. Further, there are more aboriginal students than non-aboriginal students with lower ability levels. When the BC Ministry of Education reduced the total scores into proficiency levels, one would hope that there was an equally high degree of measurement accuracy for each population at the cut-scores that make this determination, especially for the cut-score determining the lowest proficiency level. Although I did not have the cut-scores to determine if this was the case, the difference in measurement accuracy at the lower end of the ability scale is an indication that there was probably different degrees of measurement accuracy for the two populations in determining which students were classed into the lowest proficiency level.

With educational decisions being made based on students' FSA scores, it is the responsibility of the BC Ministry of Education to demonstrate the extent to which inferences based on these scores are valid and comparable for all identifiable sub-groups. The implication of the findings of this study is that scores from the Grades 4 and 7

Reading FSAs are comparable for the aboriginal and non-aboriginal populations, meaning that the inferences based on these scores are valid and comparable. Further, the scores from the Grades 4 and 7 Numeracy FSAs are not strictly comparable, meaning that the inferences based on these scores are not valid and comparable. In the case of the Grades 4 and 7 Numeracy FSA scores, there appear to be factors that are inherent in the membership in the aboriginal population that influence the academic assessment scores. When the influence of group membership on assessment scores is strong enough to produce unequivalent factor structures, differing item information functions, or DIF items; some type of change needs to be made to either the testing instrument or the interpretation of the scores if the score interpretations are to be valid and comparable. Some suggestions for changes are discussed in the *Methodological Implications* section below.

Methodological Implications

This study represents the first stage of a multi-staged process in that it *identified* problems with FSA score comparability across the aboriginal and non-aboriginal populations. This study did not find high levels of DIF for the FSAs examined. Regardless, it is important to understand the methodological implications of DIF items because they point to differences between the performance of students from the two populations on these items, sources of which we may not know yet.

For the Grade 4 Reading FSA, the sub-content area and context of items was found to be an influencing factor on the finding of DIF; for the Grade 7 Reading FSA, context of items was found to be an influencing factor on the finding of DIF. For the Grades 4 and 7 FSAs, the sources of DIF were more complex than the identifiable

variables listed above could explain. Other possible sources of differences for the two populations could be the familiarity with the FSA testing techniques, item formats, test conventions, and testing procedures (Hambleton & Jong, 2003). Further sources of DIF could be related to differences in the two populations with regard to individuals' interpretations of the relevance of the items, their intrinsic interest in the item, and their familiarity with the item content (Ercikan, 1998). To investigate the complicated nature of understanding the sources of DIF, techniques for judgemental review of DIF items and procedures for examining student cognitive processes have been developed (Ercikan et al., 2002). The process of a judgemental review is to discover why items are performing differently between groups. The judgemental review uses a panel of reviewers to examine items for inconsistencies in meaning across groups or cultures. Both of these methods provide insights about sources of DIF and why test items are functioning differentially for students from different socio-cultural backgrounds.

If assessment scores are found not to be comparable across populations, the adaptation or revision of items, via the judgemental reviewers' summary critique, are two legitimate solutions. They are both relatively inexpensive and faster alternatives to preparing a completely new test for a second cultural or language group (Hambleton & Patsula, 1998). One benefit of adapting a test across identifiable sub-populations is that it can enhance the fairness by enabling persons to take tests in their preferred manner of format, context, or language. But with the development of different adapted versions for identifiable sub-populations, come new challenges in terms of verifying that each version, administered to the appropriate population, produces scores that are valid and comparable across all versions. The other alternative suggested above in examining

sources of DIF is to revise the items in such a way that the biased nature of the item is removed. The benefit of altering the DIF items is that individuals from the identifiable cultural groups will all be given the same items; hence, comparable scores would result.

The real focus of the implication of this study should be on the necessity of the undertaking of construct validity investigations for all large-scale assessments when comparisons are to be made about scores from identifiable sub-groups. Test developers bear the responsibility of demonstrating that their tests produce scores that have valid and comparable interpretations for all persons who they were designed to assess. It should be a routine part of their development procedures; without it, no appropriate inferences can be made.

Limitation of Findings

There were certain limitations of my study that I would like to discuss; I will attempt to convey how these limitations may or may not have affected my findings. One limitation was related to the manner in which the students were categorized as either aboriginal or non-aboriginal. This categorization was based on the self-declaration of aboriginal heritage by the student; there were no other criteria applied for this declaration. The self-declaration process allowed for students who were of aboriginal heritage to choose *not* to declare themselves as aboriginal, and for students who were not of aboriginal heritage to choose to declare that they were. Considering this self-declaration variable for categorizing students into the two populations (aboriginal and non-aboriginal) for comparison, the composition of the two populations might have been different if another criteria, such as Indian Status, had been used to define the aboriginal population. But then I have to wonder, would only including Status Indian students as

aboriginal cause error in the comparison by having the Métis and other non-status aboriginal groups in the non-aboriginal population. It would be interesting to follow the self-declaration pattern of students who where actually Status Indians to see if this self-declaration variable was static or dynamic. If this variable was dynamic, and students changed their mind from year to year about deciding to declare themselves as having aboriginal ancestry, the comparison of the scores across the aboriginal and non-aboriginal populations would become unstable and impossible to understand or track over the years.

Another limitation of this study is the lack of demographic variables for the students. As mentioned above, I would have liked to have other culturally embedded variables such as socio-economic status to examine the effects of it on score performance. Another limitation of the study was that the data were received with *no* identifiable missing data. Before the data were delivered to me, all non-response data (the missing data) were coded as zeros. In an effort to clean the data, I removed the cases in which every response was zero, as these cases most likely did not reflect real people. But doing this, I may have deleted some cases in which a student answered all the questions incorrectly.

Future Directions

From this study came the realization that there are directions that interested researchers could pursue in an effort to fully understand the nature of the differences in large-scale assessment scores for the aboriginal and non-aboriginal student populations. Further, with regard to the application of the construct validation aspect of the study, there are research directions that could lead to greater specificity regarding how much

and what kind of differences between identifiable populations on assessments justify changes to the assessment instrument or different interpretations.

One research direction would be to include the examination of convergent and discriminant correlations with external variables in determining the comparability of the score inferences for the aboriginal and non-aboriginal populations. This would broaden the scope of the investigation, but would have to be done cautiously as possible measurement error and biases may exist in the external data.

Another research direction would be to interview groups of students from both aboriginal and non-aboriginal groups with regard to examining student cognitive processes as well as their beliefs about the appropriateness of the contexts of the items that make up the assessment in question. This type of research could inform the test designers as to sources of DIF and construct incomparability.

Another research direction would be to apply a two-stage DIF approach in an effort to reduce the contamination of the matching criterion across the populations. In typical DIF analysis, the matching criterion is, as was in this study, the individual's total score. This means that the total score was used as the proxy for ability, when students' scores were matched on ability. When there is a high level of DIF items found, the associated errors related to the DIF items are embedded in the total score that is used to match the scores; this leads to the circularity of the error (Zenisky, Hambleton, & Robin, 2003). This type of contamination "is likely to result in less than optimal identification of DIF items and complicate efforts to interpret the findings" (Zenisky, Hambleton, & Robin, p. 52). The two-stage approach refines the total score by removing items that were initially identified as DIF. Once examinees are matched on the revised criterion, the DIF

analysis is repeated to identify newly emerged DIF items. This process may be repeated until some a priori condition is reached (Zenisky, Hambleton, & Robin). With the use of this thorough DIF detection method, one could be assured that they did in fact identify all DIF items.

Finally, and of most interest to me, another research direction is regarding how much and what kind of psychometric-property differences between identifiable populations on assessments justify the categorization of *comparable* and *not-comparable* when performing a construct comparability study. This aspect provided me with one of the biggest hurdles in terms of interpreting the results of this study. Sireci, Xing, Bastari, Allalough, and Fitzgerald (1999) spoke to the lack of appropriate statistical tests in determining the degree of structural equivalence, I found the same void when I tried to determine the degree of DIF, the degree of internal consistency differences, and the degree of item information function differences. With the attention being drawn to the need for construct validation and comparability studies with assessment scores, comes the need for a set of guidelines by which researchers can make claims about their findings in a manner that is acceptable by their peers, and ultimately leads to a consistency in the improvement of assessment.

REFERENCES

Alexander, K. L., Entwisle, D. R., & Kabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record, 103,* 760-823.

Allalouf, A., Hambleton, R., K., & Sireci, S. (1999). Identifying causes of DIF in adapted Verbal items. *Journal of Educational Measurement, 36,* 185-198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association. *Joint committee on testing practices.* Retrieved December 4, 2002, from http://www.apa.org/science/jctpweb.html

Ayoungman, V. (1991). Siksika language renewal efforts: Description and assessment (Indians, Canada) (Doctoral dissertation, Arizona State University, 1991). *Dissertations Abstract International, 52,* 1188.

Barnhardt, C. (1999). *Kuinerrarmiut Elitnaurviat: The school of the people of Quinhagak, a case study.* Portland, OR: Northwest Regional Educational Lab. (ERIC Document Reproduction Service No. ED437252)

Barnhardt, R. (1990). Two cultures, one school: St. Mary's, Alaska. *Canadian Journal of Native Education, 17,* 54-65.

British Columbia Ministry of Education. *Interpreting and communicating British Columbia Foundation Skills Assessment results.* Retrieved November 10, 2002a, from http://www.bced.gov.bc.ca/assessment/fsa/fsa_interpretation_2003.pdf

British Columbia Ministry of Education. *Reports and Publications*. Retrieved November 10, 2002b, from http://www.bced.gov.bc.ca/reportfinder/publicschools.php

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Burket, G. R. (1993). FLUX [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Burket, G. R. (1998). PARDUX (Version 4.01) [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Cattell, R. B. (1966). The scree test for the number of factors. *The Journal of Multivariate Behavioral Research, 1,* 245-276.

Cattell, R. B. & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *The Journal of Multivariate Behavioral Research, 13,* 289-325.

Committee on Foundations of Assessments. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Council of Ministers of Education, Canada. *Education indicators in Canada.* Retrieved November 15, 2002, from http://www.cmec.ca/stats/pceip/1999old/pceipmac/english/pages/page26e.html

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

deMarrais, K. B. (1992). Meaning in mud: Yup'ik Eskimo girls at play. *Anthropology & Education Quarterly, 23,* 120-144.

Demmert, W. G. (2001). *Improving academic performance among native American students: A review of the research literature*. Retrieved November 12, 2002, from http://www.edrs.com/Webstore/Download.cfm?ID=692162

Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29*, 543-553.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2(3&4),* 199-215.

Ercikan, K. (2003). Are the English and French Versions of the Third International Mathematics and Science Study Administered in Canada Comparable? Effects of Adaptations. *International Journal of Educational Policy, Research and Practice.*

Ercikan, K., Gierl, M. McCreith, T., Puham, G. & Koh, K. (in press). Comparability of English and French versions of SAIP for reading, mathematics and science items. *Applied Measurement in Education.*

Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items. In D. Robitaille & A. Beaton (Eds.), *Secondary Analysis of TIMSS Results* (pp. 391-405). Dordrecht The Netherlands: Kluwer Academic Publisher.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with MC and constructed-response item types. *Journal of Educational Measurement, 35*, 137-154.

Evvard, E., & Weaver, Jr., R. R. (1966). Testing—Some implications of counsellors and teachers. *Journal of American Indian Education, 5(3)*. Retrieved October 1, 2002, from the Journal of American Indian Education Electronic Journal Collection (via the Xwi7xwa Library, UBC).

Gierl, M, J., & Khaliq, S., N. (2001). Identifying sources of differential item functioning on adapted achievement tests: A confirmatory approach. *Journal of Educational Measurement, 38*, 164-187.

Gorsuch, R. L. (1983). *Factor Analysis.* Hillsdale, NJ: Lawrence Earlbaum Associates, Publishers.

Gould, S. J. (1995). Mismeasure by any measure. In R. Jacoby & N. Glauberman (Eds.), The bell curve debate (pp. 3–13). New York: Random House/Times Books.

Hambleton, R. K. (1989). Principles and selected applications of Item Response Theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). Washington, DC: American Council on Education.

Hambleton, R. K., & De Jong, J. H. A. L. (2003). Advances in translating and adapting educational and psychological tests [Electronic version]. *Language Testing, 20*, 127-134.

Hambleton, R. K., & Jones, R. W. (1994). Comparison of empirical and judgemental procedure for detecting differential item functioning. *Educational Research Quarterly, 18*, 21-36.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures [Electronic Version]. Social Indicators Research, 45, 153-171.

Hambleton, R. K., Swaminatham, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publications.

Hakstian, R. A. (1971). A comparative evaluation of several prominent methods of oblique factor transformation. *Psychometrika, 36,* 175-193.

Hakstian, R. A., Farrell, S., & Tweed, R. G. (2002). The assessment of counterproductive

tendencies by means of the California Psychological Inventory. *International Journal of Selection & Assessment, 10,* 58-86.

Hakstian, R. A., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rule with simulated data. *Multivariate Behavioral Research, 17,* 193-219.

Harman, H. H. (1976). *Modern factor analysis*. Chicago: The University of Chicago Press.

Harpur, T. J., Hakstian, A. R., & Hare, R. D. (1988). Factor structure of the psychopathy checklist. *Journal of Consulting and Clinical Psychology, 56,* 741-747.

Indian and Northern Affairs Canada. *United Nations working group on indigenous populations: Statement by the observer delegation of Canada on the theme of "Children and Youth"*. Retrieved November 6, 2002, from http://www.ainc-inac.gc.ca/nr/spch/2000/chay_e.html

James, K., Chavez, E., Beauvais, F., Edwards, R., & Oetting, G. (1995). School achievement and dropout among Anglo and Indian females and males: A comparative examination. *American Indian Culture and Research Journal, 19,* 181-206.

Keeler, W. W. (1961). Challenges in Indian education. *Journal of American Indian Education,1(2).* Retrieved October 01, 2002, from the Journal of American Indian Education Electronic Journal Collection (via the Xwi7xwa Library, UBC).

Kirkness, V. J. (1999). Aboriginal education in Canada: A retrospective and a prospective. *Journal of American Indian Education, 39,* 14-30.

Leveque, D. M. (1994). *Cultural and parental influences on achievement among*

Native American students in Barstow unified school district. Paper presented at the National Meeting of the Comparative and International Educational Society, San Diego, CA.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group measurement on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

Lipka, J., & McCarty, T. L. (1994). Changing the culture of schooling: Navajo and Yup'ik cases. *Anthropology & Education Quarterly, 25,* 266-84.

Lloyd, D. O. (1961). Comparison of standardized test results of Indian and non-Indian in an integrated school system. *Journal of American Indian Education,1.* Retrieved October 03, 2002, from the Journal of American Indian Education Electronic Journal Collection (via the Xwi7xwa Library, UBC).

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

McInerney, D. M., McInerney, V., Ardington, A., & De Rachewiltz, C. (March 1997). *School success in cultural context: Conversations at Window Rock. Preliminary Report.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

McLaughlin, D. (1992). *When literacy empowers: Navajo language in print.* Albuquerque: University of New Mexico Press.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., p.p. 13-103). Washington, DC: American Council on Education.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741-749.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algoritym. *Applied Psychological Measurement, 16,* 159-176.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8,* 111-120.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Rubie, C. (1999). *Kia Kaha: Improving classroom performance through developing cultural awareness.* Paper presented at the Joint Conference of the Australian Association for Research in Education and the New Zealand Association for Research in Education, Melbourne, Australia. (ERIC Document Reproduction Service No. ED441651)

Rummel, R. J. (1970). *Applied factor analysis*. Evanston: Northwest University Press.

Sireci, S. G., Xing, D., Bastari, B., Allalouf, A., & Fitzgerald, C. (1999). Evaluating construct equivalence across tests adapted for use across multiple languages. Unpublished manuscript, University of Massachusetts at Amherst.

Slaughter, H. B., & Lai, M. (1994). *Indigenous language immersion as an alternative form of schooling for children of Hawaiian ancestry: Lessons from a six-year study.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service

No. ED375637)

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

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

Swisher, K., & Deyhle, D. (1989). The styles of learning are different, but the teaching is just the same: Suggestions for teachers of American Indian youth. *Journal of American Indian Education, Special Issue,* 1-14.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. Northridge, CA: Harper Collins College Publishers.

The Fraser Institute. *Report Card on British Columbia's Elementary Schools: 2003 Edition.* Retrieved October 15, 2003, from http://www.fraserinstitute.ca/shared/readmore.asp?sNav=pb&id=536

Watahomigie, L. J., & McCarty, T. L. (1994). Bilingual/bicultural education at Peach Springs: A Hualapai way of schooling. *Peabody Journal of Education, 69*(2), 26-42.

Wesley, C. (1961). Indian education. *Journal of American Indian Education,1(1)*. Retrieved October 01, 2002, from the Journal of American Indian Education Electronic Journal Collection (via the Xwi7xwa Library, UBC).

Witherspoon, Y. T. (1962). The measurement of Indian children's achievement in the academic tool subjects. *Journal of American Indian Education,1.* Retrieved October 01, 2002, from the Journal of American Indian Education Electronic Journal Collection (via the Xwi7xwa Library, UBC).

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-214.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach [Electronic version]. *Educational and Psychological Measurement, 63*, 51-64.

APPENDIX A

Grade 4 Numeracy Item Details

Item Number	Number Of Words	Context	Item Type	Sub-content area
1	21	Field Trip To Nature Park	MC	Number
2	26	Field Trip To Nature Park	MC	Patterns & Relationships
3	31	Field Trip To Nature Park	MC	Number
4	32	Field Trip To Nature Park	MC	Shape & Space
5	11	Field Trip To Nature Park	MC	Shape & Space
6	7	Field Trip To Nature Park	MC	Statistics & Probability
7	11	Field Trip To Nature Park	MC	Number
8	41	Field Trip To Nature Park	MC	Number
9	25	Field Trip To Nature Park	MC	Shape & Space
10	26	Field Trip To Nature Park	MC	Shape & Space
11	25	Field Trip To Nature Park	MC	Number
12	25	Field Trip To Nature Park	MC	Number
13	9	Field Trip To Nature Park	MC	Statistics & Probability
14	37	Field Trip To Nature Park	MC	Patterns & Relationships
15	20	Field Trip To Nature Park	MC	Patterns & Relationships
16	26	Field Trip To Nature Park	MC	Shape & Space
17	35	Field Trip To Nature Park	OE	Patterns & Relationships
18	40	Field Trip To Nature Park	OE	Statistics & Probability
19	17	Activity Day For Students & Parents	MC	Number
20	29	Activity Day For Students & Parents	MC	Patterns & Relationships
21	18	Activity Day For Students & Parents	MC	Shape & Space
22	25	Activity Day For Students & Parents	MC	Patterns & Relationships
23	19	Activity Day For Students & Parents	MC	Patterns & Relationships
24	26	Activity Day For Students & Parents	MC	Shape & Space
25	18	Activity Day For Students & Parents	MC	Number
26	19	Activity Day For Students & Parents	MC	Number
27	19	Activity Day For Students & Parents	MC	Number
28	26	Activity Day For Students & Parents	MC	Number
29	23	Activity Day For Students & Parents	MC	Number
30	25	Activity Day For Students & Parents	MC	Statistics & Probability
31	19	Activity Day For Students & Parents	MC	Statistics & Probability
32	20	Activity Day For Students & Parents	MC	Statistics & Probability
33	16	Activity Day For Students & Parents	MC	Shape & Space
34	26	Activity Day For Students & Parents	MC	Number
35	15	Activity Day For Students & Parents	OE	Number
36	30	Activity Day For Students & Parents	OE	Number

APPENDIX B

Grade 7 Numeracy Item Details

Item #	# Of Words	Context	Item Type	Sub-content area
1	36	Ski Trip	MC	Number
2	34	Ski Trip	MC	Statistics & Probability
3	37	Ski Trip	MC	Number
4	25	Ski Trip	MC	Number
5	32	Ski Trip	MC	Number
6	19	Ski Trip	MC	Number
7	37	Ski Trip	MC	Number
8	28	Ski Trip	MC	Patterns & Relations
9	25	Ski Trip	MC	Statistics & Probability
10	31	Ski Trip	MC	Patterns & Relations
11	20	Ski Trip	MC	Shape & Space
12	15	Ski Trip	MC	Shape & Space
13	17	Ski Trip	MC	Shape & Space
14	16	Ski Trip	MC	Shape & Space
15	44	Ski Trip	MC	Statistics & Probability
16	33	Ski Trip	MC	Number
17	49	Ski Trip	OE	Number
18	40	Ski Trip	OE	Patterns & Relations
19	27	School Fun Fair	MC	Number
20	35	School Fun Fair	MC	Shape & Space
21	27	School Fun Fair	MC	Statistics & Probability
22	35	School Fun Fair	MC	Number
23	17	School Fun Fair	MC	Number
24	26	School Fun Fair	MC	Patterns & Relations
25	32	School Fun Fair	MC	Number
26	19	School Fun Fair	MC	Patterns & Relations
27	30	School Fun Fair	MC	Patterns & Relations
28	20	School Fun Fair	MC	Shape & Space
29	19	School Fun Fair	MC	Number
30	34	School Fun Fair	MC	Number
31	23	School Fun Fair	MC	Shape & Space
32	25	School Fun Fair	MC	Shape & Space
33	28	School Fun Fair	MC	Statistics & Probability
34	46	School Fun Fair	MC	Number
35	50	School Fun Fair	OE	Shape & Space
36	20	School Fun Fair	OE	Patterns & Relations

APPENDIX C

Grade 4 Reading Item Details

Item #	# Of Words	Context	Item Type	Sub-content area
1	6	Crime Solving	MC	Identify & Interpret Key Concept & Main Idea
2	7	Crime Solving	MC	Locate, Interpret, & Organise Details
3	3	Crime Solving	MC	Locate, Interpret, & Organise Details
4	27	Crime Solving	MC	Locate, Interpret, & Organise Details
5	8	Rain	MC	Identify & Interpret Key Concept & Main Idea
6	6	Rain	MC	Locate, Interpret, & Organise Details
7	7	Rain	MC	Locate, Interpret, & Organise Details
8	12	Rain	MC	Locate, Interpret, & Organise Details
9	33	Rain	OE	Locate, Interpret, & Organise Details
10	9	House Pets	MC	Locate, Interpret, & Organise Details
11	8	House Pets	MC	Locate, Interpret, & Organise Details
12	4	House Pets	MC	Locate, Interpret, & Organise Details
13	9	House Pets	MC	Locate, Interpret, & Organise Details
14	7	House Pets	MC	Critical Analysis
15	8	Polar Bears	MC	Identify & Interpret Key Concept & Main Idea
16	8	Polar Bears	MC	Locate, Interpret, & Organise Details
17	11	Polar Bears	MC	Locate, Interpret, & Organise Details
18	16	Polar Bears	MC	Locate, Interpret, & Organise Details
19	23	Polar Bears	OE	Locate, Interpret, & Organise Details
20	5	Frogs & Toads	MC	Locate, Interpret, & Organise Details
21	7	Frogs & Toads	MC	Locate, Interpret, & Organise Details
22	9	Frogs & Toads	MC	Locate, Interpret, & Organise Details
23	12	Frogs & Toads	MC	Locate, Interpret, & Organise Details
24	10	Frogs & Toads	MC	Critical Analysis
25	6	Rabbits	MC	Locate, Interpret, & Organise Details
26	19	Rabbits	MC	Locate, Interpret, & Organise Details
27	11	Rabbits	MC	Locate, Interpret, & Organise Details
28	12	Rabbits	MC	Critical Analysis
29	19	Rabbits	OE	Critical Analysis
30	6	Memory	MC	Identify & Interpret Key Concept & Main Idea
31	8	Memory	MC	Critical Analysis
32	6	Memory	MC	Locate, Interpret, & Organise Details
33	12	Memory	MC	Locate, Interpret, & Organise Details
34	8	Memory	MC	Locate, Interpret, & Organise Details
35	8	Tree Growth	MC	Identify & Interpret Key Concept & Main Idea
36	10	Tree Growth	MC	Locate, Interpret, & Organise Details
37	9	Tree Growth	MC	Locate, Interpret, & Organise Details
38	11	Tree Growth	MC	Critical Analysis
39	26	Tree Growth	OE	Locate, Interpret, & Organise Details

APPENDIX D

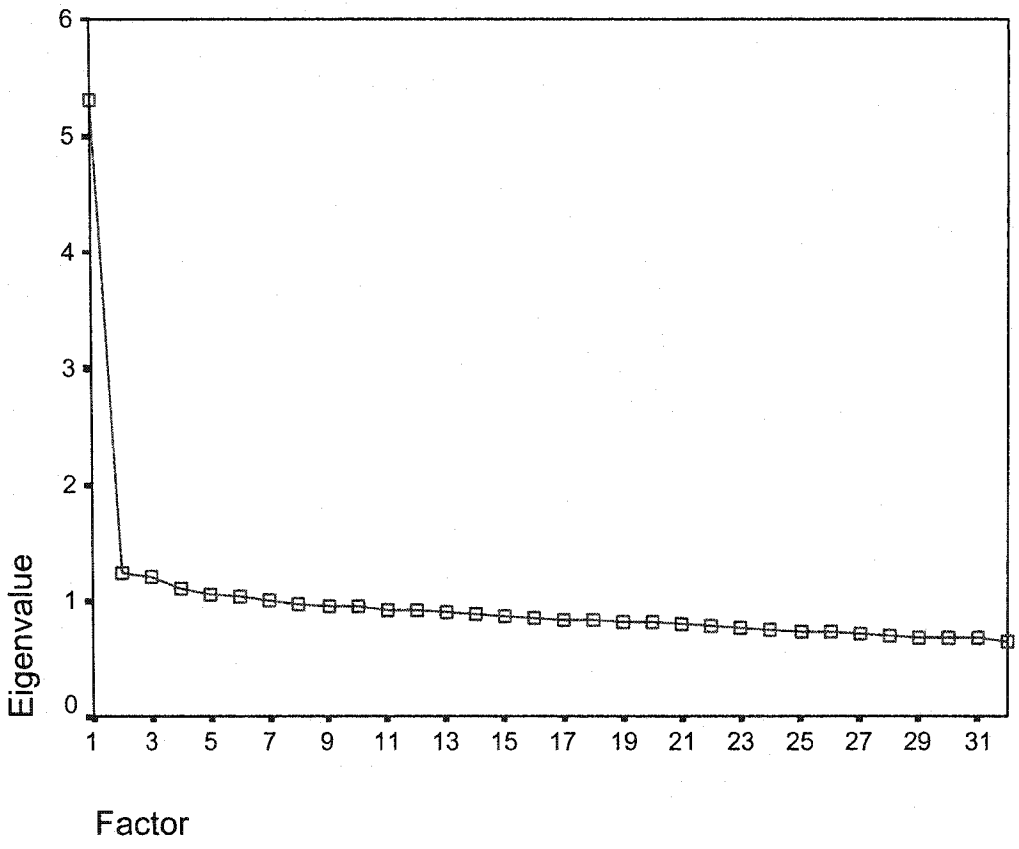
Grade 7 Reading Item Details

Item #	#Of Words	Context	Item Type	Sub-content area
1	15	Baseball Game	MC	Critical Analysis
2	13	Baseball Game	MC	Locate, Interpret, & Organise Details
3	21	Baseball Game	MC	Locate, Interpret, & Organise Details
4	10	Baseball Game	MC	Critical Analysis
5	10	Baseball Game	MC	Locate, Interpret, & Organise Details
6	38	Baseball Game	OE	Identify & Interpret Key Concept & Main Idea
7	10	Whales	MC	Locate, Interpret, & Organise Details
8	13	Whales	MC	Critical Analysis
9	15	Whales	MC	Locate, Interpret, & Organise Details
10	10	Whales	MC	Locate, Interpret, & Organise Details
11	8	Ponies	MC	Identify & Interpret Key Concept & Main Idea
12	12	Ponies	MC	Critical Analysis
13	15	Ponies	MC	Locate, Interpret, & Organise Details
14	14	Ponies	MC	Critical Analysis
15	9	Ponies	MC	Locate, Interpret, & Organise Details
16	26	Ponies	OE	Locate, Interpret, & Organise Details
17	5	Frogs & Toads	MC	Locate, Interpret, & Organise Details
18	7	Frogs & Toads	MC	Locate, Interpret, & Organise Details
19	9	Frogs & Toads	MC	Locate, Interpret, & Organise Details
20	12	Frogs & Toads	MC	Locate, Interpret, & Organise Details
21	10	Frogs & Toads	MC	Critical Analysis
22	8	Goldfish	MC	Locate, Interpret, & Organise Details
23	8	Goldfish	MC	Locate, Interpret, & Organise Details
24	11	Goldfish	MC	Critical Analysis
25	10	Goldfish	MC	Critical Analysis
26	14	Goldfish	MC	Critical Analysis
27	9	Goldfish	MC	Critical Analysis
28	9	Weeping Willow Tree	MC	Locate, Interpret, & Organise Details
29	11	Weeping Willow Tree	MC	Locate, Interpret, & Organise Details
30	17	Weeping Willow Tree	MC	Locate, Interpret, & Organise Details
31	13	Weeping Willow Tree	MC	Locate, Interpret, & Organise Details
32	17	Weeping Willow Tree	MC	Locate, Interpret, & Organise Details
33	12	Weeping Willow Tree	MC	Critical Analysis
34	36	Weeping Willow Tree	OE	Locate, Interpret, & Organise Details
35	5	Snakes	MC	Identify & Interpret Key Concept & Main Idea
36	8	Snakes	MC	Locate, Interpret, & Organise Details
37	5	Snakes	MC	Critical Analysis
38	9	Snakes	MC	Locate, Interpret, & Organise Details
39	6	Snakes	MC	Locate, Interpret, & Organise Details
40	6	Snakes	MC	Locate, Interpret, & Organise Details
41	9	Egypt	MC	Locate, Interpret, & Organise Details
42	16	Egypt	MC	Critical Analysis

43	13	Egypt	MC	Locate, Interpret, & Organise Details
44	6	Egypt	MC	Locate, Interpret, & Organise Details
45	6	Egypt	MC	Locate, Interpret, & Organise Details
46	27	Egypt	OE	Locate, Interpret, & Organise Details

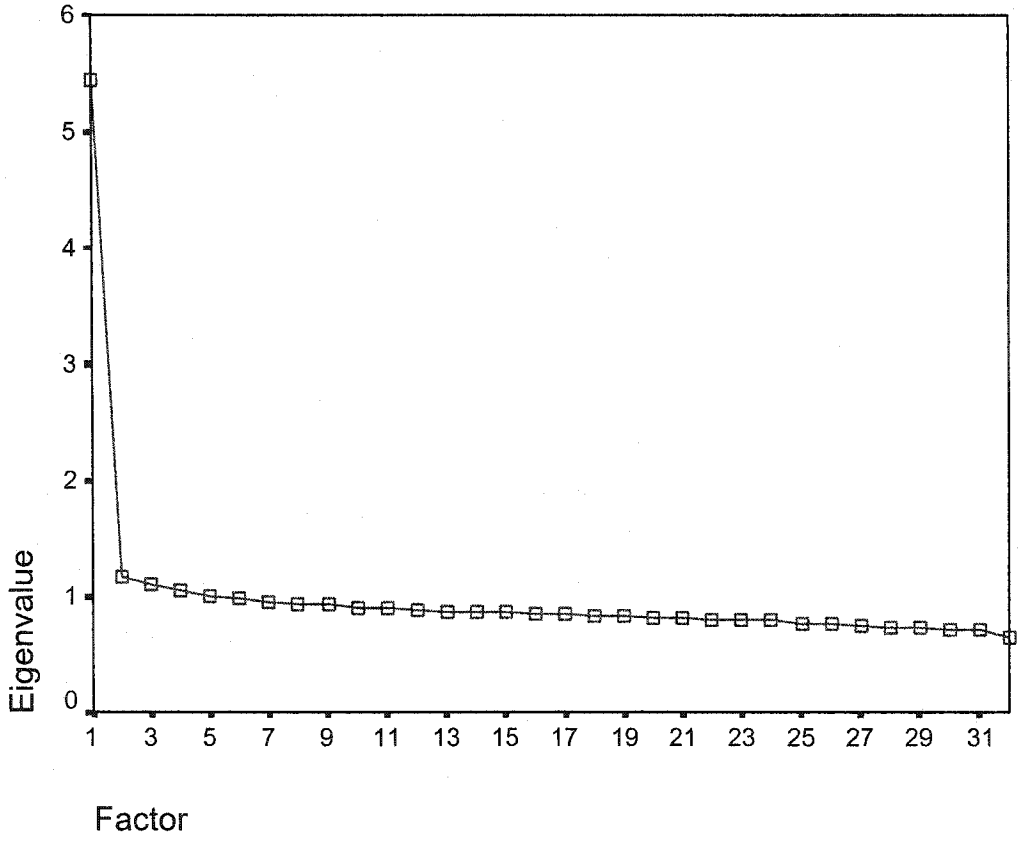
APPENDIX E

Aboriginal Grade 4 Numeracy Scree Plot



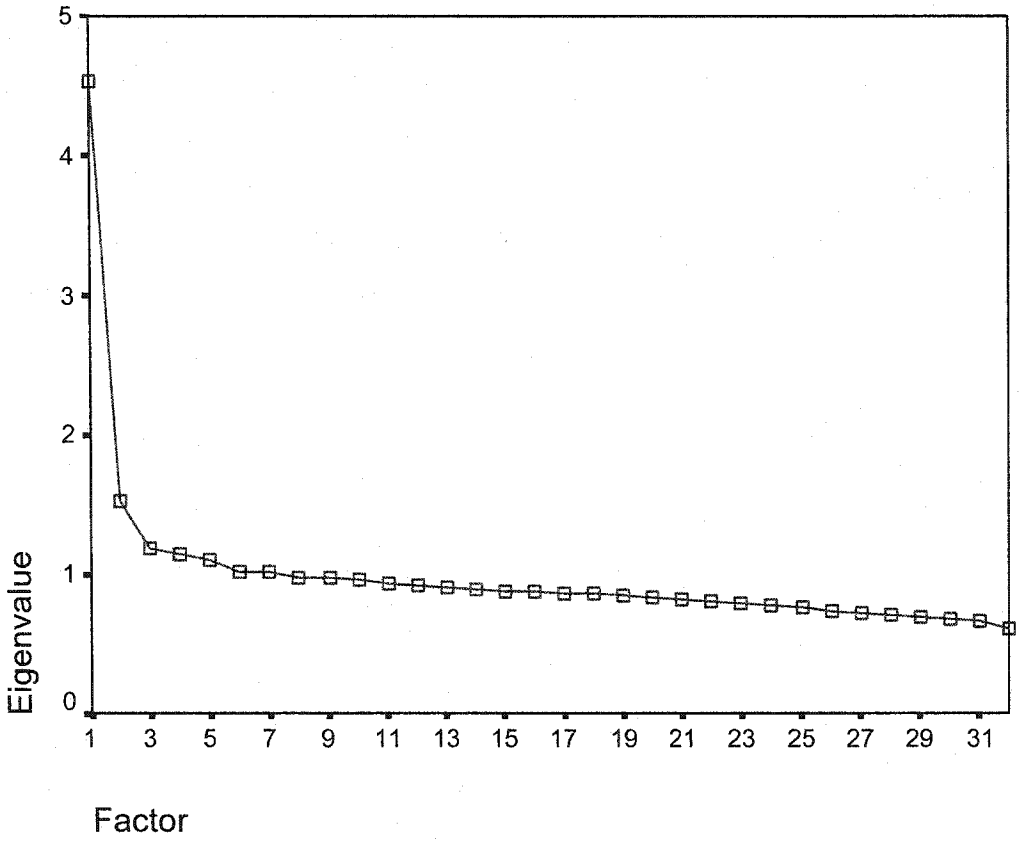
APPENDIX F

Non-Aboriginal Grade 4 Numeracy Scree Plot



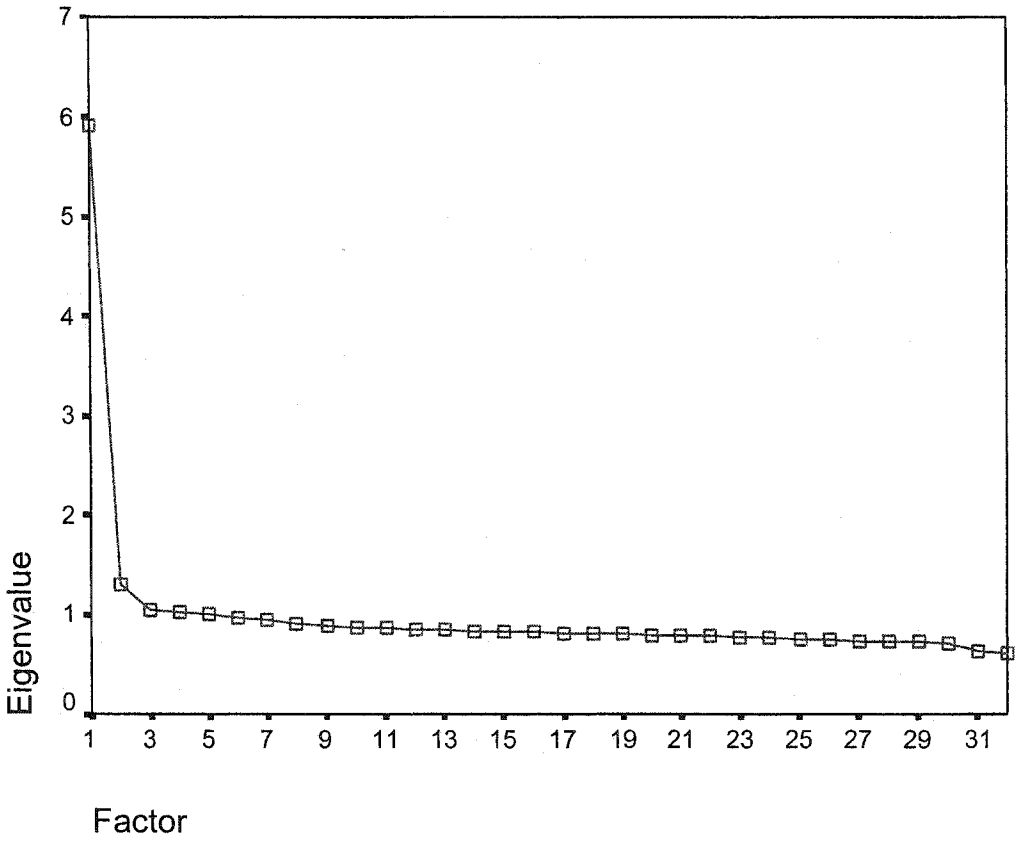
APPENDIX G

Aboriginal Grade 7 Numeracy Scree Plot



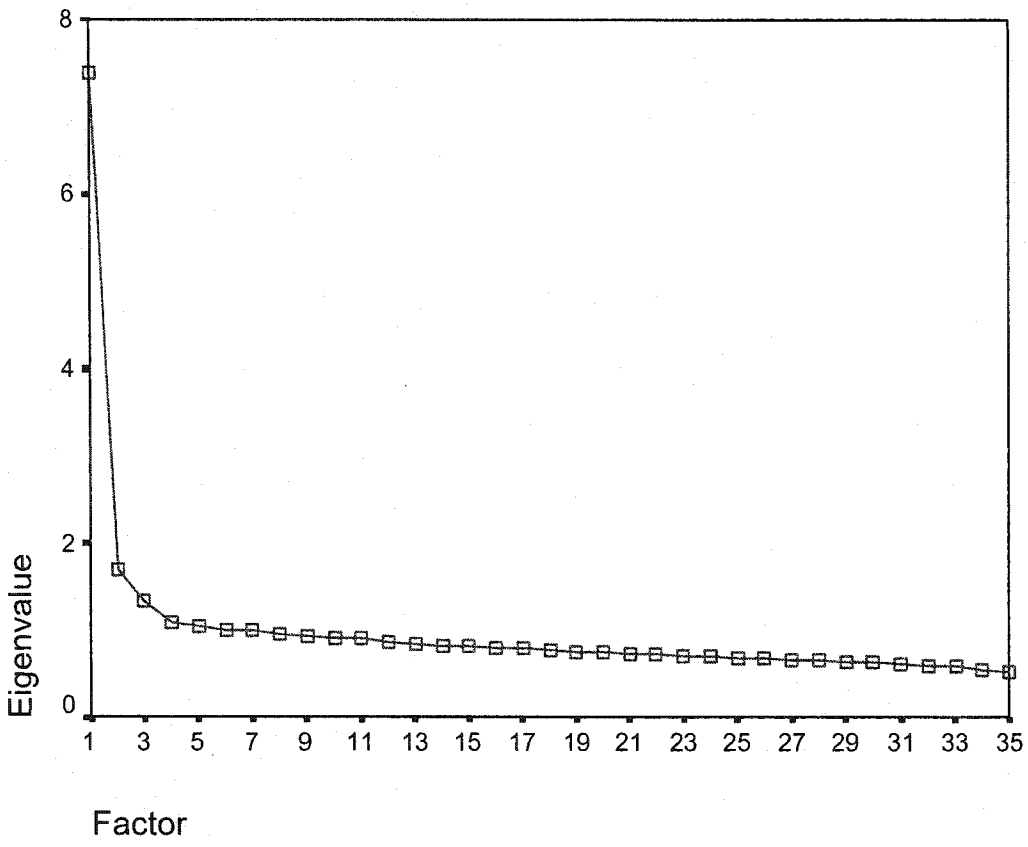
APPENDIX H

Non-Aboriginal Grade 7 Numeracy Scree Plot



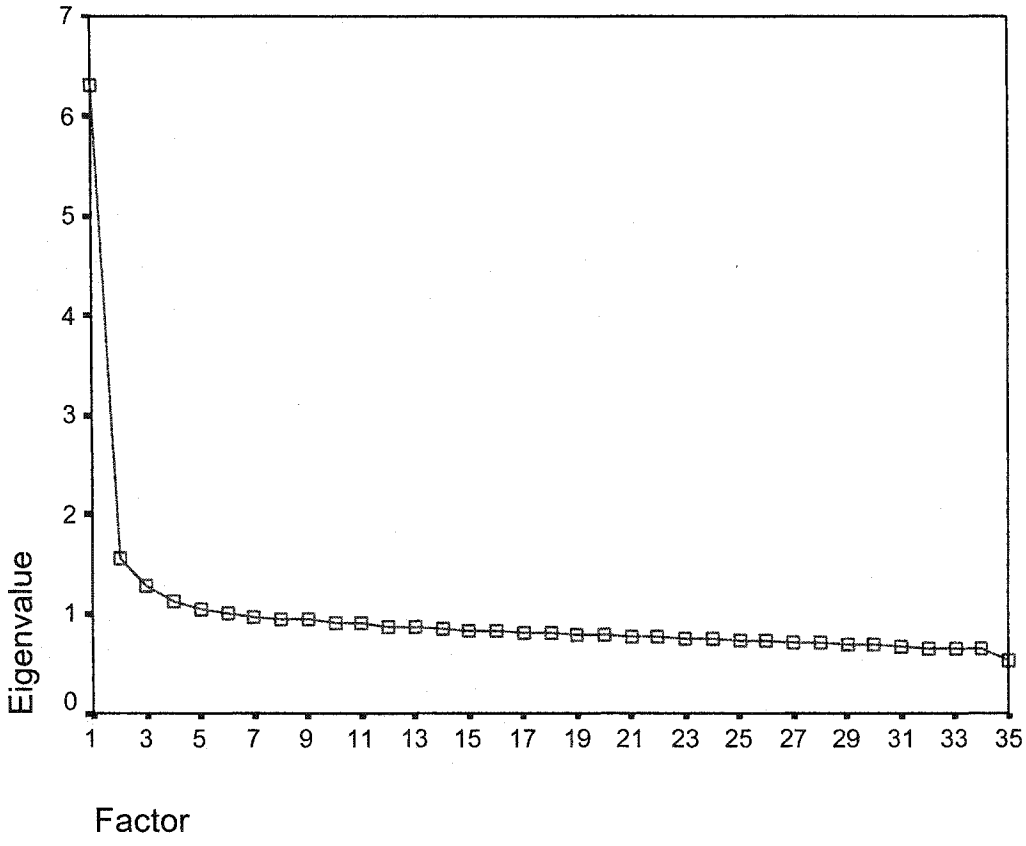
APPENDIX I

Aboriginal Grade 4 Reading Scree Plot



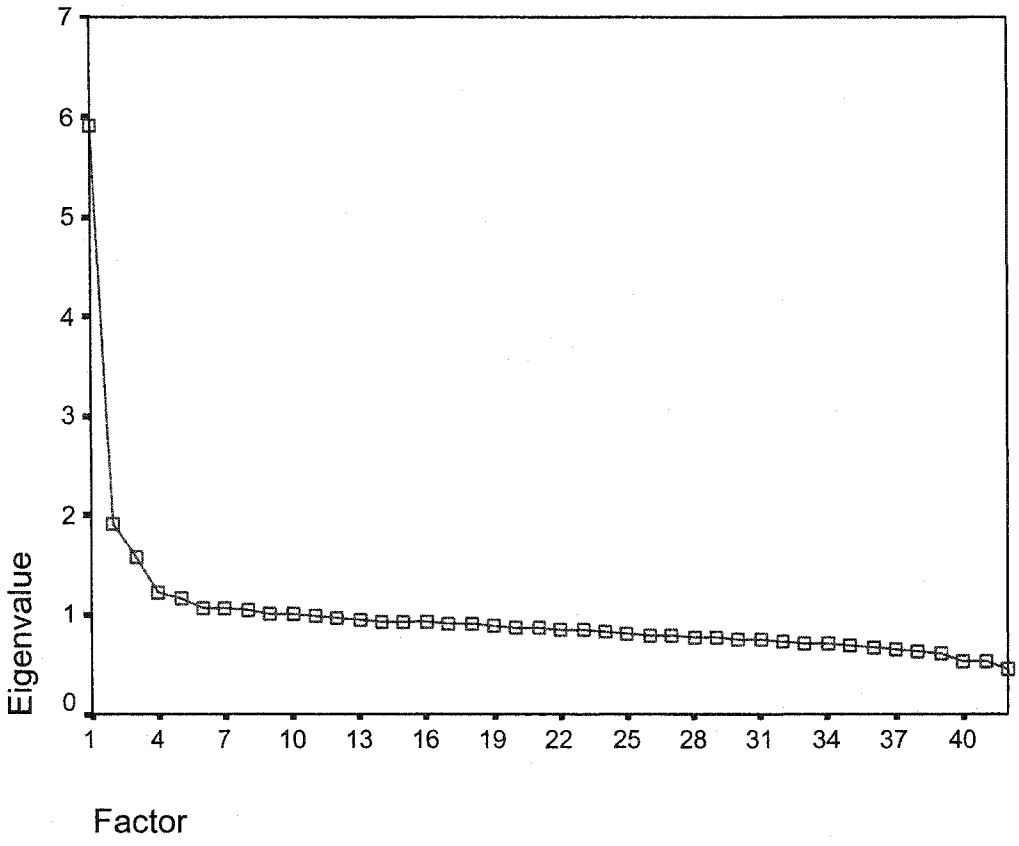
APPENDIX J

Non-Aboriginal Grade 4 Reading Scree Plot



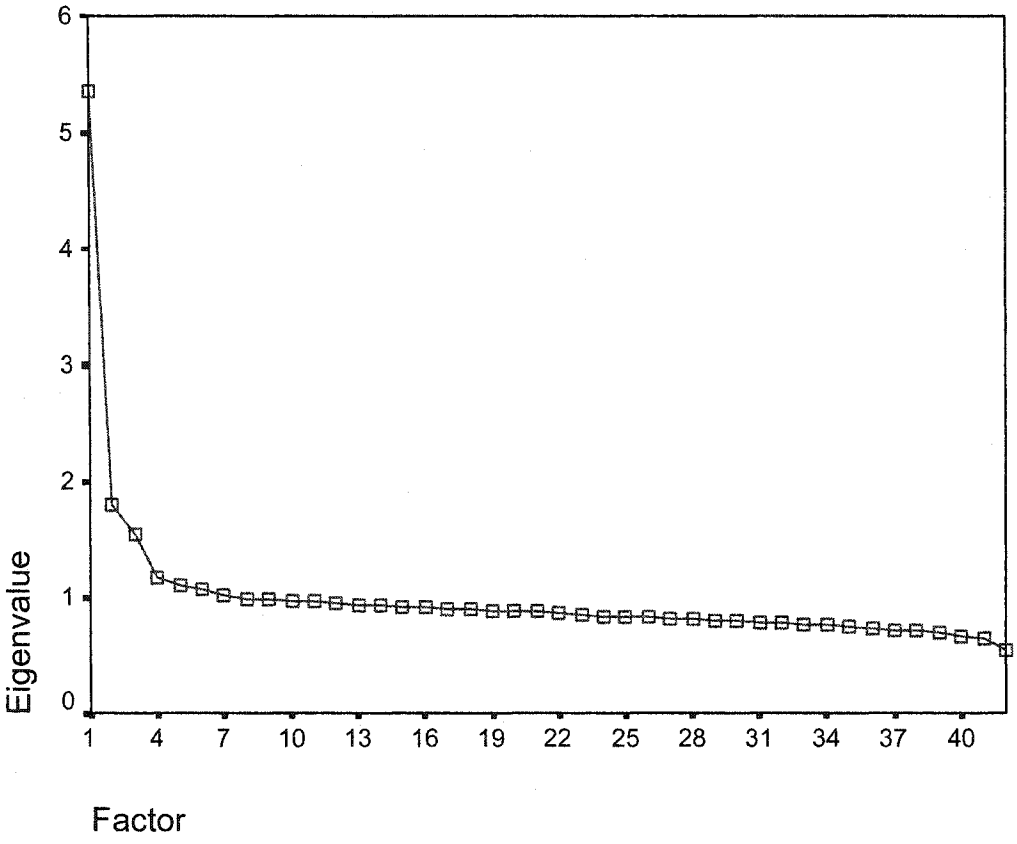
APPENDIX K

Aboriginal Grade 7 Reading Scree Plot



APPENDIX L

Non-Aboriginal Grade 7 Reading Scree Plot



APPENDIX M

Numeracy: Eigenvalues Greater than One

Grade 4 Numeracy: Eigenvalues Greater than One

Population	Factor	Eigenvalue
Aboriginal		
	1	5.304
	2	1.237
	3	1.202
	4	1.093
	5	1.045
	6	1.031
Non-Aboriginal		
	1	5.433
	2	1.175
	3	1.099
	4	1.055

Grade 7 Numeracy: Eigenvalues Greater than One

Population	Factor	Eigenvalue
Aboriginal		
	1	4.534
	2	1.525
	3	1.183
	4	1.148
	5	1.104
	6	1.023
	7	1.019
Non-Aboriginal		
	1	5.916
	2	1.308
	3	1.041
	4	1.034
	5	1.007

APPENDIX N

Reading: Eigenvalues Greater than One

Grade 4 Reading: Eigenvalues Greater than One

Population	Factor	Eigenvalue
Aboriginal		
	1	7.391
	2	1.696
	3	1.327
	4	1.093
	5	1.045
Non-Aboriginal		
	1	6.301
	2	1.556
	3	1.286
	4	1.118
	5	1.048
	6	1.005

Grade 7 Reading: Eigenvalues Greater than One

Population	Factor	Eigenvalue
Aboriginal		
	1	5.918
	2	1.923
	3	1.585
	4	1.221
	5	1.170
	6	1.074
	7	1.063
	8	1.045
	9	1.016
Non-Aboriginal		
	1	5.918
	2	1.923
	3	1.585
	4	1.221
	5	1.170
	6	1.074
	7	1.063

APPENDIX O

Numeracy: Maximum Likelihood Estimations for the Number of Factors

Grade 4 Numeracy: Maximum Likelihood Estimations for the Number of Factors

Population	Factors	Chi-Square	Df	*p*-value
Aboriginal				
	1	1322.79	464	0.00
	2	1022.29	433	0.00
	7	387.49	293	0.00
	8	326.68	268	0.01
	9	276.19	244	0.08
Non-Aboriginal				
	1	9033.63	464	0.00
	2	5979.04	433	0.00
	18	88.89	73	0.10
	19	59.99	59	0.44
	20	43.44	46	0.58

Note. The critical *p*-value was set at 0.05.

Grade 7 Numeracy: Maximum Likelihood Estimations for the Number of Factors

Population	Factors	Chi-Square	Df	*p*-value
Aboriginal				
	1	1618.11	464	0.00
	2	1009.92	433	0.00
	8	329.97	268	0.01
	9	282.18	244	0.05
	10	234.76	221	0.25
Non-Aboriginal				
	1	11906.19	464	0.00
	2	7440.72	433	0.00
	18	137.09	73	0.00
	19	91.35	59	0.00
	20	60.77	46	0.07

Note. The critical *p*-value was set at 0.05.

APPENDIX P

Reading: Maximum Likelihood Estimations for the Number of Factors

Grade 4 Reading: Maximum Likelihood Estimations for the Number of Factors

Population	Factors	Chi-Square	Df	*p*-value
Aboriginal				
	1	3329.90	560	0.00
	2	1930.52	526	0.00
	9	398.24	316	0.00
	10	342.81	290	0.02
	11	287.95	265	0.16
Non-Aboriginal				
	1	28714.35	560	0.00
	2	15039.99	526	0.00
	18	209.69	118	0.00
	19	144.52	101	0.00
	20	99.03	85	0.14

Note. The critical *p*-value was set at 0.05.

Grade 7 Reading: Maximum Likelihood Estimations for the Number of Factors

Population	Factors	Chi-Square	Df	*p*-value
Aboriginal				
	1	4329.02	819	0.00
	2	2323.94	778	0.00
	8	652.92	553	0.00
	9	581.99	519	0.03
	10	517.56	486	0.16
Non-Aboriginal				
	1	40415.89	819	0.00
	2	19827.57	778	0.00
	22	232.16	168	0.00
	23	191.67	148	0.01
	24	154.00	129	0.07

Note. The critical *p*-value was set at 0.05.

APPENDIX Q

Pattern Matrix for Aboriginal Grade 4 Numeracy Scores

Item	Factor		
	1	2	3
1	0.14	0.02	0.27
2	0.25	0.11	0.14
3	0.35	0.10	0.15
4	0.16	0.01	0.25
5	-0.04	-0.09	0.44
6	0.02	0.01	0.34
7	0.08	-0.04	0.29
8	0.05	0.00	0.32
9	-0.02	-0.06	0.37
10	-0.11	-0.02	0.34
11	0.01	-0.03	0.45
12	0.13	0.14	0.38
13	0.09	0.00	0.33
14	-0.02	0.04	0.36
15	0.31	0.10	0.20
16	0.28	0.10	0.13
19	0.42	-0.04	0.00
20	0.44	-0.02	-0.09
21	0.29	-0.04	0.04
22	0.50	-0.04	-0.02
23	0.44	-0.04	0.05
24	0.29	0.00	-0.05
25	0.44	0.04	-0.06
26	0.30	0.00	0.20
27	0.18	-0.08	0.20
28	0.38	-0.07	0.08
29	0.23	-0.20	0.07
30	0.14	-0.43	0.12
31	0.13	-0.40	0.16
32	0.33	-0.18	0.05
33	0.10	-0.11	0.05
34	0.31	-0.06	0.00

APPENDIX R

Pattern Matrix for Non-Aboriginal Grade 4 Numeracy Scores

Item	Factor			
	1	2	3	4
1	0.34	-0.04	0.14	0.09
2	0.18	-0.08	0.13	-0.06
3	0.24	-0.16	0.27	-0.15
4	-0.05	-0.09	0.34	-0.20
5	-0.01	0.09	0.48	0.21
6	0.11	-0.07	0.24	-0.12
7	0.01	0.02	0.30	-0.05
8	0.28	-0.11	0.18	-0.01
9	0.22	-0.05	0.29	0.11
10	-0.08	-0.07	0.34	-0.01
11	0.08	0.07	0.40	0.10
12	0.75	-0.18	-0.01	0.10
13	-0.03	0.00	0.37	-0.04
14	0.36	-0.08	0.12	0.06
15	0.19	-0.16	0.18	-0.26
16	0.25	-0.10	-0.01	-0.24
19	0.11	0.03	-0.01	-0.28
20	0.18	0.02	-0.12	-0.29
21	-0.08	-0.05	0.06	-0.44
22	0.18	0.06	-0.06	-0.32
23	0.15	0.10	-0.10	-0.32
24	-0.14	-0.07	-0.04	-0.56
25	-0.04	-0.13	-0.06	-0.69
26	0.82	-0.06	-0.22	0.02
27	0.53	0.07	-0.17	-0.04
28	0.00	0.08	0.20	-0.22
29	0.13	0.17	-0.05	-0.11
30	-0.13	0.78	-0.10	0.12
31	-0.15	0.67	0.00	0.06
32	-0.14	0.35	0.06	-0.15
33	0.03	0.08	-0.09	-0.15
34	0.25	0.02	-0.13	-0.22

APPENDIX S

Pattern Matrix for Aboriginal Grade 7 Numeracy Scores

Item	Factor				
	1	2	3	4	5
1	0.38	0.24	-0.02	-0.11	-0.25
2	0.14	0.15	-0.06	0.06	0.06
3	-0.06	0.31	0.00	0.39	-0.15
4	-0.09	-0.06	-0.47	0.05	-0.03
5	-0.22	-0.13	-1.03	-0.14	-0.01
6	0.47	0.04	0.00	-0.10	-0.25
7	0.04	0.30	-0.05	0.03	-0.03
8	0.11	0.52	0.03	0.02	-0.06
9	0.08	0.40	0.00	-0.05	-0.01
10	-0.04	0.43	0.00	0.04	-0.03
11	-0.16	0.38	0.08	-0.01	0.04
12	0.29	0.19	0.09	-0.12	-0.03
13	0.22	0.29	0.03	-0.06	-0.04
14	0.21	0.10	-0.01	-0.02	0.04
15	0.33	0.10	-0.02	-0.10	-0.02
16	0.15	0.11	-0.02	-0.08	0.12
19	0.40	0.10	0.06	0.10	0.07
20	0.52	-0.06	0.08	0.02	-0.01
21	0.59	-0.15	0.02	0.07	-0.06
22	0.00	-0.03	-0.02	-0.05	0.24
23	0.20	0.16	0.03	0.00	0.03
24	0.13	-0.05	-0.02	0.31	0.10
25	0.18	0.04	0.01	0.07	0.18
26	0.39	0.01	0.08	0.05	0.11
27	0.50	-0.08	0.02	0.05	0.02
28	0.19	0.22	0.06	0.05	0.10
29	0.04	0.10	0.00	0.13	0.24
30	0.03	0.09	-0.03	0.11	0.26
31	0.42	-0.06	0.04	-0.05	0.10
32	0.31	0.00	0.05	0.03	0.17
33	0.12	0.11	-0.03	0.00	0.28
34	0.20	0.08	0.00	0.06	0.20

APPENDIX T

Pattern Matrix for Non-Aboriginal Grade 7 Numeracy Scores

Item	Factor	
	1	2
1	-0.14	0.53
2	0.40	0.07
3	0.48	-0.13
4	0.36	-0.02
5	0.43	0.11
6	-0.24	0.54
7	0.32	0.12
8	0.47	0.10
9	0.19	0.28
10	0.47	0.00
11	0.51	-0.14
12	0.03	0.31
13	0.19	0.31
14	0.31	0.07
15	0.06	0.34
16	0.26	0.17
19	0.14	0.38
20	-0.01	0.36
21	-0.19	0.52
22	0.35	-0.12
23	0.13	0.28
24	0.40	-0.13
25	0.25	0.16
26	0.17	0.25
27	-0.05	0.43
28	0.41	0.09
29	0.48	-0.09
30	0.48	-0.07
31	0.07	0.29
32	0.12	0.25
33	0.44	0.07
34	0.33	0.18

APPENDIX U

Pattern Matrix for Aboriginal Grade 4 Reading Scores

Item	Factor		
	1	2	3
1	0.02	-0.46	0.22
2	0.14	-0.52	0.03
3	0.17	-0.63	0.07
4	0.08	-0.35	0.08
5	-0.04	-0.34	0.09
6	0.05	-0.49	0.20
7	-0.03	-0.42	0.21
8	0.02	-0.47	0.08
10	0.11	-0.66	0.23
11	0.12	-0.65	0.11
12	0.17	-0.77	0.25
13	0.01	-0.58	0.14
14	0.05	-0.61	0.01
15	-0.04	-0.41	0.19
16	-0.12	-0.44	0.13
17	-0.05	-0.31	0.12
18	-0.13	-0.31	0.13
20	0.39	0.02	0.13
21	0.39	0.08	0.12
22	0.33	0.01	0.17
23	0.24	0.06	0.07
24	0.50	0.01	0.01
25	0.33	-0.09	0.06
26	0.50	0.04	0.03
27	0.71	0.24	-0.17
28	0.54	-0.09	-0.03
30	0.49	0.08	-0.03
31	0.36	-0.02	0.14
32	0.42	0.06	0.13
33	0.44	0.01	0.09
34	0.63	0.12	-0.09
35	0.63	0.08	-0.03
36	0.36	-0.01	0.34
37	0.27	0.01	0.24
38	0.42	-0.09	0.17

APPENDIX V

Pattern Matrix for Non-Aboriginal Grade 4 Reading Scores

Item	Factor		
	1	2	3
1	-0.03	-0.49	-0.09
2	-0.14	-0.40	0.09
3	-0.10	-0.40	0.09
4	-0.05	-0.39	0.05
5	-0.03	-0.36	0.03
6	-0.08	-0.56	0.07
7	-0.07	-0.59	0.09
8	-0.05	-0.51	0.00
10	-0.03	-0.29	0.45
11	-0.02	-0.36	0.33
12	-0.06	-0.33	0.59
13	0.03	-0.36	0.19
14	0.00	-0.45	0.19
15	0.06	-0.38	0.01
16	0.08	-0.47	0.02
17	0.02	-0.36	0.03
18	0.00	-0.44	0.03
20	0.26	-0.10	0.09
21	0.32	-0.09	0.09
22	0.29	-0.15	0.14
23	0.21	0.09	0.01
24	0.45	0.01	0.06
25	0.31	-0.07	0.08
26	0.44	-0.05	0.06
27	0.55	0.16	0.01
28	0.53	-0.05	-0.03
30	0.35	0.07	-0.02
31	0.26	-0.22	0.14
32	0.32	-0.05	0.12
33	0.38	-0.09	0.13
34	0.53	0.12	-0.03
35	0.55	0.11	-0.01
36	0.24	-0.34	0.23
37	0.20	-0.23	0.19
38	0.37	-0.16	0.13

APPENDIX W

Pattern Matrix for Aboriginal Grade 7 Reading Scores

Item	Factor		
	1	2	3
1	0.11	-0.04	0.27
2	0.19	-0.13	-0.01
3	-0.06	-0.05	0.36
4	0.03	-0.06	0.34
5	0.08	0.03	0.25
7	0.16	-0.16	0.07
8	-0.10	-0.08	0.20
9	0.11	-0.08	0.23
10	0.13	-0.10	0.16
11	0.12	-0.08	0.09
12	-0.08	0.03	0.45
13	0.02	-0.11	0.26
14	0.02	-0.11	0.31
15	0.05	-0.13	0.19
17	-0.08	-0.68	-0.03
18	-0.15	-0.84	-0.08
19	-0.08	-0.55	0.12
20	-0.05	-0.29	-0.04
21	-0.13	-0.82	-0.10
22	0.54	0.05	-0.01
23	0.59	0.07	-0.07
24	0.25	0.05	0.11
25	0.49	0.11	0.07
26	0.20	0.05	0.06
27	0.32	0.10	0.08
28	0.27	0.02	0.35
29	0.00	0.03	0.09
30	0.18	0.02	0.28
31	0.28	0.07	0.28
32	0.08	0.05	0.20
33	0.40	0.07	0.16
35	0.50	0.05	0.02
36	0.28	0.00	0.11
37	0.66	0.01	-0.14
38	0.33	0.01	0.16
39	0.73	0.00	-0.22
40	0.47	-0.03	-0.26
41	0.34	0.04	-0.02
42	0.27	0.04	0.15
43	0.22	0.07	0.28
44	0.10	0.00	0.23
45	-0.05	0.02	0.22

APPENDIX X

Pattern Matrix for Non-Aboriginal Grade 7 Reading Scores

Item	Factor		
	1	2	3
1	-0.01	-0.01	0.33
2	-0.10	-0.16	0.02
3	0.11	0.06	0.51
4	0.03	0.02	0.45
5	-0.05	-0.01	0.19
7	-0.09	-0.11	0.10
8	0.10	0.04	0.37
9	-0.03	0.00	0.36
10	0.00	-0.09	0.19
11	-0.03	-0.04	0.12
12	0.12	0.06	0.50
13	0.02	-0.01	0.36
14	0.05	-0.03	0.38
15	0.03	-0.10	0.27
17	0.12	-0.65	-0.10
18	0.15	-0.92	-0.23
19	0.10	-0.63	0.01
20	0.11	-0.31	0.02
21	0.17	-0.90	-0.24
22	0.49	0.09	0.02
23	0.57	0.08	-0.07
24	-0.19	0.08	0.17
25	0.41	0.11	0.07
26	0.13	0.07	0.16
27	0.29	0.05	0.05
28	0.26	0.11	0.39
29	0.07	0.05	0.27
30	-0.10	0.08	0.39
31	-0.16	0.12	0.41
32	-0.08	0.07	0.23
33	0.38	0.10	0.19
35	0.45	0.08	0.02
36	0.24	0.08	0.18
37	0.58	0.07	-0.06
38	0.34	0.08	0.19
39	0.70	0.05	-0.17
40	0.42	0.03	-0.23
41	-0.24	0.04	0.06
42	-0.20	0.07	0.19
43	-0.08	0.08	0.41
44	-0.03	0.04	0.36
45	0.02	0.05	0.27