Gathering Evidence for Construct Validity: A Case for Large-Scale Educational Assessments

by

Nancy Walt
B.A., University of Victoria, 1984

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF ARTS

In the Department of Educational Psychology and Leadership Studies

We accept this thesis as conforming to the required standard

[redacted]

Dr. J. O. Anderson, Supervisor (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. W. J. C. Walsh, Departmental Member (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. R. E. Graves, Outside Member (Department of Psychology)

[redacted]

Dr. A. Preece, External Examiner (Department of Curriculum and Instruction)

Supervisor: Dr. John O. Anderson

ABSTRACT

The purpose of this study was to identify and evaluate the validity evidence related to British Columbia's Reading Comprehension Assessment. To gather validity evidence, the framework outlined in the *Standards for Education and Psychological Testing* (1999) was utilized. The methods included both qualitative procedures (document analysis, personal interviews) and quantitative procedures (reliability analyses, classical and item response theory analyses, factor analyses, and correlational analyses). It was determined that the methods employed in this study were generally adequate for providing a sampling of validity-related evidence. Overall, it was found that the evidence supporting assessment score interpretation was mixed; however, the evidence was deemed to be more positive than negative. Of substantial interest was the lack of support for the reporting of sub-domain reading scores.

Examiners:

[redacted]

Dr. J.O. Anderson, Supervisor (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. W.J.C. Walsh, Departmental Member (Department of Educational Psychology and Leadership Studies)

[redacted]

Dr. R.E. Graves, Outside Member (Department of Psychology)

[redacted]

Dr. A. Preece, External Examiner (Department of Curriculum and Instruction)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The author would like to thank and acknowledge the following individuals for their expertise, advice, and support during the development and conduct of this study:

John Anderson, John Walsh, and Roger Graves, as members of the Thesis Committee; Gerald Morton, Jim Gaskill, Britta-Gundersen-Bryden, Jerry Mussio, Darryl Hunter, Becky Matthews, Dehui Xing, John Eastaugh, Valerie Collins, Marcus Baer, Diane Lalancette, and Jemie Li from the British Columbia Ministry of Education; Mike Marshall of Applied Research and Evaluation Services, University of British Columbia; Arnold Toutant of A. Toutant Consulting; and Sharon Jeroski, Horizon Research and Evaluation Inc.

CHAPTER 1: INTRODUCTION

Large-scale Assessment

Large-scale educational assessment is a form of student assessment, which is group-administered and involves standardized administration procedures. Large-scale assessments, which broadly survey learning, are external to regular classroom assessment. These assessments, which involve either a census or sample of students at particular grade levels, measure student learning in specified subjects at a particular point in time.

Generally, large-scale assessments are developed and administered by departments of education, national and international assessment groups, publishers, or research groups. Administration of these externally prepared assessments is on a "large scale" in comparison to classroom assessments and can allow for use of results at the provincial, district, school, and student levels. Generally, each department of education makes decisions on whether or not to administer large-scale assessments (Walt, 1999).

In many United States and Canadian jurisdictions there is widespread use of large-scale assessment programs, and in a number of cases, jurisdictions have administered assessments of student learning since the 1970s.

In Canada, the majority of provinces and territories have introduced large-scale assessment programs that monitor student

achievement in core subject areas (i.e., language arts, mathematics) at various grade levels. In addition, most Canadian jurisdictions administer exit or senior examinations in a number of subject areas. These large-scale assessment programs are administered for a variety of purposes, including, collecting information for improving programs, evaluating programs and schools, instructional planning, addressing the need for public accountability, and certifying the achievement of students. Provincial assessments tend to monitor the overall education system, they are usually administered in the midst of a student's program, and generally do not affect a student's mark. Exit examinations, on the other hand, are often used for summative evaluation purposes at the end of a student's program and tend to count towards graduation (Walt, 2002).

In 1998, the Council of Chief State School Officers reported that 48 states had a state-wide assessment program in one or more subject area. In addition, 22 states reported having a policy mandating that students pass a high school exit examination and 7 states were in the process of developing exit tests (CCSSO, 1998).

The widespread and growing use of large-scale assessment programs can be attributed to the fact that parents, and the public in general, are demanding more information about student progress from the school system. In addition, schools and districts are in the need of solid data in order to monitor student growth and improvement plans (Walt, 1999).

Monitoring Large-scale Assessment Processes

Large-scale assessment programs are information systems that require monitoring for validity, accuracy, and consistency. Because of greater use and emphasis being placed on test results by schools, districts, and the public at large, it is important that information be provided to support valid test interpretation and use. As noted by Haladyna (2002a), "testing programs have an obligation to communicate with their constituency about the purpose of the test and the validity of test score interpretations and uses" (p. 89). In addition, when the stakes are raised for students, expectations for providing validity evidence are more demanding (Linn, 2002).

There are a number of professional organizations in Canada and the United States that promote good testing practices. These organizations have developed technical standards or "codes of behavior about testing that involve just about everyone in testing" (Haladyna, 2002b, p.40). These guidelines and standards for educational testing outline requirements for test developers, particularly that test developers are responsible for providing specific information at each test administration. The majority of these standards have to do with validity and responsible test interpretation.

British Columbia's Student Assessment Program

In British Columbia (BC), a large-scale provincial student assessment program has been in existence since the mid-1970s. The assessment is currently titled the Foundation Skills Assessment (FSA) and it purports to measure students' reading comprehension, writing, and numeracy skills. This annual assessment is administered in each skill area to students in grades four, seven, and ten. Reports are generated for the province, districts, schools, and individual students.

From 1976 to 1998, a technical report was produced to accompany each BC provincial assessment administration (see Appendix A). This report included information about how each assessment was designed, developed, and administered, and provided technical information about the psychometric properties of the assessment instrument. While technical information tended to focus on the reliability of each assessment instrument, validity information was provided in the form of the assessment design and development descriptions.

Since 1998, a technical report has not been produced to accompany BC provincial assessments. Detailed information about test development, validity or reliability is currently not documented or readily available. This has largely been due to lack of time and staffing resources (J. Gaskill, personal communication, December 2002).

As the purpose of the BC provincial assessments has changed and the stakes for students are somewhat higher than earlier assessments

(i.e., individual results are generated), the need for validity information is even greater.

Focus of This Research Study

This research study will explore the topic of measurement validity with respect to British Columbia's provincial student assessment program. Because of the issue outlined above, that is, an absence of validity evidence for BC's provincial assessment program, this study will attempt to address the following question—what evidence can be gathered to judge the validity of the BC FSA for its main purpose of evaluating how well students are achieving basic skills? To answer this question, the 2001 Grade 4 reading comprehension component of the FSA will constitute the focus of this study.

Although the FSA has several purposes, the validity evidence gathered in this study is mainly in support of interpreting results at a provincial level. Because validating test score interpretation and uses is a very extensive and lengthy process (some argue it is ongoing), the procedures proposed in this study are also limited to those most feasible for a department of education to conduct and sustain following annual assessment administrations.

CHAPTER 2: REVIEW OF THE LITERATURE

What is Validity?

Validity is generally defined as the extent to which an instrument measures what it claims to measure for a given purpose. Validity is concerned both with what a test measures and how well it does so, therefore, it tells us what can be inferred from test scores. As Messick (1995) states "validity is not a property of the test or assessment as such, but rather of the meaning of the test scores" (p. 741). The 1985 edition of the *Standards for Educational and Psychological Testing* states that validity "refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (AERA et al., 1985, p. 9). The most recent edition of the *Standards for Educational and Psychological Testing* defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA et al., 1999, p. 9). The main change in the definition of validity is from the nature of the test score to the use of the test score.

Since validity is not considered solely a characteristic of a test, it is not appropriate to say a test has high or low validity. Because no test is valid for all purposes, validity "must be established with reference to the particular use for which the test is being considered" (Anastasi & Urbina, 1997, p. 113).

To many, validity is the most important quality to consider in testing (Gronlund, 1993; AERA et al., 1999). Ultimately, one is after test scores that provide a representative and relevant measure of the construct under consideration for the purpose at hand. According to the Joint Committee on Standards for Educational Evaluation (1981), "validity is the most fundamental concern in the use of any measurement process. It matters little how reliably something is measured if it results in the wrong inferences." (p. 117).

Traditional Approaches for Evaluating Validity

Traditional approaches for evaluating validity were classified into one of three categories—content validity, criterion-related validity, and construct validity. For some researchers, these distinctions still hold true (see Sireci, 1998).

Content validity is concerned with an instrument's content relevance to, and representativeness of, the domain area of interest. Content validation, therefore, evaluates the degree to which an instrument adequately "covers" the domain of interest (Nunnally, 1967). Content validity is evaluated by comparing test content with the content domain it is designed to measure. Content validation approaches are mainly descriptive in nature and involve expert opinion.

Content validation is typically attended to during test development through the utilization of systematic procedures for specifying and

constructing test items. For example, subject-matter experts are consulted and curricula are reviewed regarding the domain area of interest, the actual test items, and appropriate item format. Then a table of specifications is developed which guides item writers in developing an adequate range and proportion of items for the domain of interest. A table of specifications shows the content or topic areas to be covered, processes or skills to be tested, and the relative importance of individual content and skills. The final specifications should indicate the number of items of each kind to be prepared for each topic (Anastasi & Urbina, 1997).

While content validation procedures are built in during the test development process through the use and documentation of systematic procedures, content validation studies can also be conducted following the development process. Typical procedures include having a group of independent experts (aside from the item writers) judge whether the items adequately represent the domain of interest (see Crocker & Algina, 1986).

Criterion-related validity indicates the degree to which test scores relate to some other measure of interest (criterion) (Suen, 1990). Criterion-related validation is used for situations where one wishes to draw an inference from an examinee's test score to performance on some real behavioural variable (e.g., using SAT scores as predictors of college performance) (Crocker & Algina, 1986). Criterion measures against which

test scores are validated may be obtained at the same time as the test scores (concurrent validity) or at a later date (predictive validity). Predictive validation is concerned with drawing inferences about examinees' future performance. It answers the question, "Does John have the prerequisites to become a satisfactory pilot?" Concurrent validity indicates the extent to which the test scores estimate an individual's present standing on the criterion rather than predicting future outcomes. It answers the question, "Does John qualify as a satisfactory pilot?" (Anastasi & Urbina, 1997).

Criterion-related validity evidence is often obtained through correlational studies. Validation indices are usually reported in terms of the correlation of test scores with the criterion measure(s).

Finally, construct validity is the extent to which a test may be said to measure a theoretical construct or trait (e.g., reading comprehension). As Haladyna (1999) stated, "the process of defining and measuring a construct and then validating the interpretation or use of a measure of a construct is the concern in construct validation" (p. 5).

Evidence in support of construct validity can take many forms. One approach is to demonstrate that the items within a measure are inter-related and therefore measure a single construct. As noted by Rudner (1993), inter-item correlations and factor analysis are often used to demonstrate relationships among items.

As may be evident, the distinction or lines drawn between the three traditional forms of validation are not clear-cut and are often blurred. For example, both content and criterion-related validity procedures provide evidence for construct validity. Because of these interrelationships, the traditional view that there are three discrete categories of validity has waned.

Construct Validity—A Unified Definition

Largely due to the work of Messick (1989, 1995), the traditional conception of validity divided into three separate types (content, criterion and construct) has been discarded. Construct validity is now viewed as the overarching category or unitary concept, which is based on various kinds of evidence (Gronlund & Linn, 1990). As Messick (1995) states, "construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 742). Accumulating evidence from a variety of sources helps to determine what a test measures and how test scores relate to other significant variables.

Messick (1995) describes six aspects of construct validity (content, substantive, structural, generalizability, external, consequential) and suggests that they might "function as general validity criteria or standards for all educational and psychological measurement" (p. 744). Messick's six aspects of construct validity are outlined as follows:

1. Content refers to the specification of the boundaries of the construct domain to be assessed; it relates to the content of the test, including its relevance to the construct and the representativeness of the sampling of the domain of interest.
2. Substantive relates to evidence of cognitive processes underlying performance; it is the connection of test behaviour to the theoretical rationale behind test behaviour.
3. Structural refers to the logical connection between item formats and scoring to the construct interpretation; the internal structure of the test should be consistent with what is known about the internal structure of the domain of interest.
4. Generalizability relates to how test scores remain consistent across different samples, tasks, and scorers; it includes issues related to reliability.
5. External refers to the patterns of relationships among test scores; test constructs should rationally account for the external pattern of correlations.
6. Consequential refers to the intended and unintended consequences of testing (Messick, 1995; Haladyna, 1999).

The 1999 edition of the *Standards for Psychological and Educational Testing* incorporates Messick's view of construct validity. Five of Messick's validity aspects are incorporated into chapter one on

validity. Messicks' generalizability aspect is covered in chapter two on reliability. The following summarizes the five lines of validity evidence outlined in the *Standards for Psychological and Educational Testing* (AERA et al., 1999):

1. Evidence based on test content.
2. Evidence based on response processes.
3. Evidence based on internal structure.
4. Evidence based on relations to other variables.
5. Evidence based on consequences of testing.

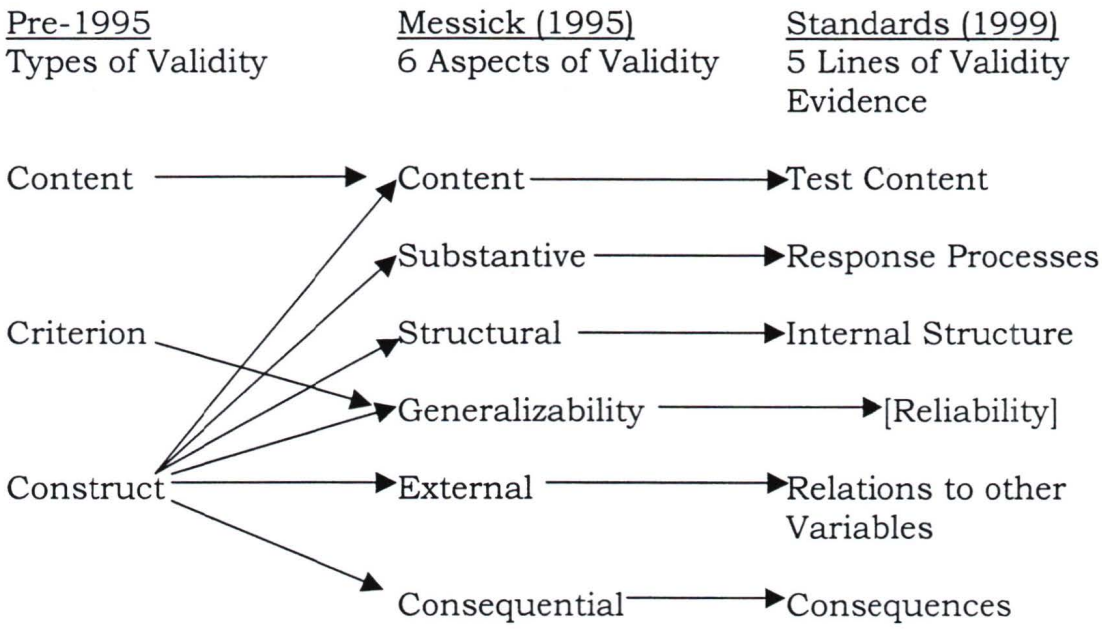


Figure 1. The evolving concept of validity.

In summary, the area of measurement validity has evolved over time (see Figure 1). Currently, construct validation is viewed as the process of compiling evidence that supports the use and interpretations

to be made of a given measurement score in a particular context. Construct validation, therefore, involves the collection of information from a variety of sources that will shed light on the nature of the construct, trait, or ability under consideration. Validity is not measured directly but rather inferred from all available evidence. "We take this evidence collectively as supporting or not supporting interpretations or uses to some degree" (Haladyna, 1999, p.12).

The latest edition of the *Standards for Psychological and Educational Testing* incorporates the idea of lines of validity evidence (as opposed to types of validity) and provides guidelines for gathering construct-related evidence of validity.

What is a Construct?

Traditionally, a construct has been defined as a theoretical, unobservable trait or attribute used to explain human behaviour (Anastasi & Urbina, 1997). Because of dispute over whether some tests are measures of constructs, the *Standards for Psychological and Educational Testing* (AERA et al., 1999) departs from a historical use of the term and defines a construct in a very broad sense "as the concept or characteristic that a test is designed to measure" (AERA et al., 1999, p.5). Examples of constructs, therefore, can include reading, writing, speaking, listening, creative thinking, critical thinking, problem solving, anxiety, mathematical aptitude, and intelligence (Haladyna, 1999;

Anastasi & Urbina, 1997). In the case of the FSA, the skills of reading comprehension, writing, and numeracy would be considered the constructs measured by each assessment. Both a unified view of construct validity and a broad use of the term "construct" will be adopted in this research study.

A Brief History of Provincial Assessments in BC

Between 1976 and 2003, there have been assessments of student achievement conducted by the BC Ministry of Education nearly every year. See Appendix A for a summary of BC provincial assessments.

From 1976 to 1998, the Provincial Learning Assessment Program (PLAP) developed assessments in the form of subject area assessments, each designed as an evaluation of student achievement within a particular curricular area. These assessments, delivered on a rotating cycle, were focused on the areas of language arts (reading, writing), mathematics, and science. However, there were also some exceptions—two social studies assessments, a French immersion assessment, a physical education assessment, and a kindergarten needs assessment were also conducted during this time.

These assessments were administered each spring to students generally in grades four, seven, and ten, roughly on a four-year cycle, with one subject-area assessment delivered each year. Reports of the results were typically produced for schools, districts, and the province.

The assessments from 1976 to 1998 were intended to assess how well the curriculum was being achieved by assessing students' knowledge and skills in a particular subject area. These assessments were designed as large-scale research projects with a program evaluation emphasis. These assessments included teacher and student background questionnaires, which provided interesting contextual and demographic information for answering specific research questions.

Beginning in 1999, the assessment program was redesigned to focus on critical foundation skills, namely reading comprehension, writing, and numeracy. The program, called the Foundation Skills Assessment (FSA), is now on an annual cycle, and students in Grades 4, 7, and 10 write the same three assessments each spring (British Columbia Ministry of Education, 2003c).

There are a number of notable differences between the PLAP and the FSA, such as the purpose for which the assessments are conducted, the way results are reported, and the cycle of assessment administration.

As noted, PLAP assessments were designed for the purpose of evaluating a particular curriculum area. These assessments provided information about students' achievement levels, classroom practices, and curriculum implementation.

These assessments, in general, placed a heavy emphasis on a multiple-choice format. All students wrote the multiple-choice questions, whereas, only a random sample of students wrote the open-ended

questions on a given assessment. While provincial, district, and school reports were typically produced, reports that were generated for schools and districts only included results for the multiple-choice component (J. Gaskill, personal communication, December 2002). In several instances (e.g., 1991 & 1993), the whole assessment employed a random sampling design, restricting the reporting of results to the provincial level only. This left schools and districts without performance trend information. PLAP assessments were also administered on a rotating basis and there were at least four years between repeat administrations of a subject area (see Appendix A).

The FSA on the other hand, is designed to provide annual information to the Ministry, school districts, schools, parents, students and the public about the performance of students in foundation skill areas (reading comprehension, writing, and numeracy).

All students receive the same questions relevant to their grade level, which include both multiple-choice and open-ended components. Reports to schools and districts are now based on both components. Results are reported in relation to provincial standards and data are provided regarding performance trends over time. Also, starting in 2000, students for the first time ever received individual results.

With the onset of the annual FSA in 1999, a greater emphasis was placed on supporting accountability initiatives. With the establishment of an annual administration, the assessment program has become more

stabilized and institutionalized. Because students in grades four, seven, and ten write the same three assessments every year, schools are becoming familiarized with the annual assessment procedures and are beginning to rely on annual data for setting goals and monitoring trends in student performance. Up until the cancellation of the BC School Accreditation Program in 2002, the use of the FSA results for monitoring student achievement was actually a legislative requirement for schools. However, schools and districts continue to draw heavily on the FSA results for school growth planning and district accountability contracts (British Columbia Ministry of Education, 2003b).

Because of the annual nature of the assessment, the FSA is becoming stabilized as an accountability measure and is being linked more heavily to school and district improvement efforts. The assessment has changed from "an event" to an ongoing process (J. Gaskill, personal communication, December 2002).

The change in provincial assessments from a cyclical program evaluation to an annual accountability measure is a trend consistent with other jurisdictions in Canada. In the early 1990s, many Canadian departmental assessment programs assessed subjects on a rotating cycle; however, annual administrations are now more common (Walt, 2002).

In summary, both assessment programs (PLAP and FSA) focus on measuring student achievement in particular areas, however the PLAP

had a research and program evaluation emphasis, intending to provide information about curricular attainment and implementation. In regards to the FSA, the more general look at a program area has fallen away. The FSA is concerned with measuring foundation skills that are considered important for all other areas of learning.

Purpose of the BC Foundation Skills Assessment

The Foundation Skills Assessment is an annual "assessment of BC students' academic skills, and provides a snapshot of how well BC students are learning foundation skills in Reading Comprehension, Writing, and Numeracy. The main purpose of the assessment is to help the province, school districts, and individual schools evaluate how well students are achieving basic skills, and make plans to improve student achievement. A secondary purpose is to provide teachers, students, and parents or guardians with an additional, external source of information about a student's performance in these important foundation skill areas." (British Columbia Ministry of Education, 2001b).

Reading Comprehension Assessment Component

In the case of the BC reading comprehension assessment, students' reading performance is reported at the provincial level, implying an underlying assumption that the test measures one construct or skill (i.e., reading comprehension). However, the table of specifications

(see Appendix C) also implies that the test is made up of three components or subcategories (identify and interpret key concepts and main ideas; locate, interpret, and organize details; and critical analysis), which are also somewhat distinct from one another. These categories appear to be somewhat of a carry-over from earlier (1988 & 1998) PLAP assessments of reading (see Gaskill, 1999).

Because the reading assessment presents students with a variety of reading passages along with multiple-choice and open-ended questions designed to assess comprehension skills, the reading comprehension construct under consideration appears to be skill-based rather than a measure of straight content knowledge. However, Kendall and Marzano (1997) describe three kinds of "content knowledge", procedural knowledge (involves skills and processes), declarative knowledge (involves understanding), and contextual knowledge (knowledge acquired in a unique context). As the BC reading comprehension assessment attempts to measure what students can do (versus what they know or understand), the skill of reading comprehension would be considered largely "procedural knowledge".

British Columbia Provincial Assessment Technical Reports

From the inception of the BC provincial assessment program in 1976 through to 1998, technical reports were produced to accompany each assessment administration. These technical reports provided

information to support interpretation of results at the provincial level; however, some of the information was also applicable to the interpretation of school and district results. For example, these reports were rich in information regarding test design, item development, pilot testing, administration, and marking. Psychometric characteristics of each test administration, such as the internal consistency reliability, were also presented.

Nine technical reports produced during a ten-year span (1988 to 1998) were reviewed for validity information (see Appendix B). Of the nine reports, six explicitly refer to test validity and three reports include implicit references to validity. However, references to validity are often made in passing through a sentence or short paragraph. For example, the 1989 assessment technical report states "content validity of the achievement forms was ensured by the item development procedures ..." (Bognar et al., 1989, p.17). While mention of validity is often brief or not explicit, all nine reports provide rich procedural information related to test design and development, which ultimately supports construct validity. For example, in each report, information is provided about the subject and technical experts involved in designing the table of specifications, writing items, and reviewing items. In eight out of nine cases, pilot-testing information is also presented. The 1991 technical report does not explicitly mention validity, however specific validation procedures such as "think-alouds" and factor analysis were utilized (see

Bateson, 1992). In the 1998 technical report, a section is devoted specifically to construct validity and represents the best example of attempting to present evidence of construct validity (see Gaskill, 1999).

Since 1998, no technical information is available or has been published to accompany BC provincial assessments. As previously mentioned, this has largely been due to shortages of resources. Another barrier to the development of a technical manual is the need for a format that is useful for a wide range of audiences (J. Gaskill, personal communication, December 2002). Detailed descriptions regarding the development procedures or documentation of the assessment technical qualities are currently not available or accessible to the public. This situation leaves the process open for criticism and allows for the mis-reporting of information (see Angus, 2002).

According to a number of recognized principles and guidelines for testing practices (AERA et al., 1999; Joint Advisory Committee, 1993; Canadian Psychological Association, 1996), validity evidence should be provided at each administration in the form of both descriptive procedures and empirical analyses. Because the results of BC assessments have higher stakes than in the past (i.e., student level results are now provided) and are being utilized for a broader range of purposes, the need for information about validity and reliability, has never been greater.

The following is an excerpt from the *British Columbia Foundation Skills Assessment 2001 Highlights* document. Aside from a table of specifications, it represents the extent of readily available information on BC's assessment development procedures:

- Test development begins in the fall prior to the administration with the creation/review of the tables of specification for each component of the FSA.
- Item-writing and item-review teams of approximately 100 practicing classroom teachers are selected in spring and fall prior to the administration.
- Test items and booklets are developed in accordance with the tables of specification for each component of the FSA and the prescribed learning outcomes listed in the provincial curricula.
- Additional design and development activities include the development of marking keys, social equity reviews of test items and booklets, and technical reviews by subject and measurement specialists.
- Test items and booklets are field tested in more than 180 classrooms throughout British Columbia.
- Upon completion of all of these activities, final test booklets are produced in the spring prior to administration (British Columbia Ministry of Education, 2001b).

What do other Jurisdictions do?

A review of other Canadian jurisdictions can shed light on the existence of large-scale assessment programs as well as the availability of supporting documentation on provincial-level validation procedures.

Most Canadian jurisdictions conduct provincial large-scale assessments, generally on an annual basis (see Walt, 2002). All provinces and territories, except Prince Edward Island, administer assessment programs of some kind. Assessment programs are "administered for a variety of purposes, including, collecting information for improving programs, evaluating programs and schools, instructional planning, addressing the need for public accountability, and certifying the achievement of students" (Walt, 2002, p. 3). Assessments are typically conducted in core areas (e.g., language arts (reading/writing), mathematics).

A review of other Canadian Ministry of Education websites, however, revealed little procedural or supporting technical information about existing assessment programs. Only one jurisdiction (Newfoundland) presented technical information about their assessment program, such as, the reliability and standard error of measurement (see Newfoundland and Labrador Department of Education, 1998). Five jurisdictions (Newfoundland, Ontario, Alberta, Saskatchewan, and the Yukon) discussed the development process, such as, the development of

a table of specifications and the field-testing of items. While a number of these jurisdictions publish general information about the assessment development process, the information is often bereft of details.

Websites were also reviewed in four state departments of education: Washington, New York, California, and Texas. It was noted that all four state department websites provided technical information about their programs, usually including sections on validity. For example, a Texas document titled *2000-2001 Technical Digest*, included an extensive section on test development procedures, and sections on both reliability and validity. In terms of validity, the traditional areas of content, construct, and criterion-related validity were each addressed. Content and construct validity were intertwined and were considered to be addressed during test development. A number of empirical studies were conducted as evidence of criterion-related validity (see Texas Education Agency, 2003).

The Washington state department website also included information on test development, item analyses, and evidence of validity. Validity evidence was the in form of internal and external validation using factor analysis and correlational techniques as well as comparisons of performance across groups (see Washington State Department of Education, 2001).

Testing Guidelines

A number of testing standards and guidelines each stress the importance of test validation and the attention that should be devoted to gathering validity evidence.

The Guidelines for Educational and Psychological Testing developed by the Canadian Psychological Association (1996) provides criteria for the evaluation of tests, testing practices, and the effects of test use. The document devotes sections to both validity and reliability.

The *Standards for Evaluation of Educational Programs, Projects, and Materials* developed by the Joint Committee on Standards for Educational Evaluation (1981), provides standards and guidelines for valid and reliable measurement.

The Principles for Fair Student Assessment Practices for Education in Canada (1993) contains a set of principles to guide fair assessment practices. Examples of validity-related principles directed to assessment developers are outlined below:

- Warn against common misuses of the assessment method.
- Describe the process by which the method was developed.
- Provide evidence that the assessment yields results that satisfy its intended purpose(s).
- Report evidence of the consistency and validity of the results produced by the assessment method for subgroups.

- Provide evidence of the effects on assessment results of such factors as speed, test-taking strategies, and attempts by students to present themselves favourably in their responses (Joint Advisory Committee, 1993, p.15-18).

The International Association for the Evaluation of Educational Achievement published a document titled *Technical Standards for IEA Studies.* This document includes a standard for validating constructs stating, "that every effort should be made to establish the validity and reliability of all measurement scales reported ..." and "to ensure that all measurement scales ... are valid measures of the intended constructs, so that consumers of study reports may have confidence in the validity of the results." (IEA, 1999, p. 75).

Finally, the *Standards for Education and Psychological Testing* devotes the first chapter to validity, in which it describes sources of validity evidence and outlines 24 validity-related standards. Chapter two is devoted to the topic of reliability and chapter six describes the test documentation standards required of test developers (AERA et al., 1999).

The Standards for Educational and Psychological Testing

The *Standards for Education and Psychological Testing* (herein called the Standards) is a document jointly produced by the American Educational Research Association, the American Psychological

Association, and the National Council for Measurement in Education. This document serves as a major reference for test developers, testing publishers, and test users and is recognized both in North America and internationally.

The *Standards* were initially produced in 1954 but have undergone four revisions. The latest edition, published in 1999, contains updated information reflecting the measurement trends in validity theory. As noted by Smith, "the current view on validity is based on the seminal work of Samuel Messick (1989, 1995) and is reflected in chapters 1 and 2 of the current 1999 edition of the Standards" (Smith, 2001, p.6).

The latest edition of the *Standards* incorporates the idea of lines of validity evidence (as opposed to types of validity) and provides guidelines for gathering construct-related evidence of validity.

Gathering Evidence to Support Construct Validity

As outlined by the *Standards*, the following types of evidence for construct validity can be gathered—1) evidence based on test content, 2) evidence based on response processes, 3) evidence based on internal structure, 4) evidence based on relations to other variables, and 5) evidence based on testing consequences.

Although validity is related to a specific test use and some interpretations of test scores may only require one or two types of

evidence, validity claims are stronger when evidence from all categories is present (Gronlund & Linn, 1990).

Evidence Based on Test Content

The *Standards* state that "important validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure" (AERA et al., 1999, p. 11). Sireci (1998) suggests that gathering content-related evidence utilizes either judgemental or statistical procedures. Judgemental methods involve subject matter experts in evaluating the representativeness and relevance of test items in relation to the content domain of interest. Statistical procedures involve the analysis of data resulting from test administration. Haladyna (1999) suggests that validity evidence can be statistical, empirical, or *procedural* in nature. He further suggests that procedural evidence would show that certain steps in test development were completed. Others also suggest that content-related matters are usually addressed during test development through appropriate test construction procedures (Gronlund & Linn, 1990; Sireci, 1998). Evidence from test development processes can, therefore, be procedural and descriptive in nature.

While inferences about content are linked to the process of test construction they are also related to the process of establishing evidence of validity after the test has been developed (Canadian Psychological

Association, 1996). For example, empirical or statistical procedures can supplement procedural information by checking for evidence of irrelevant content or factors affecting performance. Irrelevant content can be determined by identifying the types of errors typically made. Factors affecting performance might include motivation, test anxiety, fatigue, and speed.

Differential Item Functioning (DIF) is a statistical technique that reveals systematic differences among groups on a test score or test item response that are attributable to group membership instead of true differences in the construct being measured. "A test item displays DIF if examinees with the same trait level have different probabilities of responding to an item correctly" (Reise, 1999, p. 230). The existence of DIF, or bias, in item responses lowers the validity of test score interpretation and uses. This bias is a threat to valid interpretation or use of test scores, because bias favours one group of test-takers over another. "The study of DIF is essential for any test with significant consequences" (Haladyna, 1999, p. 183).

In summary, content-related evidence of validity is addressing two questions, "1) Does the test cover a representative sample of the specified skills and knowledge? 2) Is test performance reasonably free from the influence of irrelevant variables?" (Anastasi & Urbina, 1997, p.116). Content-related lines of evidence can come from judgemental, empirical, statistical, or procedural methods.

Evidence Based on Response Processes

As stated in the *Standards*, "...analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by examinees" (AERA et al., 1999, p. 12). Evidence based on response processes generally refers to evidence of cognitive processes underlying performance. It is the connection of test behaviour to the theoretical rationale behind test behaviour. For example, if one were claiming to be measuring problem-solving skills, then evidence that test takers are actually engaging in problem solving would be desirable. Evidence is generally derived from analyses of individual responses. Usual methods include questioning of test takers regarding their strategies or responses to particular items, or observation of the work methods employed by test takers.

Evidence Based on Internal Structure

As noted in the *Standards*, "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 1999, p. 13). Since "sources of evidence involve structural considerations based on investigations of the inter-item correlations and test dimensionality" (Ryan, 2002, p. 10), analyses such as internal-consistency reliability and

factor analysis techniques are common sources of internal structure evidence.

In addition, Haladyna (1999) suggests that item responses themselves are an integral aspect of construct validation. He stated, "not only are we concerned with the appropriate development of test items, but we are also concerned about responses to items that mimic our test scores" (p. 13). Responses to test items usually develop patterns. Items with desirable response patterns contribute to developing an effective test. For example, higher ability students should select the correct option in a multiple-choice question more often than lower ability students. Lower ability students should select the incorrect multiple-choice options (distractors) more often than higher ability students. In these cases, the items are functioning correctly and would lead to valid interpretations of test scores.

Items with undesirable item response patterns, on the other hand, weaken the validity of test score interpretations. For example, if all students obtain the correct answer, then the item is providing little information about differential student ability and should be eliminated. Additionally, if higher ability students select a distractor more often than lower ability students, then there may be a problem with the correct response or there may be two possible answers to the question. In either case, the item should be reviewed, revised, or eliminated.

By examining item performance, one can eliminate problematic questions or identify specific questions that may need revision. Item analysis often involves several steps, including an examination of item difficulty and item discrimination indices, and an evaluation of the effectiveness of each multiple-choice question option (distractor analysis) (Oosterhof, 1990).

Item response theory (IRT) is increasingly being used for construct validation. "Comparative fit indices for different IRT models can provide interpretations about the constructs that are measured" (Embretson, 1999, p. 2). For example, fit statistics can provide validity evidence by determining person or item departures from the general model underlying the test. Fit indices for persons determine which persons are responding in an unexpected way. Some of the factors that affect person-fit indexes are cheating, inattention, lack of motivation, random answering, plodding (speededness), and language deficiency. Item fit statistics determine which items do not fit within the unidimensional construct and which might be implying a different construct.

Evidence Based on Relations to Other Variables

The *Standards* state that the "analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence" (AERA et al., 1999, p. 13). For example, the scores of any particular test can be expected to correlate highly with scores of

other tests that presumably measure the same thing (convergent validity). Conversely, test scores can be expected to have lower correlations with measures purported to measure different abilities or underlying constructs (discriminant validity).

Substantial relationships of a test to other measures purportedly of the same construct, and the absence of relationships to measures purportedly of different constructs, support both the identification of constructs and distinctions among them.

Evidence Based on Consequences of Testing

Consequential aspects of construct validation "are concerned with score meaning and the intended and unintended consequences of assessment use" (Ryan, 2002, p. 10). A fundamental purpose of validation is to indicate whether specific benefits of testing are realized (AERA et al., 1999). Intended benefits or consequences might include public confidence, improved student learning, and teacher professional development. Unintended consequences might include teaching to the test, excessive test preparation, narrowing of the curriculum, imposition on instructional time, decreased student confidence, lack of student motivation, test anxiety, and cheating.

As noted by Ryan (2002), "questionnaires, classroom observations, and case studies are the most typical methods used to study consequences" (p. 10). Additionally, a study of test consequences might

evaluate the accuracy of inferences drawn from the results. For example "content analyses of media reports...could be used to look for evidence of inappropriate causal inferences from standards-based score reports" (Haertel, 2002, p. 20).

Summary

As noted earlier by Gronlund & Linn (1990), a stronger case for validity can be made when evidence from all categories is present. However, the area of validity is extremely vast and the validation process is viewed as an ongoing effort. As described in the following section, this study will limit the investigations to those analyses related to provincial-level interpretation of results and those feasible and sustainable for a department of education to conduct annually.

CHAPTER 3: METHOD

The purpose of this study was to identify and analyze the evidence that could be gathered for judging the validity of BC's reading comprehension assessment in relation to provincial-level interpretation of results. The 2001 FSA grade four English reading comprehension assessment served as the focus for this study and the Standards for Education and Psychological Testing (AERA et al., 1999), provided a framework for gathering evidence of construct validity.

As noted, construct validation involves the collection of information from a variety of sources that will shed light on the nature of the construct, trait, or ability under consideration. Validity is inferred from all available evidence in support of test score interpretation or use. Since the purpose of BC's reading comprehension assessment is to evaluate students' reading comprehension skills and report this information to the public, one would be interested in obtaining as much evidence as possible that the assessment does in fact measure reading comprehension skills. While a comprehensive approach to gathering validity-related evidence would arguably be best, a number of analyses and methods particularly related to response processes and testing consequences were outside the scope of this study.

To address the research question, analyses deemed most useful for interpreting provincial-level results were selected. Analyses were also limited to those considered feasible and sustainable for a department of

education to conduct annually following an assessment administration and do not include analyses which would involve lengthy studies or complicated research designs. The analyses meeting these criteria relate to the areas of test content, internal structure, and relations to other variables.

The methods employed in this study were both qualitative (procedural evidence) and quantitative (statistical evidence) in nature. All statistical analyses were conducted on existing student response files. Qualitative methods included analyses of published documentation and discussions with Ministry of Education staff.

At the time of this research study, the author was an employee of the Ministry of Education and a staff member of the department responsible for the FSA development and marking. During the conduct of this study, the author made every effort to remain unbiased and objective as the methods were employed.

The FSA Reading Comprehension Assessment Instrument

The 2001 Grade 4 FSA English reading comprehension assessment instrument consisted of 39 items, 35 of which were multiple choice and four of which were open ended. These items were spread across eight reading passages. The reading passages consisted of stories, poems, and articles representing a mixture of literary and informational genres. The questions asked students to locate information or details from the

passage, to decipher the main idea of the passage, or to draw conclusions from the passage.

The multiple-choice questions included a prompt and four possible options. The multiple-choice questions were scored as correct (assigned a score of 1) or incorrect (assigned a score of 0). The open-ended questions presented students with a prompt and required a written answer. The open-ended questions were scored according to a four-point rating scale. In total, the maximum score possible on the assessment was 51.

The eight reading passages were presented in two sections (Part 1 and Part 2). Each section was expected to take 45 minutes to complete. Schools were provided with guidelines for administering the assessment and were encouraged to give students a break between each section or to spread the assessment across different days (British Columbia Ministry of Education, 2001c).

The Datasets

Existing student level data available at the British Columbia Ministry of Education was obtained from the 2001 English Reading Comprehension component of the FSA. The Grade 4 reading data was the focus for the study; however, to complete the proposed analyses, Grade 4 student-level data from all 2001 FSA assessment components (reading, writing, and numeracy) were utilized. The Ministry of Education FSA datasets required for the analyses included the "item-level response

file" and the "student summary file". Permission to access these datasets was granted to the author of this study by the Student Assessment and Program Evaluation Branch of the British Columbia Ministry of Education.

The reading comprehension item-level response file includes students' responses to each item on the assessment. For multiple-choice questions, a student's response is recorded (A, B, C, D) and this response is also scored as another variable (1=correct, 0=incorrect). Students' responses to open-ended questions are scored on a scale of zero to four. Total scores on the multiple-choice component, the open-ended component, and on the full assessment are also provided. Finally, an analysis code is applied to each student record. Students who complete limited amounts of the assessment are flagged and are included in, or excluded from, various aspects of provincial analyses and reporting.

The student summary file includes students' scores on each component of the FSA. Students' raw score percent and students' score on a 3-point scale (not yet meeting expectations, meeting expectations, exceeding expectations) are presented for each assessment component (reading, writing, and numeracy). Additionally, a 2-parameter Item Response Theory (IRT) scaled score and error score are also provided for each student on the reading and numeracy assessments (writing cannot be scaled using IRT as the writing assessment is based on only two items). A variable identifying whether or not students were included in

the provincial analyses and reporting (based on percentage of the assessment completed) is also included. See Appendix D for further information about the datasets.

Gathering Evidence Based on Test Content

Test Development, Administration and Marking

As previously mentioned, validity evidence based on test content can be of a procedural nature, including evidence that particular test development steps were taken. The methods for gathering validity evidence based on test content included a review of published procedural information regarding test development and discussions with staff at the Ministry of Education.

For the 2001 FSA reading comprehension assessment, published print and website information was examined for procedural information about test development. The documentation included the following Ministry of Education references:

1. FSA test design and development (www.bced.gov.bc.ca)
2. Interpreting and communicating British Columbia foundation skills assessment results, 2001
3. British Columbia foundation skills assessment, highlights, 2001
4. Foundation skills assessment, instructions for teachers/invigilators, 2001

Informal in-person discussions were held with six Ministry staff members to obtain information about each step of the test development, administration, and marking process. The manager and two staff members responsible for test development and the manager and two staff members responsible for marking were informally interviewed. While no set interview protocol was employed, staff members were asked to describe in detail the activities that they coordinate and manage. Some follow-up email queries were conducted to clarify points. Internal working documentation was also obtained from these staff members where applicable or available.

To provide content-related support for construct validity, the steps in test development, administration, and marking were documented from all sources listed above. These steps were then compared with the 11 steps proposed by Haladyna (1999) in his chapter titled "Validity from item development procedures" (p. 145-161).

The purpose of this analysis was not to judge the quality of the FSA development procedures, but rather to identify and describe the procedures that are in place as they relate to test content validity evidence. While the description of test development procedures supports the provincial-level interpretation of results, knowing if systematic development procedures are in place supports all levels (district, school, and student) of interpretation.

Factors Irrelevant to Test Content

Gathering evidence based on test content can also include examining the response data for evidence of factors affecting performance that would be considered irrelevant to test content, sometimes called test-score pollution (Haladyna, 2002b). A reasonable check following an assessment administration is for the effect of speed. It is assumed that "examinees have sufficient time to attempt all items on the test and thus incorrect responses or omissions are due to limited ability rather that lack of the opportunity to attempt the questions. Speed in answering questions introduces another ability that would underlie test scores ..." (Goodman, 2001, p. 22).

Evidence of speed as a factor in the interpretation of test scores can be gathered by observing the number of omissions by question over the length of the assessment. If by the end of the assessment the number of omissions increases dramatically (e.g., 10-20%), it might be evidence that speed is a factor, which would raise validity concerns of the assessment's ability to measure students' reading comprehension skills. Following a procedure used in the 1998 Assessment of Reading Comprehension (see Gaskill, 1999), the percentage of omissions on each multiple-choice question was calculated and reviewed.

To conduct this analysis, the 2001 English reading comprehension item level student response dataset was utilized. This file consisted of 47,207 records, however, some initial cleaning was required. This

entailed removing students who were missing one of the components (open-ended or multiple-choice) of the assessment, students who did not complete 50% of the multiple-choice items, and students who did not score at least 15% of the total score on the test. These data cleaning criteria are used by the Ministry of Education to ensure that the reported results are based on a reasonable amount of information and do not include spurious student responses, such as students randomly responding test questions (J. Gaskill, personal communication, December 2002). This data-cleaning step omitted 1,820 students and left 45,387 student records. (The removal of 1,820 students is in itself a validity issue; that is, we know nothing about the ability of these students or whom they represent).

Using SPSS version 10.0, a frequency analysis was conducted on the multiple-choice responses. If a student did not answer a particular question, this is denoted as a blank in the dataset. The number of "blanks" on each question, as a percentage of the total responses, was calculated to determine the percentage of students who omitted each multiple-choice question. Finally, the percentage of omissions on each multiple-choice question was graphed using Microsoft Excel 2000.

This analysis is designed for interpretation of provincial-level results; however this analysis would be especially important to conduct at school and district levels. For example, if it were found that a large proportion of students did not complete the assessment in a school or

district, this would bear important information on the interpretation of reading comprehension results at those aggregate levels.

Gathering Evidence Based on Internal Structure

As noted in the *Standards*, "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 1999, p.13). For example, the conceptual framework for an assessment may imply a single dimension of behaviour, or it may imply several components that are expected to be homogeneous, but that are also distinct from each other. In this case, a factor analysis would be appropriate for identifying test dimensionality or underlying test structure.

Analyses investigating the assessment's internal structure included 1) reliability analyses, 2) a principal components analysis, 3) classical item analyses (item difficulty, item discrimination, and distractor analysis), and 4) item response theory analyses (item fit). The item-level dataset was utilized for this set of analyses.

Reliability Analyses

Reliability refers to the degree to which test scores are free of errors of measurement for a given group. Sources of measurement error include fatigue, nervousness, guessing, misinterpretation of instructions, and

content sampling. Measurement error reduces the usefulness of the measure and limits the confidence that can be placed in any given test score. The reliability must be sufficiently high in order to warrant the particular test interpretation and use.

The reliability measure for tests delivered in a single administration is called an internal consistency reliability analysis. The internal consistency reliability is an index of the extent to which the items "hold together" in the sense that if one responds in one way to say item 1, then one would tend to respond similarly to other items on the test.

Internal consistency reliability can be estimated by determining the relationship among test responses. A reliability analysis, which is based on inter-item correlations, provides information on the inter-item consistency of examinees' responses. The reliability index is, therefore, a measure of this consistency, or lack thereof.

The conventional internal consistency index is known as Cronbach's alpha or coefficient alpha (Crocker & Algina, 1986) and is denoted with the symbol "α". The Cronbach alpha statistic was utilized for this analysis as it can be used for tests consisting of both dichotomously scored (i.e., 0 and 1) and polychotomously scored (e.g., 1, 2, 3, 4, and 5) items. This index can vary from a minimum of 0.0 to a maximum of 1.0. An alpha approaching 1.0 suggests that the items comprising a test do have high internal consistency.

To make sound interpretations based on test scores, one is after high reliability indices. Haladyna (2002b), defines reliability indexes of greater than .90 as very reliable; indexes of .80 to .90 as good; and indexes below .80 as not very good (not much confidence can be placed in these scores). While it depends on the importance of the decision to be made on the basis of the test scores (Dick & Hagerty, 1971; Cronbach, 1971), other researchers (see Anastasi & Urbina, 1997) also suggest that reliabilities of over .80 are desirable. Therefore, reliabilities greater than .80 will be considered acceptable for this analysis.

Using SPSS version 10.0, an internal consistency reliability analysis (Cronbach's alpha) was conducted to determine the interrelatedness of the assessment items. Because the FSA reports overall student reading comprehension scores based on the full set of assessment items, Cronbach's alpha was calculated on all assessment items (multiple-choice and open-ended). In addition, student ability is also reported on subdomains, therefore, items referenced to each of the reporting categories were also examined for inter-item consistency; that is, reliability analyses were also conducted on these subsets of items.

Principal Components Analysis

Principal components analysis (PCA) or factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables. Factors are thought to

represent underlying processes that have created the correlations among the variables (Tabachnick & Fidell, 1996).

Because most achievement tests are considered to be a test of some ability (i.e., reading), an analysis providing evidence of unidimensionality would strengthen test score interpretation. Because overall reading scores are reported, there appears to be an underlying assumption that the FSA reading comprehension assessment is unidimensional. To test for unidimensionality or a dominant factor, a PCA using SPSS version 10.0 was conducted on the 39 items from the FSA reading assessment. A PCA in SPSS was chosen, as this analysis is based on Pearson Product correlations. These correlations are the most applicable to the FSA item format, which includes both multiple-choice (dichotomous) and open-ended (polychotomous) items. Additionally, a PCA was chosen over other factor analysis procedures, as results tend to be virtually the same especially with large data sets and a large number of variables (Tabachnick & Fidell, 1996).

Following a procedure outlined by Hambleton, Swaminathan, & Rogers (1991), the plot of the eigenvalues (screeplot) was examined to determine whether a dominant first factor is present. An eigenvalue is an index of the total variation explained by each factor (Tabachnick & Fidell, 1996).

Hambleton, Swaminathan, & Rogers further suggest that dominance of the first factor can be assumed if the eigenvalue of this

first factor is approximately five times larger than the second largest eigenvalue. Additionally, the second largest eigenvalue and remaining eigenvalues should be similar. Further, Reckase (1979) suggested that essential unidimensionality could be assumed if the first factor accounts for at least 20% of the test variance. These criteria were applied against the FSA data to test for a dominant factor and essential unidimensionality.

The principal components analysis was also used to evaluate underlying test structure and to determine if the items cluster together in a predictable fashion. When the items that are intended to be manifestations of the same dimension do "belong together" within the same factor, evidence of the internal structure aspect of construct validity is obtained (Suen, 1990). Three factors were extracted, equal to the number of theoretical subcategories within the internal structure of the construct. The principal components analysis was chosen as it is used when one it interested in reducing a large number of variables down to a smaller number of components (factors) (Tabachnick & Fidell, 1996). In this case 39 variables (or items) were being reduced to three hypothetical factors.

To allow for interpretation of the components related to the underlying test structure, the PCA solution was rotated. As noted by a number of researchers, judgement of the factor loadings should not be performed until the factors are rotated (Suen, 1990; Tabachnick & Fidell,

1996; Crocker and Algina, 1986). To aid in component (factor) interpretation, the principal components analysis solution was rotated using both oblique and orthogonal (varimax) rotations. While the varimax rotation allows for the easiest interpretation (Sapp, 2002) and is the most common (Suen, 1990), an oblique rotation is used when it is assumed that the underlying factors might be correlated. Following a common rule of thumb, only items with factor loadings higher than .30 were interpreted (Nunnally, 1978).

Item Difficulty, Item Discrimination and Distractor Analysis

The functioning of individual items can affect the validity of test score interpretations. Classical item analyses can identify items that are functioning correctly or incorrectly. Using ITEMAN version 3.5 (Assessment Systems Corporation, 1998), item difficulty (p values) and item discrimination (point-biserial correlations) indices were calculated on the multiple-choice responses from the item-level dataset. For the open-ended questions, item discrimination indexes (item-scale correlations) were also calculated using ITEMAN version 3.5 (Assessment Systems Corporation, 1998) and item average proportion scores (approximated p-values) were hand-calculated.

The difficulty of an item is defined as the percentage (or proportion) of examinees correctly answering the item. This index, which can range from 0 to 1, is also known as the p-value of the item. Easier items have

higher p-values. For example, items with p-values of .80 would be considered relatively easy, whereas items with p-values of .20 or .30 would be considered difficult. Items with p-values close to 0 or 1 should be revised or discarded because they do not provide information about differences among examinees' abilities.

In an achievement test like the FSA, one would expect a wide range of items at varying difficulty levels. That is, because the purpose of the assessment is to determine student achievement across a range of performance, items should reflect a wide range of difficulty. As noted by Allen & Yen (1979), "generally, item difficulties of about .30 to .70 maximize the information the test provides about differences among examinees" (p.121). Ministry of Education staff confirmed that a range of p-values of .30 to .80 is desired for the FSA. However, occasionally items with p-values over .80 are retained but their discrimination must also be acceptable (above .20) (J. Gaskill, personal communication, February 2003).

According to classical test theory (which is used for FSA item development), items with p-values of .50 offer the most information about differences among examinees, that is, item score variance is maximized. However, if all items had p values of .50 there would be no fine distinction among examinees' ability levels. Therefore, it is best to choose items with a range of p values that average around .50 (Allen & Yen, 1979). Because p-values are also affected by guessing they, in fact,

seldom have a floor of zero. For example, if random responding is occurring, a multiple-choice question with four options has a chance level of 25%. Therefore, for the greatest variation, it is suggested that the optimal item difficulty level is about halfway between the chance level (i.e., .25) and one, which in this 4-option scenario would be about .63 (Allen & Yen, 1979; Crocker and Algina, 1986; British Columbia Ministry of Education, 2003d).

It is important to note that p-values centring around .50 (or .63) would not necessarily be the desirable outcome in Item Response Theory (IRT). In IRT, one would generally desire a greater number of items centring on the test cut points(R. Graves, personal communication, July, 2003). However, the FSA is developed in a classical test theory context and, therefore, multiple-choice item difficulty values were reviewed and compared to the desired range of p-values (.30 to .80) and to the optimal average p-value (.63).

P-values (item average proportion scores) for the open-ended items were also estimated and reviewed. Although, multi-point open-ended items do not have p-values per se, an approximation can be calculated by taking the average number of points assigned to examinees and dividing by the maximum number of points available (British Columbia Ministry of Education, 2003d). For example, if 100 examinees answered a 4-point question and earned points distributed as: 0 (20 examinees), 1

(30 examinees), 2 (15 examinees), 3 (30 examinees), 4 (5 examinees), the p-value for this item would be:

$$(0)(20) + (1)(30 + (2)(15) + (3)(30) + (4)(5) / (4)(100) = 170/400 = .425$$

Item discrimination is another item characteristic that describes an "item's ability to measure individual differences that truly exist among test-takers" (Haladyna, 1999, p.166). Since "the purpose of testing is to differentiate among people with different ability levels, a good item should be able to discriminate those with high from those with low ability" (Suen, 1990, p. 75). There are over 50 item discrimination indices (Anastasi & Urbina, 1997); however, one commonly used item discrimination index is the point-biserial correlation. The point-biserial is the correlation between scores on the item and the total test score and provides an index of the discrimination value of the item (i.e., whether or not it differentiates between persons who scored high on the test and those who scored low). Point-biserial item discrimination values can range from -1.00 to 1.00. Because items should discriminate between ability levels, item discrimination indexes should be positive, indicating that a greater number of high scoring examines than low-scoring examinees answered the item correctly. A zero or negative discrimination value is undesirable and may indicate problems with the scoring of the item or the item wording.

In general, point-biserial correlation coefficients should be greater than .20. Items with coefficients less than .20 do not yield much information about differences among the abilities of the persons tested (De Ayala & Kelley, 1987).

Using ITEMAN version 3.6 (Assessment Systems Corporation, 1998) item discrimination indexes were calculated on the multiple-choice items and open-ended questions. FSA item discrimination coefficients were reviewed for any negative or zero correlations and to determine whether they were over .20.

Finally, FSA multiple-choice item distractors were reviewed for any problematic options. For example, it is undesirable to have distractors that are seldom or never chosen. Distractors should appeal to low-scoring test-takers and not appeal to high scoring test-takers. The point-biserial discrimination index (correlation between distractor performance and total test score) on each option should reveal a negative correlation with total test score for each distractor and a positive correlation for correct choices; meaning that lower ability students tend to select distractors more often than higher ability students and higher ability students are choosing the correct choice more often than lower ability students. These desirable item characteristics were compared to the results of the ITEMAN distractor analysis.

Item Response Theory

Item Response Theory (IRT) analyses can identify examinees or items that do not fit the model underlying the assessment. This is important in construct validation, as items or people that do not fit the model weaken test score interpretation.

For item response theory analyses, the Ministry of Education utilizes the PARSCALE software program, as this program can analyze items of both a dichotomous and polychotomous nature. Because PARSCALE does not calculate person fit statistics, only item fit statistics were calculated. Item fit statistics were calculated on the item-level data set using PARSCALE version 3.2 (Muracki & Bock, 1998). Item fit statistics in PARSCALE are based on Chi square statistic. Because the chi square statistic is very sensitive to sample size, the resulting item fit statistics are not very meaningful. That is, "even minor empirical departures from the model will result in many items being identified as misfitting" (Hambleton, Swaminathan, & Rogers, 1991, p.54-55). One solution is to convert the Chi square statistic to the Cramer's V statistic.

Cramer's V statistic is calculated with the following formula:

$$V = \sqrt{(X^2/N(L-1))}$$

where X^2 is the Chi square value, N is the number of students, and L is the number of score points for the item (2 for multiple-choice, 5 for open-ended). Because Cramer's V takes into consideration the sample size, the statistic brings the test to more reasonable fit decisions. The Cramer's V

statistic can run from 0 to 1, however, the smaller the value, the better the item fit. A Cramer's V greater than .20 has been recommended as indication of a misfitting item (M. Marshall, personal communication, March 2003). In this study, the Chi square statistic was converted to the Cramer's V statistic and considered against a benchmark of .20.

The analyses related to internal structure are most applicable to the interpretation of provincial-level results; however, a number of these analyses (e.g., item statistics) are relevant to the interpretation of reading comprehension results at all levels.

Gathering Evidence Based on Relations to Other Variables

Messick suggests that "a wide variety of correlational analyses are relevant to construct validation" (Messick, 1989, p.51). These analyses often examine the structure of test or construct scores in relation to other variables. For example, the reading comprehension construct as measured should be dicriminable from measures of similar or related constructs (i.e., writing ability).

To provide evidence for discriminant validity, correlations between students' FSA reading, writing, and numeracy scores were examined. The student summary file from the grade four 2001 English FSA was utilized for the correlational analysis. As noted earlier, this file contains students' summary scores on each of the reading, writing, and numeracy assessment components.

While this file contains 50,035 student records, some initial steps were necessary to prepare the records for comparison. First, the file was filtered on a provincial reporting variable. This variable identifies whether or not a student was included in the provincial-level results report. The students excluded during this filtering process are generally federally-funded band school students who participate in the assessment but because they do not fall under provincial jurisdiction, they are not included in the provincial report. This filtering process eliminated 404 students, leaving 49,631 student cases.

Second, the file was filtered on an analysis and reporting flag. As noted earlier, students who complete particular percentages of each assessment are categorized using a number of flags. For example, students may only complete the multiple-choice component and not the open-ended questions or students may only complete a small proportion of the assessment. In addition, some students are excluded from one or more component of the assessment. It was determined from Ministry of Education documentation, that students who receive a reporting flag of "5" are included in the provincial level reporting (D. Xing, personal communication, December 2002). Therefore, students with a reporting flag other than "5" on each assessment component were removed from the dataset. This process filtered another 5,334 students out of the dataset, leaving 44,297 student cases. The result of removing over 5,000

students from the dataset is due to the fact that not all students write all three assessment components (reading, writing, numeracy).

Using the resulting student summary file, students' overall scores on each of reading, writing, and numeracy were correlated. With SPSS version 10.0, a Pearson's bivariate correlation was calculated on the students' raw score percentages and scatterplots were produced. To aid in interpretation, the correlation coefficients were squared as a measure of the strength of the relationship (Tabachnick & Fidell, 1996). A squared correlation coefficient shows the percentage of variance explained in the first test score by the second test score. Haladyna suggests that squared correlations in the range of .30 to .40 would be considered moderate (Haladyna, 2002b). In terms of this analysis, low squared correlations were desirable as evidence that the tests measure separate constructs.

This analysis is designed for interpretation of provincial-level results; however similar analyses could be conducted at school and district levels. For example, a measure of convergent validity might entail correlations between students' FSA scores and letter grades in Language Arts.

Summary of Methods

Because this study is developing and evaluating procedures to judge the validity evidence related to the FSA's main purpose of evaluating British Columbia students' reading comprehension skills, the methods focus on the interpretation of provincial-level aggregated

results. While much of the validity evidence gathered in this study also supports district, school, and student level interpretation of results, additional information would be required for each of these interpretations and is outside the scope of this study. In summary, the methods address the following questions:

1. Are there systematic assessment development procedures in place? (Test content).
2. Are there irrelevant factors that might be interfering with test score interpretation at the provincial level? (Test content).
3. Are the item characteristics supporting sound interpretation at the provincial level? (Internal structure).
4. Are the provincial-level test scores relatively free from error? (Internal structure).
5. Do the items seem to be measuring a single dimension (i.e., reading comprehension)? How do items relate to the table of specifications? (Internal structure).
6. How do the provincial reading test scores correlate with other tests of different measures? (Relations to other variables).

Evidence gathered from these analyses will be used to judge the construct validity of the 2001 FSA English reading comprehension assessment for its main purpose of evaluating students' reading comprehension skills.

CHAPTER 4: RESULTS

The results of this study to evaluate the evidence supporting construct validity are organized under the following headings: evidence based on test content, evidence based on internal structure, and evidence based on relations to other variables.

Evidence Based on Test Content

Test Development, Administration, Marking, and Reporting

As noted by a number of researchers (Haladyna, 1999; Gronlund, 1993), content-related evidence of construct validity can be in a procedural form, showing that systematic test development procedures were adhered to. The following procedural evidence related to the FSA development, administration, marking, and standard setting was identified from published information and through discussions with Ministry of Education staff responsible for the development and marking of the FSA.

Test Development

The test development process starts with a conceptual framework document that guides all further development. The conceptual framework document is developed by the Ministry of Education in conjunction with contracted academic researchers knowledgeable in the area of language arts and assessment and measurement. The conceptual

framework document includes 1) a definition of reading, which is taken from the National Council of Teachers of English, 2) a table of specifications which outlines the categories assessed in the FSA reading comprehension component and the percentage weighting of items for Grades 4, 7, and 10, 3) linkages of the assessment to the BC Language Arts Curriculum and the BC Performance Standards (classroom assessment tools), 4) a description of what the assessment is intended and not intended to accomplish, and 5) the types of reading passages that are included in the reading comprehension assessment (e.g., literary & informational). An excerpt of this conceptual framework document is provided in Appendix E and the table of specifications is provided in Appendix C. The conceptual framework document is reviewed and updated from time to time by academic content and measurement experts, identified as knowledgeable in the field of language arts and measurement.

Next, teacher committees responsible for selecting reading passages and developing both multiple-choice and open-ended items, are established. Teacher committee members are provided with the framework document (mentioned above) to guide test development. Teachers include both experienced and new teachers, those who represent both genders, those who teach ESL, those who have high and low end students, those from public and independent schools, and those from different parts of the province geographically. This mix of teacher

background and experience is deemed important in the development of fair tests for students representing a wide variety of backgrounds and abilities; that is, the wide range of teacher experience is intended to represent the variability in the student population.

The teacher committees next select a large number and range of reading passages which would be considered age appropriate, engaging, and relevant. Theses reading passages, which are authentic and not written for the test, must represent a wide range of materials that a student might encounter both inside and outside of school. A mixture of literary (e.g., stories, poetry) and informational (e.g., charts, maps, recipes, schedules) works is chosen. Because the assessment is intended to measure reading across the curriculum, passages are chosen which are cross-curricular in nature, that is, one passage might deal with a science topic such as tree growth. Additionally, since the assessment is intended for all schools and students across the province, reading passages must be selected which will appeal to, or engage, a range of students.

Next, a social considerations review is conducted on the selected passages. This review is conducted by BC educators, knowledgeable in the areas of gender equity, special education, aboriginal education, English as a second language, and diversity. This review identifies potentially controversial, sensitive, biased, or offensive elements that may exist in the content or presentation of material proposed for use. An

example of the social considerations review criteria is provided in Appendix F.

This type of review is sometimes called a sensitivity review or a review for bias. This review identifies, for example, cultural stereotyping, gender-sensitive language, and appropriateness for students with special needs. Reading passages should be age appropriate, should appeal to both genders, be engaging, and reflect a range of difficulty. Passages should not be disturbing and should not produce an emotional response that would impede a student's ability to answer the question. Because the reading passages are scrutinized in this way, validity is built into the development process by ensuring that the assessment content is relevant for a wide range of students. Through this process, every effort is made to eliminate potential bias that that might result in unfairness of test score interpretation. Passages that do not meet the social considerations criteria are removed.

Following the social considerations review, teacher committees receive item writing training and then develop both open-ended and multiple-choice items for each passage. Answer keys for each question are also developed. This step is followed by a second social considerations review on both the passages and the items. Items undergo a technical review by measurement experts and an editorial review is also conducted.

Items and passages are revised based on the reviews and are compiled into field-test booklets. Field-testing is then conducted in classrooms at Grades 5, 8, and 11. Between 100 and 200 students are involved in the field-testing at each grade level. Aside from the testing of the items, the field-testing provides teacher feedback about the appropriateness of the material, problem areas, and wording or concepts that were unfamiliar to students. Sometimes the field-testing includes observations by Ministry of Education staff. Following field-testing, the student responses are marked and classical item statistics (difficulty and discrimination indices) are reviewed. The difficulty of items should fall within an acceptable range and the spread of scores should ensure that students are able to achieve high results. Multiple-choice distractors are also reviewed for any problems or implausible options.

In the past, timing pilots, separate from the field-testing, were also conducted. Timing pilots were conducted solely to determine how long it takes students to complete the assessment. However, it has been determined that enough information is known about the types of passages and the amount of material a student can complete in the time allotted, so timing pilots are no longer deemed necessary. (V. Collins, personal communication, January 2003).

Revisions are made based on the results of field-testing. Items are selected or revised based on classical item statistics. Items are selected

which demonstrate an appropriate level of difficulty, a balance of topics, and correct weighting according to the table of specifications.

Near to final test booklets are produced based on field-testing results. At this time, passages from previous years (for measuring change over time) are incorporated. Ministry staff and committee members conduct a final review to ensure overall balance as per the table of specifications. Final booklets are produced.

During the development process, two advisory groups meet to provide advice and recommendations to the Ministry of Education. The Advisory Committee on Provincial Assessments (ACPA) consists of members of educational stakeholder organizations. The ACPA meets three to four times a year to provide advice on the reporting and interpretation of the assessment results. The Technical Working Group (TWG) consists of technical and subject area measurement experts. The TWG was meeting about four to five times a year to advise the Ministry on overall technical aspects of the FSA and to reflect on the implications of the results. Since most of the major technical issues have been dealt with, the TWG meetings have currently ceased (J. Gaskill, personal communication, February 2003).

As noted by Haladyna (1999), "the various activities that comprise item development and review are essential in construct validation" (Haladyna, 1999, p.146). The ability of test makers to write test items properly and to review and improve these items is key to valid tests and

test score interpretation. Any factor contributing to the increased or reduced difficulty of the test is a form of bias and can contaminate the inferences we make from test results. By adhering to systematic development procedures, as appears to be the case with the FSA development, much of this bias is attended to at the outset.

Administration and Security

Standardized administration procedures and test security also play a role in validity. For instance, standardized administration procedures ensure that all students receive the same assessment, at the same time, under the same conditions across the province. In the case of the FSA, final test booklets are printed and distributed to schools. School personnel administer the assessment, following standardized procedures set by the Ministry of Education. Schools are asked to ensure administration is in an appropriate test-taking area, free from distractions, for example. Test invigilators are provided with guidelines for administering the assessment. It is recommended that the assessment be administered in two 45-minute blocks either on the same day or over two days.

Lack of test security through the exposure of test items, can compromise valid interpretation and uses of test scores. Therefore, school principals are instructed to keep the FSA materials in a secure location before, during, and after the assessment. All material (used and

unused) is to be returned to the Ministry intact. Schools are instructed not to photocopy or keep copies, although this is difficult to enforce (V. Collins, personal communication, January 2003).

Marking

After the assessment, marking of the open-ended questions occurs. Teacher markers are selected who represent a balance of experienced and new teachers, but they must have recent experience at the appropriate grade level. Prior to the marking, the marking chairs select exemplar or "anchor" papers to be used as examples of each level of performance for each of the open-ended questions. At the beginning of the marking session, teachers are trained to mark the open-ended questions using these anchor papers as guides or standards.

Reliability papers, which have been pre-scored by the marking chairs, are sprinkled throughout the bundles of student papers. The markers score these papers during the marking session; however, the markers do not know the score that was pre-assigned. After markers have scored these papers, the scores are compared to the pre-assigned scores. The percentage of agreements and correlations between the sets of scores on each open-ended question are calculated to determine the relationship and the closeness of the marks assigned. These measures provide an index of the inter-marker reliability on the open-ended questions. For example, on the 1998 Grade 4 reading assessment, the

average percent agreement on the open-ended questions was 62% and the correlation between the sets of marks was .74 (Gaskill, 1999).

In addition to the reliability papers, marker agreement papers are given to markers twice a day. All markers mark these papers at the same time. The scores that the markers assign are reviewed and the differences are tallied and discussed. Chairs of the marking session work with any teachers who have veered off the marking standards. These marker agreement papers are used for training and recalibration during the marking process (B. Gundersen-Bryden, personal communication, March 2003).

Developers of the assessment attend the marking session to observe and obtain feedback on the assessment instrument. Markers provide feedback on specific items and the assessment instrument in general, providing suggestions for improvement. Multiple-choice item statistics are reviewed and the marking team may recommend the deletion of any problematic items.

Because of the procedures in place for training and recalibrating markers, the marks assigned to the open-ended questions are intended to be reliable and fair to students and, therefore, contribute to valid score interpretations. Inter-rater reliability data, however, should be published.

Standard Setting

In 2000, standard setting sessions were held. Using a modified Angoff procedure (see Angoff, 1971), panels of teachers recommended the scores students needed to receive in order to meet or exceed provincial expectations in reading comprehension. The meets expectations standard is defined as the level of performance at which a student meets or exceeds the widely held expectations for the grade on this test. With no other information, this is the level below which a teacher would want to know more about the reasons for a student's low performance. The exceeds expectations standard is defined as the level of a student's performance that is beyond that at which a teacher would say the student has fully met the expectations of the grade on this test. Student performance would be considered excellent for the grade on this test. Students who have not attained the "meets expectations" standard are considered to be "not yet within expectations" (British Columbia Ministry of Education, 2000).

These standards set in 2000 are considered a benchmark against which improvement can be measured over time. Statistical techniques involving Item Response Theory are used to enable the performance on different years' tests to be compared against these common standards (British Columbia Ministry of Education, 2000a).

Reporting

The Ministry of Education provides aggregated FSA results for schools, districts, and the province. Individual student results are also provided to students, teachers and parents. Based on the previously set standards (or cut scores), students are placed in one of three categories—not yet within expectations, meets expectations, and exceeds expectations. In the aggregated reports for schools, districts, and the province, both the number and percent of students at each performance level are provided. Noticeably lacking in these reports are descriptions of what it means to "not yet meet", "meet", or "exceed" expectations.

The provincial-level results for the 2001 reading comprehension assessment are provided in Appendix G and a sample school report is also provided for reference in Appendix H. As noted from these examples, the number and proportion of students at each level of performance are provided. The proportion of students meeting or exceeding expectations is also presented along with a confidence interval. Confidence intervals are based on the standard error of measurement (SEM) and provide the level of confidence we can place on the results. For example, the sample school report (see Appendix H) displays the proportion of students meeting or exceeding expectations. This is reported with a 90% confidence interval based on the SEM. The sample school report shows that 63 +/- 6% of students meet or exceed expectations, indicating that if the same assessment were conducted with the same students, a large

number of times, one could be confident that the proportion of students meeting or exceeding expectations would be within the reported confidence intervals (57% and 69%), 9 times out of 10.

In addition to the proportion of students at each performance level, the provincial and school report examples present the overall distribution of scores, as well as the mean scores on the reading assessment subscale categories (see Appendix G and H). Finally, the second page of each example presents the results for subpopulations of students (e.g., males); however, the results are presented without confidence intervals.

As a point of interest, a sample student-level report is also provided in Appendix I. Based on the standard error surrounding the cut scores, student performance is either reported as falling completely within one of the performance levels or spanning two performance levels. That is, based on the standard error, if a student's performance is close to the cut points, it is determined that a student's performance fell somewhere between the two levels. Although it is not published information, in 2001, 15% of students fell between the not yet within and the meets expectations levels of performance and 8% fell between the meets expectations and the exceeds expectations categories on the reading assessment (M. Marshall, personal communication, 2003).

Summary of Development Procedures

Table 1 summarizes the steps involved in the FSA development process. As a comparison, the FSA development procedures were contrasted with the steps outlined by Haladyna (1999) in his chapter titled "Validity from Item Development Procedures" (p.145-161). Ten of the 11 steps outlined by Haladyna were matched by one or more of the FSA development steps (see Table 1). In comparison to Haladyna's steps, the only step missing in the FSA development process was a review of the cognitive behaviour and thought processes actually used by students to answer questions. This can be accomplished through "think-aloud" procedures during field-testing, where students are asked what they were thinking when they responded to individual items. As noted by Haladyna, "students' perceived thought processes involved in answer selection or answer creation can be very revealing about the actual thought processes involved". (Haladyna, 1999, p.158).

This analysis of the FSA development, administration, and marking procedures did not attempt to judge the quality of these processes, but rather to identify if systematic procedures were in place. It was noted that ten of the eleven steps outlined by Haladyna (1999) for ensuring construct validity through test development are in place.

Although the methods employed involved the identification and description of the test development procedures, an evaluation of the quality of the procedures might be investigated. In particular, the validity

of the conceptual framework and the standard setting process would be useful to conduct.

Table 1: FSA Reading Comprehension Development Process

FSA Development Steps	Steps Identified by Haladyna (1999)
Assessment Framework & Table of Specifications	Content Definition and Test Specifications
Establishment of Teacher Committees	Content Review (Content Experts)
Reading Passage Selection	Content Review (Content Experts)
Social Considerations Review #1	Sensitivity Review
Item Training and Item Writing	Item-Writer Training
Social Considerations Review #2	Sensitivity Review
	Key Check
Technical and Editorial Item Reviews	Review of Items for Item-Writing Errors
	Editorial Review
	Key Check
Field testing	Test-Taker Review
N/A	Cognitive Behaviour ("think-alouds")
Administration (including security procedures)	Test Security
Consultation with Advisory Groups	Content Review

Factors Irrelevant to Test Content

Factors irrelevant to test content (e.g., cheating, test anxiety, motivation) are a threat to valid interpretations and uses of test scores. That is, test scores can be increased or decreased due to a factor unrelated to what the test is supposed to measure. These factors are sometimes referred to as test score pollution, (Haladyna, 2002b) or construct-irrelevant variance (Messick, 1989).

In reviewing the possible affect of fatigue (or speed) on the reading comprehension assessment, the percentage of omissions on each multiple-choice question was examined to determine if omissions rose sharply over the course of the assessment.

The assessment administration procedures recommend that the assessment be administered in two separate sections across two different blocks of time, therefore, each section will be considered as a "mini" assessment. Figure 2 depicts the percent of omissions on each multiple-choice question for each section of the assessment. With the exception of multiple choice question #4 and multiple choice question #10, the percent of omissions in section one remained at or under .5% until question #15. Questions 15-18 form the end of the first section of the assessment, and as can be seen, the percent of omissions starts to rise across these questions, reaching almost two percent. This might indicate fatigue was starting to affect students' ability to complete the assessment. The two exceptions (#4 & #10) involved interpreting details or making inferences from the reading passages. While the majority of students who attempted these items obtained the correct answer, a number of students obviously found these items difficult and chose not to attempt these questions.

The percentage of omissions in section two of the reading assessment remained under .5% until question 30. As Figure 2 depicts, the percentage of omissions starts to rise from question 30 to 34,

reaching over 1%. From question 35 to 38, the percentage of omissions continues to rise to almost 3%, representing approximately 1,300 students.

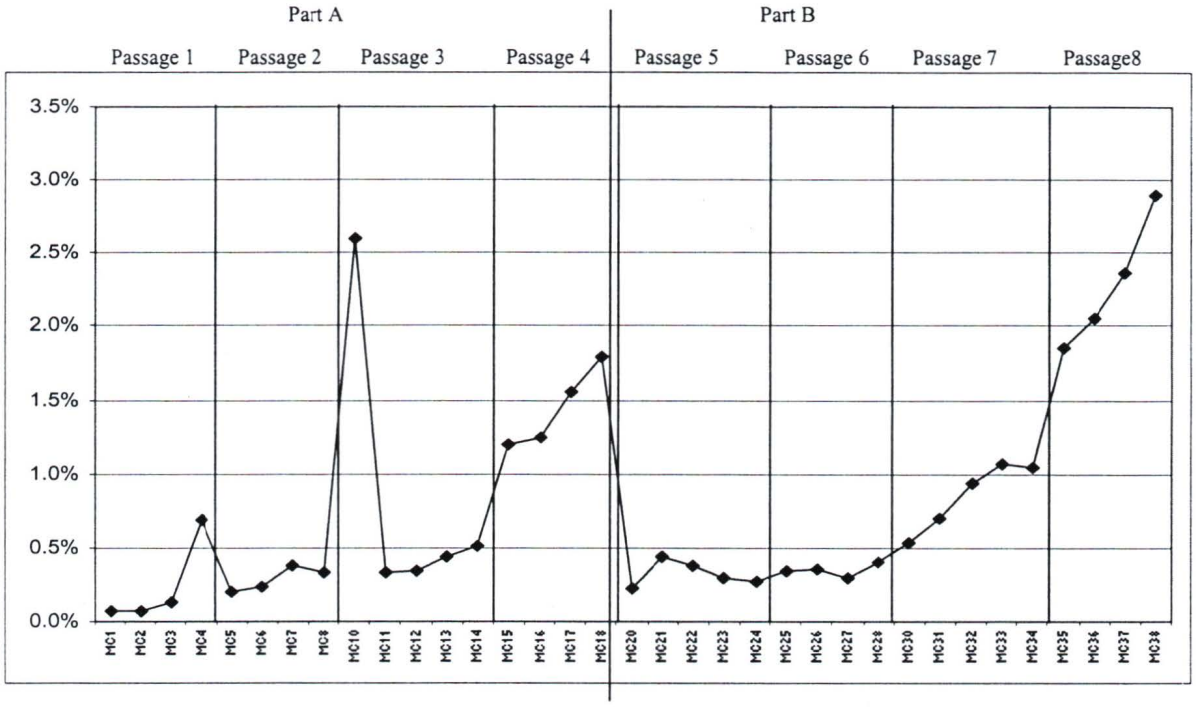


Figure 2. Percent of omissions by multiple-choice question

This finding is consistent with the pattern of results found in the 1998 BC assessment of reading comprehension; however, in 1998, the percentage of omissions rose to a high of 11 percent across the multiple-choice questions, an indication that timing was a factor (Gaskill, 1999).

The results of this analysis provide some information that an outside factor may be affecting (polluting) students' test scores. While 3% of omissions might not be considered indicative of speededeness, the rise in omissions at the end of each section of the assessment might indicate

that fatigue was interfering with some students' ability to complete the assessment.

Evidence Based on Internal Structure

Reliability Analysis

Reliability refers to the consistency of measurement and the error associated with that measurement. All test scores have random error in them; a reliability estimate offers a way to estimate this error.

An internal consistency reliability estimate (coefficient alpha) was calculated on the FSA reading assessment items. From Table 2, it is noted that the overall reliability estimate across all 39 items was .85, meeting the desirable reliability of .80 as suggested by Anastasi & Urbina, 1997 and Haladyna, 2002. The 35 multiple-choice items demonstrated the same reliability (.85) as the overall reliability; the four open-ended questions had a reliability estimate of .66. The correlation of the open-ended questions and the multiple-choice questions was .652.

Table 2

Internal-consistency reliability analysis by item type.

Item Type	No. of Items	Alpha
Multiple-Choice (MC)	35	.85
Open-Ended (OE)	4	.66
All Items (MC & OE)*	39	.85

**MC/OE Pearson Correlation = .652.*

In classical test theory, reliability is related to test length. Therefore, it is not surprising that the smaller group of four open-ended questions produced smaller reliability coefficients. However, for only four questions, .66 is a moderate reliability coefficient. As a point of interest, the Spearman-brown reliability formula was used to estimate the reliability of the open-ended questions by doubling the number items.

The Spearman-Brown formula expresses the effect of test length upon the reliability of a test, and is widely used in practice when estimating the reliability of a lengthened test. The Spearman-Brown is as follows:

$$K * \text{old reliability} / 1 + (K-1) * \text{old reliability}$$

where K is the proportional length of the new test in relation to the old test. By doubling the open-ended questions, the reliability estimate increased to .80. The open-ended items appear to be highly reliable in terms of internal consistency.

As noted earlier, items referenced to the same sections on the reading comprehension table of specifications should reveal fairly strong internal consistency. High alphas would show some support that these items are measuring the same subdomain. Standard 2.1 in the *Standards for Educational and Psychological Testing* also require that when scores are interpreted on subcategories, reliability indexes are also required for these subcategories (See AERA et al., 1999).

On the 2001 English reading comprehension assessment, the first domain (identify & interpret key concepts and main ideas) of 5 items (all multiple-choice) had an alpha of .43; the second domain (locate, interpret & organize details) of 26 items (25 multiple-choice and 1 open-ended) had an alpha of .78; and the third domain (critical analysis) of 8 items (5 multiple-choice and 3 open-ended) had an alpha of .63 (see Table 3).

Table 3

Reliability analyses related to the table of specifications.

Reporting Category	Items	Alpha
1. Identify and Interpret Key Concepts and Main Ideas	5 Multiple-Choice Items	.43
2. Locate, Interpret and Organize Details	25 Multiple-Choice Items; 1 Open-ended Item	.78
3. Critical Analysis	5 Multiple-Choice Items; 3 Open-ended Items	.63

Again, it is not surprising that the smaller groups of items produced smaller reliability coefficients. However, the reliability coefficients related to each subdomain of reading comprehension do not meet the minimum recommended acceptable standard of .80 as suggested by Anastasi & Urbina, 1997 and Haladyna, 2002b; that is, the confidence we can place in interpretation of these subscores at a provincial level is not overly high.

Principal Components Analysis

A principal components analysis was conducted to 1) determine unidimensionality and 2) to explore the assessment's underlying structure. As noted earlier, an analysis of internal structure can provide validity evidence for the theoretical construct underlying the assessment.

Tests for unidimensionality were evaluated in two ways. First, the scree plot was reviewed for evidence of a single dominant factor. The scree plot is a visual representation of the magnitude of each component's eigenvalue. Evidence of unidimensionality is determined by existence of a dominant first factor. That is, the eigenvalue of factor one should be much higher than that of factor two and the second largest eigenvalue should be hardly distinguishable from the remaining eigenvalues (Hambleton, Swaminathan & Rogers, 1991).

Second, the percent of variance accounted for by each component (factor) and the actual component eigenvalues were reviewed. Reckase (1979) suggests that, in order for a test to satisfy essential unidimensionality, factor one should account for 20% of test variance. Hambleton, Swaminathan & Rogers (1991) also suggest that the eigenvalue of component one should be 5 times larger than the eigenvalue of component two.

Results of the tests for unidimensionality are presented in Figure 3 and Table 4. Figure 3 displays the scree plot of eigenvalues for all

components (factors) of the reading comprehension assessment. Table 4 presents the eigenvalues that were greater than one and the percentage of variance accounted for by each factor.

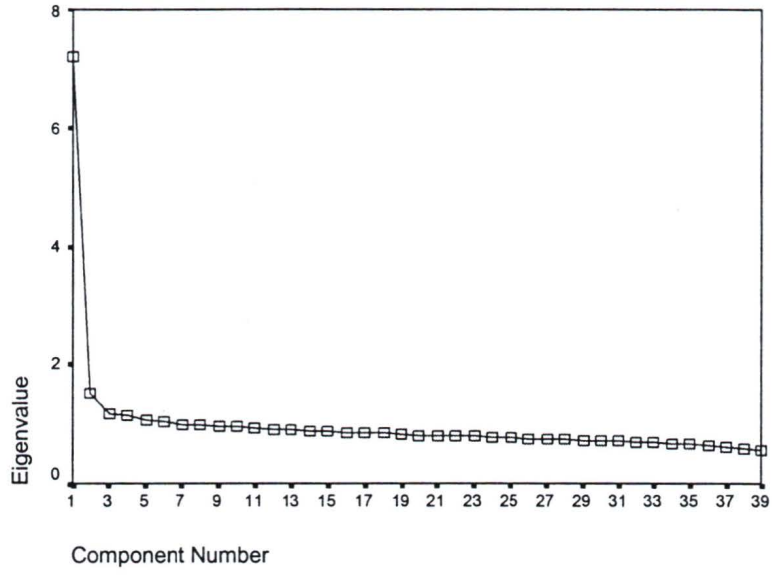


Figure 3. Scree plot of eigenvalues

As shown in Figure 3, there is evidence of a single dominant factor and the remaining factors are generally not distinguishable from one another. The percent of variance accounted for by factor one was 18.50% (see Table 4), which falls short of the 20% criterion for essential unidimensionality described by Reckase (1979). In reviewing the eigenvalues presented in Table 4, the eigenvalue for component one (7.213) was 4.7 times larger than the eigenvalue of component two (1.521), just short of the criterion for establishing clear dominance of the first component as suggested by Hambleton, Swaminathan & Rogers (1991).

Table 4

Component eigenvalues and percent of variance accounted for by each component with eigenvalues greater than one.

Component	Eigenvalues	% of Variance	Cumulative %
1	7.213	18.496	18.496
2	1.521	3.900	22.396
3	1.166	2.990	25.386
4	1.141	2.926	28.312
5	1.062	2.722	31.034
6	1.028	2.637	33.671

To explore the assessment's underlying structure, a rotated principal components analysis (PCA) solution was reviewed. The PCA produces a matrix of intercorrelations for all items on the same instrument. This matrix is then factored to determine whether item responses "cluster" together in some predictable pattern (Crocker & Algina, 1986) such as the subdomains on the assessment table of specifications.

Initially, a three-factor PCA was conducted on the reading comprehension assessment items to determine if the items clustered together according to the subdomains on the table of specifications (see Table 5). A six-factor solution was also conducted as there were six

factors with eigenvalues greater than one and a six-factor solution provided additional information as to how the items clustered.

Table 5

Assessment items referenced to subscale categories.

Subscale Category	Items
1. Identify and Interpret Key Concepts and Main Ideas (5 MC[1] items)	MC1, MC5, MC15, MC30, MC35
2. Locate, Interpret and Organize Details (26 items—25 MC; 1 OE[2])	MC2, MC3, MC4, MC6, MC7, MC8, MC10, MC11, MC12, MC13, MC16, MC17, MC18, MC20, MC21, MC22, MC23, MC25, MC26, MC27, MC32, MC33, MC34, MC36, MC37, OE39
3. Critical Analysis (8 items—5 MC; 3 OE)	MC14, MC24, MC28, MC31, MC38, OE9, OE19, OE29

[1]MC means multiple-choice
[2]OE means open-ended

Additionally, both oblique and varimax rotated solutions were conducted to determine if different information might be provided by these rotation techniques. It was found that the results of the oblique and the varimax solutions were very similar; however, the varimax solutions provided a clearer interpretation. Therefore, the results of the varimax-rotated solutions are presented in Table 7 and Table 8.

Only factor loadings greater than .30 were interpreted as factor loadings less than .30 are generally considered meaningless (Crocker & Algina, 1986). As depicted in Table 7, the items did not cluster together as anticipated and the three-factor solution provided little, or no, support

for the underlying structure as referenced in the reading assessment table of specifications (see Appendix C or Table 5). It was noted, however, that items five through eight, 10 through 14, and 15 through 18 each clustered together on various components. It was further noted that these groups of items were related to reading passages two, three, and four respectively.

The six-factor solution presented in Table 8 provided further evidence of how items clustered together. With the exception of a handful of items, the items clustered together based on the reading passage to which they were associated. For example, items one through four are associated with reading passage one on the assessment and items five through eight are associated with reading passage two. As Table 8 depicts, these items each clustered together. This general pattern was also noted for the multiple-choice items associated with reading passages three, four, six, and eight. For reference, the questions associated with each reading passage are presented in Table 6.

Table 6

Assessment items related to reading passages.

Reading Passage	Items
Passage #1	MC1, MC2, MC3, MC4
Passage #2	MC5, MC6, MC7, MC8, OE9
Passage #3	MC10, MC11, MC12, MC13, MC14
Passage #4	MC15, MC16, MC17, MC18, OE19
Passage #5	MC20, MC21, MC22, MC23, MC24
Passage #6	MC25, MC26, MC27, MC28, OE29
Passage #7	MC30, MC31, MC32, MC33, MC34
Passage #8	MC35, MC36, MC37, MC38, OE39

In the three-factor solution, the open-ended items loaded on the factor with the other items and the passage to which they were related. In the six-factor solution, 2 of the open-ended questions clustered on the reading passage to which they were associated. However, there seems to be something in common about these items, as in both cases they also clustered together on one factor (see Table 7 and Table 8).

In summary, the principal components analysis revealed no evidence of the theoretical structure underlying the assessment. Additionally, the results fell just short of the criteria proposed for determining test unidimensionality.

Table 7

Factor loadings on the three-factor varimax-rotated solution

	Item	Comp.1	Comp.2	Comp.3
Passage 1	MC1	**.474**	.189	.042
	MC2	.226	**.313**	-.047
	MC3	.214	**.348**	.041
	MC4	**.391**	.141	-.042
Passage 2	MC5	**.331**	.164	.038
	MC6	**.457**	.215	.066
	MC7	**.483**	.201	.113
	MC8	**.375**	.267	.134
Passage 3	MC10	.026	**.637**	.156
	MC11	.156	**.581**	.114
	MC12	.035	**.710**	.169
	MC13	.157	**.447**	.257
	MC14	.281	**.482**	.170
Passage 4	MC15	**.389**	.239	.131
	MC16	**.425**	.307	.224
	MC17	**.369**	.164	.057
	MC18	**.472**	.190	.024
Passage 5	MC20	**.306**	.051	.178
	MC21	**.333**	.082	.166
	MC22	**.388**	.044	.179
	MC23	-.069	.024	**.316**
	MC24	.229	.148	**.345**
Passage 6	MC25	.253	.097	**.265**
	MC26	.274	.162	**.345**
	MC27	.047	.173	**.498**
	MC28	.285	.258	**.441**
Passage 7	MC30	.074	.075	**.361**
	MC31	**.404**	.060	.205
	MC32	.242	-.061	**.385**
	MC33	.298	.004	**.401**
	MC34	.123	.094	**.511**
Passage 8	MC35	.167	.136	**.479**
	MC36	**.565**	.001	.197
	MC37	**.460**	-.062	.145
	MC38	**.417**	.093	.293
P2	OE9	**.441**	.155	.135
P4	OE19	**.647**	.110	.151
P6	OE29	**.488**	.118	.271
P8	OE39	**.450**	.080	.288

Table 8

Factor loadings on the six-factor varimax-rotated solution

	Item	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Passage 1	MC1	.34	.149	.045	**.381**	.077	-.030
	MC2	.005	-.000	.068	**.607**	.108	-.019
	MC3	.017	.054	.110	**.609**	.132	.061
	MC4	.272	.005	.020	**.381**	-.006	.064
Passage 2	MC5	.204	.122	.048	**.336**	-.008	.079
	MC6	.317	.247	.084	**.355**	-.020	.023
	MC7	.335	.254	.084	**.366**	-.038	.138
	MC8	.224	.233	.143	**.377**	.008	.135
Passage 3	MC10	.083	.038	**.706**	.051	.070	.089
	MC11	.200	.108	**.631**	.065	.017	.020
	MC12	.076	.100	**.758**	.081	.081	.036
	MC13	.046	.244	**.351**	.293	.173	.085
	MC14	.240	.209	**.442**	.204	.095	.006
Passage 4	MC15	**.394**	.096	.230	.121	.142	-.009
	MC16	**.371**	.231	.265	.208	.130	.078
	MC17	**.338**	.100	.122	.165	.062	-.021
	MC18	**.490**	.065	.180	.115	.072	-.089
Passage 5	MC20	.294	.164	.076	.040	.095	.180
	MC21	.294	**.301**	.069	.053	-.017	.051
	MC22	.334	**.408**	.033	.034	-.098	.071
	MC23	-.153	.062	.032	.125	.062	**.519**
	MC24	.148	**.414**	.099	.088	.148	.080
Passage 6	MC25	.167	**.510**	.047	.044	-.049	.012
	MC26	.174	**.509**	.091	.103	.121	.024
	MC27	-.064	**.557**	.106	.054	.237	.099
	MC28	.151	**.517**	.150	.204	.251	.064
Passage 7	MC30	.041	.209	.004	.046	**.485**	-.115
	MC31	**.363**	.232	.006	.115	.197	.041
	MC32	.227	.060	.038	.032	.044	**.656**
	MC33	.281	.106	.081	.055	.104	**.564**
	MC34	.119	.068	.094	.061	**.562**	.224
Passage 8	MC35	.180	.069	.132	.052	**.591**	.116
	MC36	**.569**	.185	.038	.050	.031	.184
	MC37	**.495**	.122	.012	-.034	-.035	.202
	MC38	**.415**	.135	.070	.097	.356	.012
P2	OE9	**.415**	.007	.106	.218	.235	.028
P4	OE19	**.647**	.105	.104	.154	.141	.061
P6	OE29	**.447**	.191	.073	.170	.263	.058
P8	OE39	**.476**	.035	.089	.086	.368	.086

Item Difficulty, Item Discrimination and Distractor Analysis

Because items and item responses are subunits of tests and test scores, validity at the item level is also important. Desirable item response patterns strengthen test score interpretations, whereas undesirable item response patterns weaken the validity of test score interpretations (Haladyna, 1999).

Because the "validity and reliability of any test depend ultimately on the characteristics of its items" (Anastasi & Urbina, 1997. p.172), the item difficulty values, the item discrimination index values, and the distractor analysis were reviewed. Specifically, the FSA item p-values and item discrimination indexes were compared with conventional "rules of thumb" for p-value ranges (.30 to .80), optimal average p-values (.63), and acceptable item discrimination values (.20 or greater) as outlined in the methods section. The item distractor analysis was reviewed for any distractors operating in an undesirable fashion (for example, no examinees selecting a particular distractor).

The multiple-choice item p-values are displayed in Figure 4 and Table 9. The average item difficulty, or p-value, across all multiple-choice items was .77. P-values ranged from .38 to .95. In comparison to our desirable range of p-values (.30 to .80), 20 of the 35 multiple-choice items fell within the desirable range and 15 did not. Because a large number of items had p-values over .80, it suggests that there may not have been enough difficult items. Figure 4 provides a histogram of the

item p-values. As depicted, there are more p-values in the higher range and the average p-value of .77 is much higher than the optimal average p-value of .63.

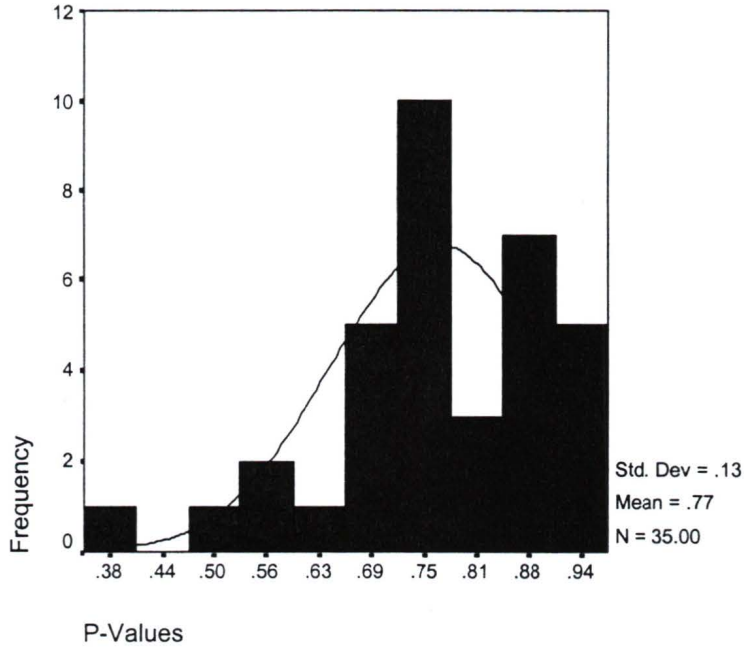


Figure 4. Histogram of multiple-choice p-values.

The point-biserial correlations revealed item discrimination coefficients in the range of .18 to .55 (see Figure 5 and Table 9). The average point-biserial was .41. Thirty-two items had point-biserial correlations over .30 and 22 of these items had point-biserial correlations over .40. Item discrimination indices also revealed no negative or zero values. Negative values would be an indication that the item is not functioning properly in that lower ability students would be selecting the correct choice more often than higher ability students. An item

discrimination value of zero would be an indication that the item is not discriminating at all between higher and lower ability students.

Figure 5 provides a histogram of the item discrimination values. As depicted, the item discrimination values are generally higher than .20 and averaged at .41, a very acceptable discrimination value. In relation to our criteria of desirable item discrimination values (.20 or greater), it can be stated that all multiple-choice items with one exception (MC23) had acceptable item discrimination coefficients; that is, 33 of 34 multiple-choice items displayed point-biserial correlation coefficients of .20 or greater.

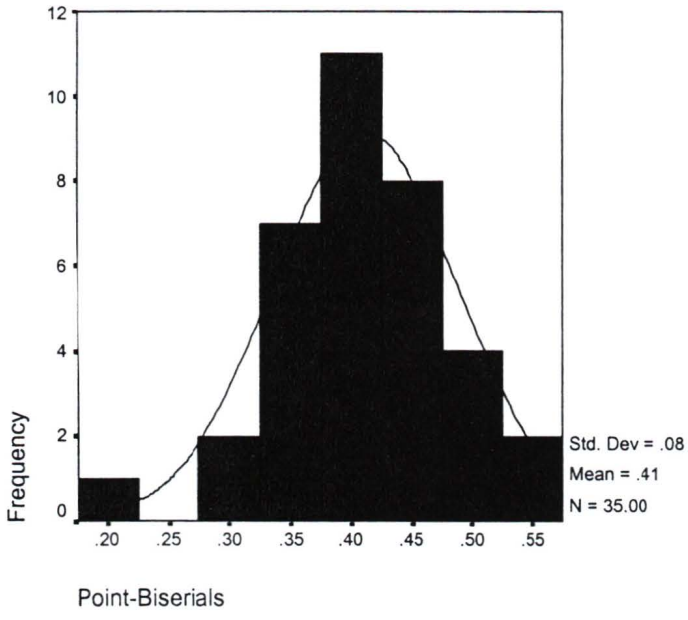


Figure 5. Histogram of multiple-choice point-biserial values.

Table 9

Classical item analysis for multiple-choice items.

Item	P-value	Point Biserial Correlation
MC1	.68	.47
MC2	.95	.28
MC3	.93	.33
MC4	.71	.36
MC5	.75	.36
MC6	.77	.47
MC7	.77	.51
MC8	.87	.46
MC10	.91	.37
MC11	.84	.42
MC12	.93	.40
MC13	.92	.42
MC14	.81	.50
MC15	.69	.47
MC16	.76	.55
MC17	.72	.40
MC18	.75	.45
MC20	.77	.36
MC21	.86	.37
MC22	.73	.41
MC23	.57	.18
MC24	.88	.40
MC25	.70	.38
MC26	.82	.44
MC27	.88	.35
MC28	.86	.52
MC30	.76	.29
MC31	.62	.45
MC32	.58	.38
MC33	.68	.44
MC34	.86	.38
MC35	.90	.40
MC36	.52	.53
MC37	.38	.41
MC38	.73	.49

The four open-ended questions also revealed estimated p-values (item average proportion scores) and point-biserial correlations in the acceptable range (see Table 10), that is, p-values were in the range of .30 to .80 and item discrimination values were each over .20. Item discrimination indices were calculated with ITEMAN version 3.5. P-values were estimated by taking the average number of points assigned to examinees and dividing by the maximum number of points available.

Table 10

Open-ended item characteristics.

Item	P-value	Discrimination
OE9	.77	.64
OE19	.42	.66
OE29	.48	.64
OE39	.62	.63

Finally, distractor analysis evaluates the effectiveness of the options on each multiple-choice question. With one exception, the multiple-choice question options revealed positive point-biserials on the correct option and negative point-biserials on each incorrect option (distractor). Multiple-choice question 23 was the only problematic item. This question had one distractor that revealed a positive point-biserial, that is; higher ability students were selecting this distractor more often than lower ability students. This item displayed a p-value of .57 but it

did not discriminate well (.18). It seems likely that this item is poorly worded, misleading, or has another possible answer.

Across all multiple-choice questions, at least 1% of the students chose one of the distractors; that is, there were no items where a distractor was not chosen at all. In all cases, the higher ability group chose the correct option more often than the lower ability group. With the exception of multiple-choice question 23, the low ability group endorsed each distractor more often than the higher ability group. See Appendix J for the results of the multiple-choice distractor analysis.

Taken as a whole, the results of the classical item analyses generally provided support for valid interpretations of reading comprehension assessment scores. While a number of items (15) were outside the range of desirable p-values, all items, with one exception, had acceptable discrimination values. With the exception of one distractor, the patterns of distractor performance were also as desired. In reviewing the range of p-values, it could be argued that the test was too easy and did not allow for higher ability students to demonstrate their true ability. However, as noted by the item discrimination values, the test is discriminating quite well between higher ability and lower ability students. In summary, the item responses are generally in support of valid test score interpretations for provincial-level reporting.

Item Response Theory

Item fit statistics support valid test score interpretations by determining whether any items do not fit within the underlying construct. The results of the PARSCALE item fit analysis are shown in Table 11. At noted, when sample sizes are large, the chi square (X^2) statistic tends to claim that many or all items are not fitting the model (Hambleton, Swaminathan, & Rogers, 1991). As can be seen in Table 11, all items were seen to have a significant lack of fit.

Because chi square is affected by sample size, an alternative index, calculated from the chi square statistic, is Cramer's V (M. Marshall, personal communication, March 2003; Rea & Parker, 1997). This statistic takes into account the large sample size and brings the test to more reasonable fit decisions.

As noted earlier, Cramer's V runs from 0 to 1. Because this is a measure of item mis-fit, the smaller the value, the better the item fit. A value under .20 indicates is an indication of item fit, whereas a value greater than .20 indicates item misfit. The results of the Cramer's V calculation are shown in the far right hand column of Table 11. Using the criterion of V greater than .20 to mean lack of fit, it is noted that there were no mis-fitting items.

This finding provides evidence related to internal structure in that no items appear to be departing from the model underlying the test.

Table 11

IRT item fit statistics.

ITEM	CHI-SQUARE	D.F.	PROB.	Cramer's V
0017	330.18967	56	0.000	0.08
0001	189.28271	58	0.000	0.06
0002	99.16830	46	0.000	0.05
0003	89.32883	47	0.000	0.04
0004	300.04880	59	0.000	0.08
0005	189.61470	59	0.000	0.06
0006	145.77716	54	0.000	0.06
0007	281.85782	53	0.000	0.08
0008	142.63858	49	0.000	0.06
0009	153.04584	50	0.000	0.06
0010	108.62302	53	0.000	0.05
0011	76.72837	42	0.001	0.04
0012	102.61121	43	0.000	0.05
0013	76.79718	52	0.014	0.04
0014	166.29440	58	0.000	0.06
0015	111.58828	52	0.000	0.05
0016	167.23401	59	0.000	0.06
0018	111.40630	58	0.000	0.05
0019	342.90292	53	0.000	0.09
0020	410.47006	58	0.000	0.09
0021	257.20703	59	0.000	0.07
0022	80.77551	50	0.004	0.04
0023	135.14108	59	0.000	0.05
0024	112.98576	53	0.000	0.05
0025	169.24095	53	0.000	0.06
0026	98.78428	46	0.000	0.05
0027	284.79111	59	0.000	0.08
0028	350.93750	59	0.000	0.09
0029	187.41502	59	0.000	0.06
0030	301.38348	58	0.000	0.08
0031	95.29511	53	0.000	0.05
0032	71.73806	48	0.015	0.04
0033	693.33105	58	0.000	0.12
0034	932.85461	59	0.000	0.14
0035	162.05833	55	0.000	0.06
0036	618.33545	216	0.000	0.06
0037	1865.45361	214	0.000	0.10
0038	1966.01489	219	0.000	0.10
0039	878.58362	222	0.000	0.07

Evidence Based on Relations to Other Variables

As noted earlier, scores of any particular test can be expected to correlate highly with scores of other tests that presumably measure the same thing. Conversely, test scores can be expected to have lower correlations with measures purported to measure different abilities or underlying constructs. If we find, for example, that test scores intended to measure the same construct do in fact correlate highly, then this provides some validity evidence that the both tests measure the construct of interest.

To provide evidence for discriminant validity, correlations between students' FSA reading, writing, and numeracy scores were examined. Because of the connection in student learning between the development of reading and writing skills (British Columbia Ministry of Education, 2000), it was expected that higher correlations between FSA reading and writing scores, versus FSA reading and numeracy or FSA writing and numeracy scores would be found. That is, students who do well in reading usually do well in writing, as these skills tend to co-develop. Additionally, because each of these assessments is purporting to measure different constructs (or skills), no correlations were expected to be overly high. This result would provide validity evidence that these tests are measuring different constructs.

Students' raw score percent on each of the reading, writing, and numeracy assessments were correlated. The results showed the highest

correlation (.67) between the students' reading and numeracy raw score percentages. Reading and writing scores showed a correlation of .42 and writing and numeracy scores showed a correlation of .37 (see Table 12). Figures 6, 7, and 8 represent these same correlations as scatterplots. As noted in Figure 7, the scatterplot depicts the strongest relationship between reading and numeracy scores.

Table 12

Correlation of reading, writing, and numeracy scores.

	Reading	Writing	Numeracy
Reading	1.000	.421	.670
Writing		1.000	.368
Numeracy			1.000

As noted by Allen & Yen (1979), "if the range of one or both of the scores involved in a correlation is restricted or reduced, the correlation will tend to be smaller than a similar correlation based on an unrestricted range of scores. This effect is called attenuation (reduction) due to restriction of range" (p.34). Because the FSA writing scores are restricted in the possible scores available (ranging from 3 to 12), the correlations involving the writing scores may be attenuated. The correlations involving writing were, therefore, adjusted by applying the following Spearman's correction for attenuation formula:

$$PT_XT_Y = pxy/\sqrt{(pxx)(pyy)}$$

where PT_XT_Y is the correlation between the true score for X and the true score for Y; pxy is the correlation of observed scores X and Y; and pxx and pyy are the reliabilities of X and Y (see Allen & Yen, 1979). Once applied, the reading/writing correlation increased to .51 from .42 and the numeracy/writing correlation increased to .44 from .37.

Regardless of the correction for attenuation, the relation of reading and writing scores did not result in the highest correlations as hypothesized. Reading and numeracy scores showed the highest correlation; reading and writing scores showed a moderate correlation; and, as expected, writing and numeracy scores showed the lowest correlation.

Because we are looking for evidence that the reading comprehension assessment does in fact measure the construct of reading, we expect low correlations with scores from tests designed to measure different constructs. Since we found low to moderate correlations, this provides some evidence of discriminant validity.

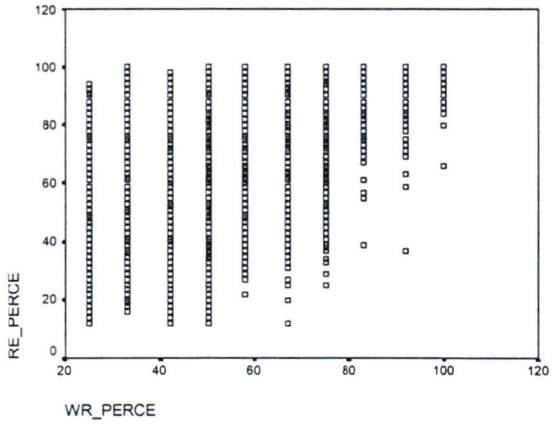


Figure 6. Scatterplot of students' reading and writing scores.

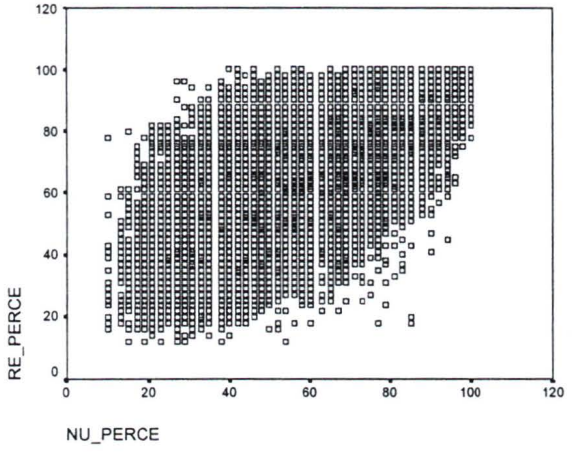


Figure 7. Scatterplot of students' reading and numeracy scores.

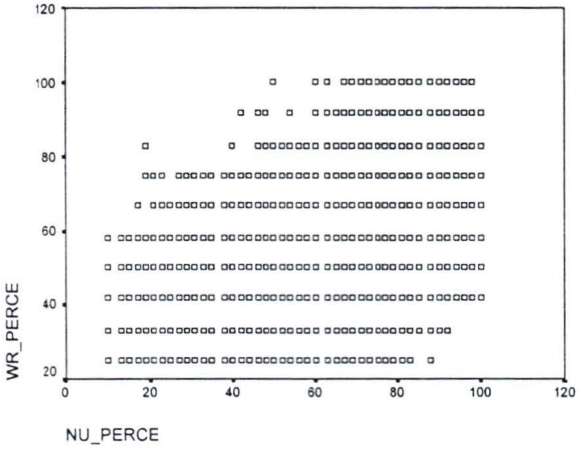


Figure 8. Scatterplot of students' writing and numeracy scores.

Finally, as a measure of the strength of the relationship, the correlation coefficients in Table 12 were squared. Reading and writing revealed a squared correlation coefficient of .18; reading and numeracy revealed a squared correlation coefficient of .45; and writing and numeracy revealed a squared correlation coefficient of .14.

With the exception of the reading/numeracy correlation, the squared correlation coefficients were found to be low. That is, little variance in one test score can be explained by the other test score. This represents some support for discriminant validity; that is, these assessments seem to be measuring different constructs. The moderately high squared correlation coefficient (.45) between the numeracy and reading raw score percentages is surprising. Because students are required to engage in reading math problems on provincial numeracy assessments, these assessments might be measuring reading skills as well as math skills. The results of this particular correlation may in fact be providing convergent validity evidence; that is, both the reading and numeracy assessments involve reading skills and, therefore, student scores on these instruments should be related. A study of the underlying cognitive processes students engage in as they respond to assessment items, coupled with convergent validity evidence, might reveal important information regarding the nature of these constructs.

CHAPTER 5: SUMMARY, DISCUSSION, AND CONCLUSION

Summary

A large-scale provincial assessment of student achievement has been in existence in BC since 1976. This assessment provides important accountability information to the public about how well students are learning important skills. In 1999, the assessment program was redesigned to focus on the foundation skills of reading, writing, and numeracy. Because the stakes are increasing, there is an increased responsibility by the Ministry of Education to ensure test scores are valid.

Prior to 1999, a technical report was produced following each BC provincial assessment administration. This report provided provincial-level validity-related information in the form of assessment development procedures and psychometric characteristics of the assessment instrument for that particular administration. Since 1999, this information has not been produced and is not readily available.

Standards and guidelines for testing state that test developers have an obligation to provide validity-related evidence for each test use and interpretation. The main authority on testing guidelines, the *Standards for Educational and Psychological Testing,* incorporates a unified view of validity and suggests that when evidence is gathered from a variety of sources, validity of test score interpretation is strengthened (AERA et al., 1999).

Based on the framework outlined in the *Standards for Educational and Psychological Testing* (1999), this study set out to gather validity-related evidence from a range of sources in order to judge the validity of the BC FSA test score interpretation at the provincial level. Specifically, by collecting a variety of evidence, this study judged the construct validity of the 2001 BC FSA Grade 4 reading comprehension assessment for its purpose of evaluating BC students' reading comprehension skills.

Discussion

A review and an interpretation of the key findings are presented below under the following headings: evidence based on test content, evidence based on internal structure, and evidence based on relations to other variables. Recommendations, limitations with the methods and procedures utilized in this study, and suggestions for further research are provided at the end of the section.

Evidence Based on Test Content

Content-related evidence of construct validity was gathered through documentation of the FSA test development process. The purpose of this analysis was not to judge the quality of each test development step, but to document the procedural steps that are in place. It was noted that the major assessment development steps include a definition of the reading comprehension construct, the preparation of a table of specifications, the

establishment of teacher committees, item training and item writing, a social considerations review, technical item reviews, field-testing, administration and security procedures, and marking reliability. The FSA assessment procedures identified are generally congruent with the important steps in the test development process as outlined by Haladyna (1999).

The Ministry of Education should be commended for having such an extensive test development process in place. However, this information should be made more available to the public and the nature of, and exact methods used, need to be documented and evaluated. In general, detailed information on each step of the test development process would be very useful to researchers and the general public and would support validity claims. Information resulting from field-testing and statistics related to test reliability are important and necessary for evaluating the validity of test score interpretation. Although a number of statistics (internal consistency reliability; inter-marker reliability) are generated each year by the Ministry of Education staff, they are not readily obtainable and should be made public.

The study of omissions on each multiple-choice question provides some evidence of content-irrelevant factors that may be affecting valid interpretation of student reading ability. It was found that the percent of omissions rises slightly at the end of each section of the assessment, suggesting fatigue may be a minor factor in valid test score

interpretation. The results were compared to results obtained in the 1998 reading assessment. The percent of omissions (under 3%) in the 2001 assessment was relatively low in comparison to the 1998 assessment.

While the percent of omissions was generally small (under 3%), there was an increase at the end of each section of the assessment, suggesting that fatigue could be interfering slightly with the measure of students' reading comprehension abilities.

Evidence Based on Internal Structure

Overall internal-consistency reliability was high (.85). However, reliability indexes on the assessment subscales did not meet the criterion set for an acceptable reliability index. These results lend support for reporting overall test scores; however, reporting student results on the subscales should be re-examined, particularly coupled with the findings from the principal components analysis.

A principal components analysis provided a measure of test unidimensionality and underlying test structure. While the scree plot demonstrated that there is evidence of a dominant factor, the percent of variance accounted for by factor one and the ratio of the factor one eigenvalues to the factor two eigenvalues fell slightly short of the criteria for test unidimensionality.

The underlying structure determined through a rotated factor solution also revealed an interesting finding. While it was anticipated that items would cluster together in relation to the way they are classified in the assessment table of specifications, it was found that the items clustered around the reading passages to which they were associated rather than the subdomains to which they were classified. This result also casts doubt regarding the fidelity of reading assessment subdomains. In fact, the reporting of overall reading scores (suggesting unidimensionality) and the reporting of results on subdomains (suggesting a multi-dimensional structure) appears somewhat contradictory from the outset.

Item responses are the basis of a test score. Finding that items operate properly in a psychometric sense ultimately supports valid score interpretations. Item analysis provides information about how well each item in the test functioned. Specifically, item analysis information provides information about the difficulty of each item, how well the item discriminated between high and low scorers on the test, and whether all of the alternatives (distractors) functioned as intended.

From this study, it was determined that all item p-values were above .30 and 24 of the 39 items (both multiple-choice and open-ended) had p-values that were in the range of .30 to .80. However, 15 multiple-choice items would be considered relatively easy as they had p-values over .80. In comparison to the optimum average p-value (.62), it was

noted that the average p-value on the multiple-choice questions was .77 and the average p-value on the open-ended questions was .64. The score distribution table in Appendix J displays the distribution of test scores on the multiple-choice items. This distribution also appears skewed, and shows that a majority of students received high total test scores on the multiple-choice component. These results indicate that the assessment may have been on the easy side, not challenging students of higher ability.

Thirty-four of the 35 items appropriately discriminated between more and less able students in terms of reading comprehension ability. With the one exception, discrimination indexes were above .20 and 32 items had item discrimination indexes over .30. As desirable, there were no negative discrimination indexes found. The average discrimination was .41, which is considered excellent by many researchers (Sapp, 2002; Crocker & Algina, 1986).

The results of the distractor analysis also showed that 34 of 35 multiple-choice items operated well. With the exception of multiple-choice question #23, the distractors had negative point-biserial correlations, meaning that the distractors were endorsed by the lower-ability students more often than the higher-ability students. In all cases, the correct choices had positive point-biserial correlations, that is, the higher ability group chose the correct option more often than the lower ability group. All distractors had at least 1% of students choosing the

option, meaning all distractors are plausible. Only one item (MC23) did not function properly. Although the item was relatively difficult (.57), the discrimination was poor (.18). This item had one positive discrimination value on one of the distractors, possibly meaning that the there was another viable option. The question should have been deleted from the overall test score or should be revised if used in the future.

Although there were 15 multiple-choice items with p-values over .80, they still had acceptable discrimination values. The item discrimination and distractor analysis results were very strong. In summary, the results of the item analysis were positive and generally support valid interpretations of these test scores.

One recommendation would be to increase the difficulty level of a number of the multiple-choice items. This will result in a wider range of total test scores and more reliable discriminations among students at various levels of achievement (Gronlund, 1993).

The results of the IRT item fit analysis based on the Cramer's V index revealed a positive finding. It was found that all items had a Cramer's V less than .20, an indication of item fit. This finding provides validity evidence related to internal structure in that no items appear to be departing from the model underlying the test.

Evidence Based on Relations to Other Variables

The correlational analysis revealed some evidence of discriminant validity between the constructs measured in the FSA. The squared correlation coefficient between reading and writing and writing and numeracy scores were low indicating these assessments are likely measures of different constructs. However, reading and numeracy scores displayed moderate squared correlations. Discriminant validity evidence does not ensure that the FSA is in fact measuring reading comprehension. Evidence of the FSA reading construct would be strengthened by demonstrating relationships to other measures of reading (convergent validity).

Recommendations

To support interpretation of results at a provincial level, the following recommendations are provided:

- Detailed documentation of important steps in the test development process is desirable and should be published.
- Reporting of subscale category mean scores on the reading comprehension assessment might be reconsidered.
- Information regarding what it means to "meet expectations" should be added to reports to aid interpretation.
- Confidence intervals could be reported along with the results for subpopulations (e.g., males and females).

- Pilot test and actual administration item analyses might be re-evaluated to improve detection and elimination of "faulty" items.
- Re-establishing pilot timing tests might be considered.
- Items with greater difficulty levels might be added to ensure a wider range of test scores and more reliable discriminations among students at various levels of achievement.
- Information such as, the item level analyses and the internal consistency and inter-marker reliability should be made available to consumers of the assessment results.
- A sustainable model for reporting validity-related information following each assessment administration should be established.

Limitations

A number of limitations to this study were observed by the author and are noted below.

While content-related validity evidence will often be evaluated through an examination of the test construction procedures, this study did not evaluate the quality of those procedures. For example, it was determined that field-testing occurs during the FSA development, but how do we know that the field-testing procedures are appropriate?

Goodness-of-fit analyses were somewhat unsatisfactory. Person fit statistics were not available with the software utilized in this study.

Additionally, item fit statistics had to be converted to another index for meaningful interpretation. Goodness-of-fit studies often place a great deal of reliance on statistical tests of model fit. Because these statistics are greatly affected by sample size, other procedures have been proposed (see Hambleton, Swaminathan, & Rogers, 1991).

The *Standards for Educational and Psychological Testing* (1999) provided a framework for collecting validity evidence. In this study, evidence based on response processes and test consequences was not addressed. Because validity arguments are strengthened when evidence from all categories is present (Gronlund & Linn, 1990), this study provides an incomplete picture of construct validity and also provides limited evidence for interpreting results at the school district, school or student levels.

Suggestions for further research

While this study provided some evidence of construct validity to support provincial-level test score interpretation, the information is far from complete. Construct validation is ongoing and there are arguably endless ways to enhance the information collected in this study. The following provides a number of additional ideas for collecting validity evidence to support the provincial-level interpretation of results.

To further evaluate the methods employed in this study, these analyses should be extended to the other FSA assessment areas (i.e.,

writing and numeracy), other grade levels (7 and 10), and other years of the assessment.

Further research in this area of content-related evidence might include involving independent content reviews of subject experts in the classification of items. This would provide evidence of the content domain and the item congruence with table of specifications. Further study might also include an evaluation of the quality of the test development procedures, such as the development of the conceptual framework. For example, there is an assumption that the FSA reading comprehension construct is made up of three categories—1) identifying and interpreting key concepts and main ideas, 2) locating, interpreting, and organizing details, and 3) critical analysis. This "construct" appears to be a carry-over from earlier PLAP assessments rather than one grounded in psychological theories of reading (J. Walsh, personal communication, 2003).

An area of construct validation might involve determining the underlying cognitive processes students are using as they answer specific items. "Think-aloud" procedures are often used to query students as they complete assessment items. As noted by the Canadian Psychological Association (1996), "detailed questioning of test takers regarding their performance strategies or responses to particular items, or the probing of raters regarding the reasons for their ratings, can yield hypotheses that enrich the definition of a construct" (p. 12). This type of

procedure would provide information about the construct in general as well as the items related to each of the reading subdomains.

Convergent validity studies should be conducted to complement the discriminant validity evidence gathered in this study. For example, convergent validity evidence could be gathered by investigating scores on tests measuring the same constructs. For instance, national and international assessments of reading are conducted in BC. It would be interesting to explore the relationships of students' scores across these measures of the same construct. Additionally, to aid in school and student level interpretations, studies of the correlations between students' classroom grades in reading and scores on the FSA reading assessments would provide further convergent validity evidence towards valid test score interpretation. This could be extended to determine if the FSA is identifying the right students who require intervention. That is, do students who do not meet expectations on the FSA also do poorly in others assessments of reading?

Differential Item Functioning (DIF) is a statistical technique that reveals systematic differences among groups on a test score or test item that are attributable to group membership instead of true differences in the construct being measured. The existence of this bias in item responses lowers the validity of test score interpretation and uses, because bias favours one group of test-takers over another. Because the FSA results are reported for separate groups of students (e.g., male &

female; Aboriginal & non-Aboriginal), a study of DIF would be important to conduct.

Finally, intended and unintended consequences of testing are typically not well investigated (Mehrens, 2002). Haertel (2002) suggests that a content analysis of the factual accuracy of media reports following the release of test results would reveal evidence of appropriate or inappropriate test score inferences. A study of how results are used would be especially important for interpretation of results at the district, school, and student levels.

Conclusion

This study set out to determine and evaluate validity-related evidence in support of assessment score interpretation at the provincial aggregate level. Overall, it was found that the evidence supporting provincial-level FSA score interpretation was mixed; however, the evidence was deemed to be more positive than negative. Table 13 provides a summary of the procedures employed and the evaluation of the evidence in support of valid score interpretation.

Test content-related evidence was generally strong. The FSA involves a systematic test development process, which ultimately supports validity claims. The evidence from the analysis of test development was, therefore, judged to be strong. The percent of omissions on each multiple-choice question was relatively low; however,

there was some limited evidence that fatigue may be polluting test score interpretation. Because the percent of omissions was relatively low, this evidence was considered moderate for provincial-level interpretation.

Table 13.
Validation procedures and rating of evidence.

Line of Evidence	Procedure	Rating of Evidence in Support of Valid Interpretations
Test Content	Description of test development	Strong
	Multiple-choice omissions	Moderate
Internal Structure	Unidimensionality	Moderate
	Test Structure	Weak
	Overall Reliability	Strong
	Subdomain Reliability	Weak
	Classical Item Analyses	Strong
	IRT item fit	Strong
Relations to Other Variables	Test score correlations	Moderate

Evidence based on internal structure was judged to be fairly strong. Item analyses, overall reliability results, and IRT infit statistics generally support valid interpretations. Only one item distractor was found to discriminate in an undesirable fashion. Overall test reliability was high and in the desirable range for valid interpretations. Additionally, the IRT goodness of fit statistics showed that there were no items misfitting the 2-parameter model.

However, reading assessment subdomain reporting is suspect due to low internal-consistency reliability values and lack of evidence of the

underlying structure as articulated in the table of specifications. This evidence was, therefore, judged to be weak. Also, evidence of test unidimensionality fell just slightly short of the criteria for essential unidimensionality and was therefore considered moderate as support for valid score interpretations.

The correlations with tests of other constructs (writing and numeracy) were not overly high, which seems to provide some evidence of discriminant validity. That is, these tests are generally measuring different constructs and therefore, this evidence was judged to be moderate.

In summary, the evidence related to test content, evidence related to internal structure, and evidence of relations to other variables was judged to be moderate to strong in support of valid test score interpretation at a provincial level. The weakest evidence related to the subdomain analysis. That is, items were not found to cluster together structurally and the internal-consistency reliability was low on these subdomains. One might conclude that the reading comprehension subdomains are not distinct from one another. That is, they may be inter-related and may not support subdomain reporting.

The *Standards for Educational and Psychological Testing* (1999) outline standards for validity, reliability, and test support documentation. Such documentation usually includes the assessment purpose, the table of specifications, item formats, scoring procedures,

and the test development process. Technical data, such as the psychometric indices of the items, reliability and validity evidence, and cut score information should be summarized. Appropriate and inappropriate uses and interpretations of test scores should also be included. "The objective of the documentation is to provide test users with the information needed to make sound judgments about the nature and quality of the test, the resulting scores, and the interpretations based on the test scores" (AERA at al, 1999, p.67). In the case of British Columbia Foundation Skills Assessment, all validity evidence, whether weak or strong, should be made available to the intended users of the assessment results.

Finally, it was determined that the methods employed in this study were generally adequate for providing a sampling of validity-related evidence in support of test score interpretation. While far from complete, this set of procedures might serve as a basis for monitoring the validity-based evidence from British Columbia's large-scale assessment program.

REFERENCES

Alberta Learning. (2003). *Achievement Subject Bulletins.* Available from, http://www.learning.gov.ab.ca/k_12/testing/achievement/bulletins

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory.* Monterey, CA: Brooks/Cole Publishing Company.

American Educational Research Association (AERA), American Psychological Association (APA), the National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

American Educational Research Association (AERA), American Psychological Association (APA), the National Council on Measurement in Education (NCME). (1985). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, NJ: Prentice-Hall Inc.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, 2nd Edition, Washington, DC: American Council on Education.

Angus, W. A. (2002). *Item types and validity/reliability in large-scale assessment: The Emperor isn't wearing any clothes.* Paper presented at the Education Symposium, UBC, September 2002.

Assessment Systems Corporation. (1998). *ITEMAN for Windows* (Version 3.6) [Computer software]. St. Paul, MN: Author.

Bateson, D. J. (1992). British Columbia assessment of science 1999: Technical report 1: Classical component. Victoria, British Columbia, Canada: British Columbia Ministry of Education and Ministry Responsible for Multiculturalism and Human Rights.

Bognar, C. J., Cassidy, W., & Clarke, P. (1997). *Social studies in British Columbia: Results of the 1996 provincial learning assessment: Technical report.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Bognar, C. J., Cassidy, W., Lewis, C., & Manley-Casimir, M. (1991). *Social studies in British Columbia: Technical report of the 1989 social studies assessment.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Bognar, C. J., Chapman, A., Jeroski, S., Tolsma, C., & Toutant, A. (1995). *The 1993 British Columbia communication skills assessment. Technical report I: The classical study.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

British Columbia Ministry of Education. (2003a). *FSA Brochure for Parents and Students.* Retrieved April 27, 2003, from http://www.bced.gov.bc.ca/assessment/fsa/brochure.htm

British Columbia Ministry of Education. (2003b). *2002-03 Accountability Contracts*. Available from, http://www.bced.gov.bc.ca/schools/sdinfo/acc_contracts/

British Columbia Ministry of Education. (2003c). *Provincial Student Assessment Program*. Available from, http://www.bced.gov.bc.ca/assessment/

British Columbia Ministry of Education. (2003d). *Item writer's manual.* Victoria, British Columbia, Canada: Author.

British Columbia Ministry of Education. (2002). *FSA test design and development*. Retrieved September 21, 2002, from, http://www.bced.gov.bc.ca/assessment/fsa/development.htm

British Columbia Ministry of Education. (2001a). *Interpreting and communicating British Columbia foundation skills assessment results*. Retrieved February 20, 2003, from, http://www.bced.gov.bc.ca/assessment/fsa/01interpret.pdf

British Columbia Ministry of Education. (2001b). *British Columbia foundation skills assessment 2001 highlights*. Available from, http://www.bced.gov.bc.ca/assessment/fsa/publications.htm

British Columbia Ministry of Education. (2001c). *Foundation skills assessment: Instructions for teachers/invigilators*. Available from, http://www.bced.gov.bc.ca/assessment/fsa/publications.htm

British Columbia Ministry of Education. (2000a). *Interpreting and communicating British Columbia foundation skills assessment results.* Victoria, British Columbia, Canada: Author.

British Columbia Ministry of Education. (2000b). *The primary program: A framework for teaching.* Victoria, British Columbia, Canada: Author.

California State Department of Education. (2000). *Alignment, validity, and reliability of the spring 2000 golden state examinations.* Retrieved September 20, 2002, from, http://www.cde.ca.gov/statetests/gse/admin/gsereliabilityrpt.pdf

Canadian Psychological Association. (1996). *Guidelines for educational and psychological testing.* Available at: http://www.cpa.ca/guide9.html

Council of Chief State School Officers (CCSSO). (1998). *Key State Education Policies on K-12 Education—Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance.* Washington, DC: CCSSO.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Holt, Rinehart and Winston, Inc.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Education Measurement* (pp. 443-507). Washington, D.C.: American Council on Education.

De Ayala, R. J., & Kelley, H. P. (1987). *MEC Item Analysis.* Austin, Texas: Measurement and Evaluation Center. Available at: http://www.utexas.edu/academic/mec

Dick, W., & Hagerty, N. (1971). *Topics in measurement: Reliability and validity.* New York: McGraw-Hill, Inc.

Education Quality and Accountability Office. (1999). *Ontario Provincial Report on Achievement.* Retrieved September 21, 2002, from, http://www.eqao.com/eqao/home_page/pdf_e/99/99P036e.pdf

Edgley, M. (1995). *The 1994 writing for specific audiences and purposes study: Technical report.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 1-15). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Erickson, G. (1992). *British Columbia assessment of science, 1991. Technical report II: Student performance component.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Gaskill, J. L. (1999). *British Columbia 1998 assessment of reading comprehension and first-draft writing: Technical report.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Goodman, D. P. (2001). *Assessing the feasibility of using item response theory to analyse British Columbia's foundation skills assessment 2000.* Unpublished master's thesis, University of Victoria, Victoria, British Columbia, Canada.

Gronlund, N. E. (1993). *How to make achievement tests and assessments.* Needham Heights, MA: Allyn and Bacon.

Gronlund, N. E., & Linn R. L. (1990). *Measurement and evaluation in teaching.* New York: Macmillan Publishing Company.

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21(1),* 16-22.

Haladyna, T. M. (2002a). Supporting documentation: Assuring more valid test score interpretations and uses. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 89-108). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Haladyna, T. M. (2002b). *Essentials of standardized achievement testing: Validity and accountability.* Boston: Allyn & Bacon.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications, Inc.

International Association for the Evaluation of Educational Achievement (IEA). (1999). *Technical standards for IEA studies.* Delft, Netherlands: Eburon Publishers.

Jeroski, S. (1989). *1988 British Columbia reading and written expression assessment: Technical report.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Joint Advisory Committee. (1993) *Principles for fair student assessment practices for education in Canada.* Edmonton, Alberta, Canada: Author.

Joint Committee on Standards and Educational Evaluation. (1981). Standards for evaluations of educational programs, projects, and materials. USA: McGraw-Hill Book Company.

Kendall, J. S., & Marzano, R. J. (1997). *Content knowledge: A compendium of standards and benchmarks for k-12 education.* Aurora, CO: Mid-continent Regional Educational Laboratory, Inc.

Linn, R. L. (2002). Validation of the uses and interpretations of results of state assessment and accountability systems. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 27-48). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan Publishing Company.

Marshall, M., Taylor, A., Brigden, S., Bateson, D., Cardwell, S., Deeter, B., & Martin, S. (1996). *The 1995 British Columbia assessment of mathematics and science: Technical report.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Mehrens, W. A. (2002). Consequences of Assessment. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 27-48). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-104). New York: American Council on Education and Macmillan Publishing Company.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Muraki, E., & Bock, R. D. (1998). *PARSCALE: Parameter Scaling of Rating Data (Version 3.2)* [Computer program]. Chicago, IL: Scientific Software, Inc.

Newfoundland and Labrador Department of Education. (1998). *Grade 9 Science Report.* Available from, http://www.gov.nf.ca/edu/dept/etc.htm#Publications

New York State Department of Education. *1999 New York state grade 4 mathematics statewide assessment technical report.* Retrieved September 20, 2002, from,

http://www.emsc.nysed.gov/ciai/testing/assesspubs/G4Ma99TchRpt.PDF

Nova Scotia Department of Education. (1999). *1998 intermediate science criterion-referenced test report.* Available from, http://www.edu.gov.nf.ca/etc/gr9scrp/master.htm

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

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.

Oosterhof, A. (1990). *Classroom applications of educational measurement.* Columbus, OH: Merrill Publishing Company.

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

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219-241). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Robitaille, D. F. (1991). *The 1990 British Columbia mathematics assessment technical report.* Victoria, British Columbia, Canada: British Columbia Ministry of Education.

Rudner, L. M. (1993). *Test evaluation.* ERIC/AE 12/93.

Ryan, K. (2002). Assessment validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21(1),* 7-15.

Sapp, M. (2002). Psychological and educational test scores: What are they? Springfield, IL: Charles C. Thomas, Publisher, Ltd.

Saskatchewan Learning. *Saskatchewan provincial learning assessment program.* Retrieved September 20, 2002, from, http://www.sasked.gov.sk.ca/k/pecs/ae/docs/plap/foundation.pdf

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45,* 83-117.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A rasch measurement perspective. *Journal of Applied Measurement, 2,* 281-311.

Suen, H. K. (1990). *Principles of Test Theories.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics. New York: HarperCollins College Publishers.

Texas Education Agency (2003). *2000-2001 Technical Digest.* Available from, http://www.tea.state.tx.us/student.assessment/resources/techdig/index.html

Walt, N. (2002). *Canadian perspectives on large-scale assessment programs.* Paper prepared for the Symposium on Large-Scale Testing, Victoria, British Columbia, Canada.

Walt, N. (1999). *International perspectives on large-scale assessment programs*: An Overview of Policies and Practices. Unpublished manuscript, University of Victoria, British Columbia, Canada.

Washington State Department of Education. (2001). *Washington Assessment of Student Learning: Grade 4, 2000, Technical Report.* Retrieved September 21, 2002, from, http://www.k12.wa.us/assessment/assessproginfo/subdocuments/techreport2000/2000TechRptGr4.pdf

Yukon Department of Education (2002). *Parent Guide to the Yukon Achievement Tests.* Retrieved September 21,2002, from, http://www.gov.yk.ca/depts/education/ess/assessment/yat_parent_guide.pdf

Appendix A

Summary of Provincial Assessments in British Columbia

Year	Program	Assessment Area	Technical Report
1976	PLAP[1]	English Language Arts	Yes
1977	PLAP	Mathematics, Social Studies, and Language Arts (Reading)	Yes
1978	PLAP	Science and Language Arts (Written Expression)	Yes
1979	PLAP	Physical Education	Yes
1980	PLAP	Language Arts (Reading) and Kindergarten Needs Assessment	Yes
1981	PLAP	Mathematics	Yes
1982	PLAP	Science	Yes
1983		N/A	
1984	PLAP	Language Arts (Reading)	Yes
1985	PLAP	Mathematics	Yes
1986	PLAP	Science	Yes
1987	PLAP	French Immersion Assessment	Yes
1988	PLAP	Language Arts (Reading and Writing)	Yes
1989	PLAP	Social Studies	Yes
1990	PLAP	Mathematics	Yes
1991	PLAP	Science	Yes
1992		N/A	
1993	PLAP	Communications (Reading & Writing)	Yes
1994	PLAP	Writing for Specific Audiences and Purposes	Yes
1995	PLAP	Mathematics and Science	Yes
1996	PLAP	Social Studies	Yes
1997		N/A	
1998	PLAP	Reading Comprehension & First-Draft Writing	Yes
1999	FSA[2]	Reading Comprehension, Writing, Numeracy	No
2000	FSA	Reading Comprehension, Writing, Numeracy	No
2001	FSA	Reading Comprehension, Writing, Numeracy	No
2002	FSA	Reading Comprehension, Writing, Numeracy	No
2003	FSA	Reading Comprehension, Writing, Numeracy	No

[1] Provincial Learning Assessment Program
[2] Foundation Skills Assessment

Appendix B

Technical Report Analysis—Psychometric Characteristics and Procedures Related to Validity

Year & Subject[1]	**Statistics Provided**	**Procedures Relevant to Validity**	**Validity Implicitly or Explicitly Mentioned**
1988 R/W	Mean, SD, and alpha for subsections and total	• Description of development process including review of the curriculum • Table of specifications • Item development/review panels • Development of marking scales and procedures • Field testing • Scoring and interpretation	Implicit through development process
1989 SS	Alpha for subsections and total	• Description of development process including review of the curriculum • Table of specifications • Item development/review panels • Development of marking scales and procedures • Field testing • Scoring and interpretation	Explicit—"content validity was ensured by the item development procedures"
1990 Ma	Mean, SD, alpha, and standard error for each form	• Description of development process including review of the curriculum • Table of specifications • Item development/review panels • Pilot testing--detailed • Interpretation (including IRT analyses)	Implicit through development process, although item development information somewhat limited
1991 Sc	CTT, IRT and think alouds at piloting phase; alpha by grade/form & item stats on final items; generalizability coefficients by goal area; standard error of mean; p values; factor analysis on attitude scales	• Description of development process including review of the curriculum • Table of specifications • Item development/review panels • Extensive piloting • Scoring and interpretation	Implicit through development/piloting process (e.g., think alouds)

[1] R/W=Reading & Writing; SS=Social Studies; Ma=Mathematics; Sc=Science; Cm=Communications; WSAP=Writing for Specific Audiences and Purposes

Appendix B continued

Year & Subject[1]	Statistics Provided	Procedures Relevant to Validity	Validity Implicitly or Explicitly Mentioned
1993 Cm	Alpha—inter-rater reliability for writing and reading	• Description of development process including review of the curriculum • Item development/review panels • Field testing • Development of marking scales and procedures	Explicit—"content validity was built into the development ..."
1994 WSAP	Alpha—inter-rater reliability for writing	• Description of development process • Item development/review panels • Field trial • Development of marking scales & procedures • Scoring and interpretation	Explicit—"content validity was built into the development ..."
1995 Ma & Sc	CTT item stats; Mean, SD, and alpha for each form; generalizability coefficients by reporting category; standard error of mean; p values	• Description of development process including review of the curriculum • Table of specifications • Item validation and selection • Pilot testing • Interpretation	Explicit in terms of the attitude scales—"the content validity of the scales was addressed...and judged to be appropriate."
1996 SS	Alpha for each form *(Note: very limited statistical information provided)*	• Description of development process including review of the curriculum • Table of specifications • Item development/item review* • Development of coding scales and procedures • Coding and interpretation *(Note: no reference to piloting)*	Explicit—"content validity of the instruments was established through a variety of methods."
1998 R/W	Alpha—inter-rater reliability and internal consistency	• Description of development process • Table of specifications • Item development/review panels • Pilot testing and associated statistics • Coding, standard setting, and interpretation	Explicit—There is a section on construct validity providing correlations between reading and writing

[1]R/W=Reading & Writing; SS=Social Studies; Ma=Mathematics; Sc=Science; Cm=Communications; WSAP=Writing for Specific Audiences and Purposes

Appendix C

Reading Comprehension Assessment Table of Specifications

FSA 2001 Reading Comprehension Specifications

	Reporting Category	Question Distribution			Total Questions (%)
		Literature (%)		Information (%)	
		Narrative	Poetry		
Grade 4	**Identify and Interpret Key Concepts and Main Ideas** *Students identify main ideas and make inferences about relationships between events and characters.*	4–8	3–6	12–20	19–34
	Locate, Interpret and Organize Details *Students locate and select relevant information to answer specific questions. They interpret simple figurative language, illustrations, graphic material and text features. They sequence ideas and events.*	15–22	6–8	15–25	36–56
	Critical Analysis *Students draw reasoned conclusions from reading selections and defend their conclusions rationally. They identify viewpoints and opinions, assess the plausibility of ideas in information (fact and opinion) and situations in literature (real and make-believe).*	3–8	0–6	7–10	10–24
Grade 7	**Identify and Interpret Key Concepts and Main Ideas** *Students identify main ideas and make inferences about relationships between events and characters.*	4–10	3–7	13–20	20–37
	Locate, Interpret and Organize Details *Students locate and select relevant information to answer specific questions, including examples of literary elements (e.g., plot, climax, conflict, theme and setting). They interpret simple figurative language, illustrations, graphic materials and text features. They sequence ideas and events.*	12–18	3–7	20–30	35–55
	Critical Analysis *Students draw reasoned conclusions from reading selections and defend their conclusions rationally. They identify viewpoints and opinions, stereotypes and propaganda, and express agreement and disagreement with ideas presented.*	3–10	2–8	7–10	12–28
Grade 10	**Identify and Interpret Key Concepts and Main Ideas** *Students identify main ideas, events and themes in reading selections. They make inferences and draw conclusions about relationships between events and characters and information presented in various text and graphic forms.*	6–10	5–8	13–20	24–38
	Locate, Interpret and Organize Details *Students locate and select relevant information to answer questions. They describe characters and their motivation by providing evidence from the text. They interpret simple figurative language, illustrations, graphic materials and text features. They sequence ideas and events.*	11–16	5–8	20–30	36–54
	Critical Analysis *Students draw reasoned conclusions from reading selections and defend their conclusions rationally. They identify authors' viewpoints and opinions, recognize bias and false reasoning, and assess the effectiveness of persuasive techniques.*	5–10	4–8	5–10	14–28

Appendix D

Datasets

2001 FSA Item Level Dataset

Variable	**Description**
mcresp1-mcrespXX	Student's response for each multiple-choice questions (A, B, C, C, or blank)
mcscor1-mcscorXX	Student's score for each multiple-choice question (0=question answered incorrectly; 1=question answered correctly)
oescor1-oescorXX	Student's score for each open-ended question from 0 to 4 (8=inappropriate response; 9=no response)
mctot	Student's multiple-choice score
oetot	Student's open-ended score
totscor	Student's total score (multiple-choice + open-ended)
analysis	An analysis code applied to each student's record to determine if student results should be included in further analyses (based on proportion of the assessment that was completed).

2001 FSA Student Summary Dataset

Variable	**Description**
prbase	Whether or not student is counted in province based numbers
nu_three	Student's three-point numeracy achievement score or reporting category (1=below expectations, 2=met expectations, 3=exceeded expectations)
nu_percent	Student's numeracy raw score percent
re_three	Student's three-point reading achievement score or reporting category (1=below expectations, 2=met expectations, 3=exceeded expectations)
re_percent	Student's reading raw score percent
wr_three	Student's three-point writing achievement score or reporting category (1=below expectations, 2=met expectations, 3=exceeded expectations)
wr_percent	Student's writing raw score percent
nu_participation	Student's numeracy participation code to determine whether to include in various levels of reporting. Code 5 is used for provincial reporting.
re_participation	Student's reading participation code to determine whether to include in various levels of reporting. Code 5 is used for provincial reporting.
wr_participation	Student's writing participation code to determine whether to include in various levels of reporting. Code 5 is used for provincial reporting.

Appendix E

Reading Comprehension Assessment
Conceptual Framework

Focus of the FSA

The Foundation Skills Assessment (FSA) is an annual 'paper and pencil' large-scale assessment. It includes multiple-choice and written-response questions. The FSA Reading Comprehension and Writing are based on the English Language Arts curriculum. While this curriculum addresses six aspects of Language Arts: reading, writing, speaking, listening, viewing, and representing, the FSA addresses only reading and writing. The other four aspects, speaking, listening, viewing, and representing can be better served through classroom assessment.

The FSA Reading Comprehension passages include the following types of text: literature (prose and poetry), and informational. The informational passages may contain discontinuous text (e.g. timetables, recipes) and material presented in visual or graphical formats (e.g. charts, maps, diagrams, schedules, numerical data, cartoons, web pages).

Definition of Reading

The FSA takes its definition of reading from the National Council of Teachers of English, (NCTE) 1997.

> Reading is the process of constructing meaning from a written text. It is an active process involving the constant interaction between the mind of the reader, the text, and the context.

The definition reflects numerous current theories, which define reading as a constructive, interpretive, and interactive process. Meaning is constructed in the interaction between reader and text in the context of a particular reading experience, and culturally and socially derived expectations. The reader brings a repertoire of skills, cognitive, and metacognitive strategies, dispositions, and background knowledge. Texts are broadly defined to include print, graphic, and digital forms. This definition of reading corresponds to that used in the Language Arts curriculum and the BC Performance Standards.

Appendix E Continued

Relationship Between the FSA, Performance Standards, and Curriculum

In British Columbia, the learning outcomes of the curriculum are presented in the form of Integrated Resource Packages (IRPs). Learning outcomes, or content standards, describe the knowledge, attitudes and skills students are expected to learn in each grade level. Performance Standards describe levels of achievement in key areas of learning. Performance Standards support teachers in making consistent and accurate judgments about how well students are performing in relation to the prescribed learning outcomes.

Both the Language Arts Curriculum and the Performance Standards focus on three aspects of reading: Strategies and Skills; Comprehension; Response and Analysis. The FSA addresses only two of these three domains: Comprehension and Response and Analysis. Strategies and Skills cannot be evaluated in an annual assessment such as the FSA.

The table below summarizes the relationship among the *BC Language Arts IRP,* the *BC Performance Standards: Reading* and the FSA Framework.

BC Language Arts IRP	***BC Performance Standards***	***FSA Framework***
Comprehend and Respond (Strategies and Skills)	Skills and Strategies	Not appropriate for large-scale testing
Comprehend and Respond (Comprehension)	Comprehension	1. Identify & Interpret Key Concepts & Main Ideas 2. Locate, Interpret and Organize Details
Comprehend and Respond (Engagement and Personal Response) Comprehend and Respond (Critical Analysis)	Response and Analysis	3. Critical Analysis

Appendix F

Social Considerations Criteria

Social Considerations	✔
1. Gender Role Portrayals of the Sexes Consider the portrayal of personal traits, circumstances, attitudes and actions of males and females with regard to: • Frequency of portrayal of each sex • Diversity in roles and relations for each sex • Recognition of the contributions, experiences and perspectives of girls and women as well as those of men Gender-sensitive language that is specific when appropriate and inclusive whenever possible	
2. Age Portrayal Consider how members of different age groups are presented as well as how society's treatment of people of various ages is portrayed. Analyze for accuracy and currency. Evaluate for bias.	
3. Cultural Diversity/Multiculturalism Historical and modern-day society are presented as culturally diverse. Diversity is presented positively/celebrated. Authentic portrayal is more important than mere representation. The totality of a culture is presented, not just the exotic. There is an absence of cultural stereotyping. (Is the material suitable for students with diverse backgrounds?)	
4. First Nations/Metis/Aboriginal Culture/Roles A balance and realistic portrayal of Aboriginal people and their culture is presented/avoids putting emphasis on traditional aspects of Aboriginal people to the exclusion of their contemporary realities. Highlights contributions of Aboriginal people. Note that many of the points presented in the Cultural Diversity section also apply here.	
5. Portrayal of Persons with Disabilities and Special Abilities Children and adults with disabilities or special abilities are portrayed in ways that highlight their capabilities and contributions to society. (Is the material suitable for students with disabilities and special abilities?)	
6. Socio-economic Reference Socio-economic references often appear as analogies and examples in problems or in situations depicted. Consider whether assumptions/bases/values and perspectives presented in the material represent authentic family and society structure (whether context is historical or modern day).	
7. Reference to Belief Systems Examine content for references to organized sets of doctrines or ideas e.g., philosophies, religions, and political ideologies. Attitudes toward a particular belief system emerge in the personal traits, circumstances, attitudes, and actions of adherents. Consider how groups or members of groups are identified e.g., by appearance, socio-economic status, and activities. Analyze content for accuracy and currency. Evaluate presentation for bias.	
8. Historical/Geographical Context Historical Context – Examine for: • Material that contains ideas/actions/language that might be considered *inappropriate or outdated in modern times.* Geographical Context – Examine for: • Material that contains ideas or content that reflects geographical bias (i.e., favors one geographical region over another/often evident in materials development in the United States or eastern Canada). Consider the extent to which bias affects the usefulness of material.	
9. Physical/Emotional Safety and Portrayal of Violence Activities portrayed in text and visuals should model safe practices, responsible actions and common sense. Where violence or other emotionally charged content appears, suitability of the material requires careful consideration about the purpose of portrayal of violence and its potential emotional impact on the student audience. • NOTE: Students in an assessment context are a "captive audience". Material with strong emotional content may better be presented in a setting in which students are engaged in discussion about the content and their response to it before and after presentation.	
10. Ethical Legal Issues Consider whether content of passage refers to/raises issues subject to debate on moral or legal grounds. Where such references/issues do appear comment on appropriateness (value/purpose) of the proposed passage for assessment purpose/considering student audience. Analyze content for accuracy and currency. Evaluate presentation for bias. Consider the emotional impact content might have on student audience. Prominent examples of current issues that could raise potential concern include: • Medial procedures. Drugs. Prostitution. Pornography. Nuclear weapons/energy. Evolution versus Creationism. Environment/Natural resources. Immigration. Child custody/adoption. Abandonment/abduction. Abuse verbal/physical/emotional. Hand guns/Other weapons. Judicial system/due process. Electronic communication/commerce.	

Appendix G

2001 Provincial Report Sample

Provincial Report

GRADE 4 - READING COMPREHENSION

Foundation Skills Assessment 2001
Detailed Provincial Report

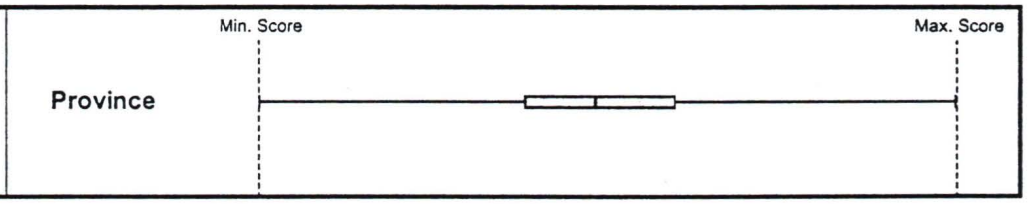
Number and Proportion of Students at Each Performance Level*

Not Yet Within Expectations | Meets Expectations | Exceeds Expectations

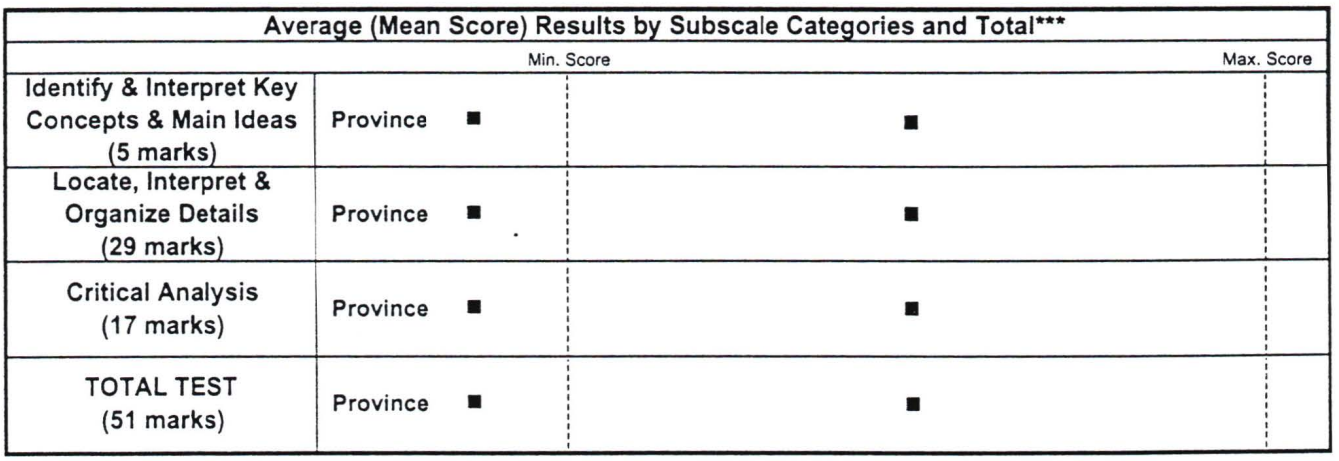
				Total Enrolment	Participated # and %	Excused # and %	% Meeting or Exceeding †
Province							
Number	10066	33175	2303	49472	46094	2055	78±0%
Proportion	22%	73%	5%		93%	4%	

† The proportion of students meeting or exceeding expectations falls within this range 9 times out of 10.

Overall Distribution of Results **



Average (Mean Score) Results by Subscale Categories and Total***



Discussions about the educational importance of these results should be based on additional information and guiding questions in *Interpreting and Communicating BC FSA Results 2001.*

*Below each bar graph are the numbers and proportions of students returning sufficient information to be placed within a performance level. Participating students were those who attempted or completed the component regardless of whether or not their responses were valid. See page 2 for further details.

**Each line and each box represents 25% of the students. See page 3 for further details.

***Results for each subscale category have been converted to a common scale. See page 3 for further details.

Appendix G Continued

PROVINCIAL REPORT

GRADE 4 : READING COMPREHENSION

Foundation Skills Assessment 2001
Results for Particular Groups of Students

Not Yet Within Expectations | Meets Expectations | Exceeds Expectations

	Total Enrolment	Participated # and %	Excused # and %	% Meeting or Exceeding
ALL STUDENTS				
Number*	10066	33175	2303	49472	46094	2055	78%
Proportion*	22%	73%	5%		93%	4%	
Distribution**							
	Min. Score		Max. Score				
MALE							
Number*	5712	16297	979	25373	23284	1283	75%
Proportion*	25%	71%	4%		92%	5%	
Distribution**							
	Min. Score		Max. Score				
FEMALE							
Number*	4354	16878	1324	24099	22810	772	81%
Proportion*	19%	75%	6%		95%	3%	
Distribution**							
	Min. Score		Max. Score				
ABORIGINAL/ FIRST NATIONS							
Number*	1447	1689	50	3893	3315	321	55%
Proportion*	45%	53%	2%		85%	8%	
Distribution**							
	Min. Score		Max. Score				
ENGLISH AS A SECOND LANGUAGE: CURRENTLY ENROLLED							
Number*	1706	3309	132	6173	5199	644	67%
Proportion*	33%	64%	3%		84%	10%	
Distribution**							
	Min. Score		Max. Score				
FRENCH IMMERSION							
Number*	311	1731	160	2260	2223	10	86%
Proportion*	14%	79%	7%		98%	0%	
Distribution**							
	Min. Score		Max. Score				

*Below each bar graph are the numbers and proportions of students returning sufficient information to be placed within a performance level. Participating students were those who attempted or completed the component regardless of whether or not their responses were valid. See page 2 for further details.

**Each line and each box represents 25% of the students. See page 3 for further details.

Appendix H

2001 School Report Sample

GRADE 4 - READING COMPREHENSION

Foundation Skills Assessment 2001
Detailed School Report

Number and Proportion of Students at Each Performance Level*

Not Yet Within Expectations | Meets Expectations | Exceeds Expectations

				Total Enrolment	Participated # and %	Excused # and %	% Meeting or Exceeding †	Compared to District ‡	Compared to Province ‡
School									
Number	15	24	2	44	44	0	63±6%	↓	↓
Proportion	37%	59%	5%		100%	0%			
District									
Number	303	997	70	1466	1387	50	78±1%	n/a	●
Proportion	22%	73%	5%		95%	3%			
Province									
Number	10066	33175	2303	49472	46094	2055	78±0%	n/a	n/a
Proportion	22%	73%	5%		93%	4%			

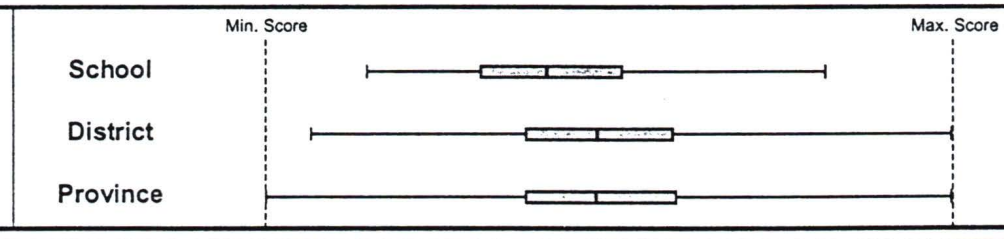
† The proportion of students meeting or exceeding expectations falls within this range 9 times out of 10.

‡ Comparisons to similar proportions in other schools in the district or province are represented using the following symbols:

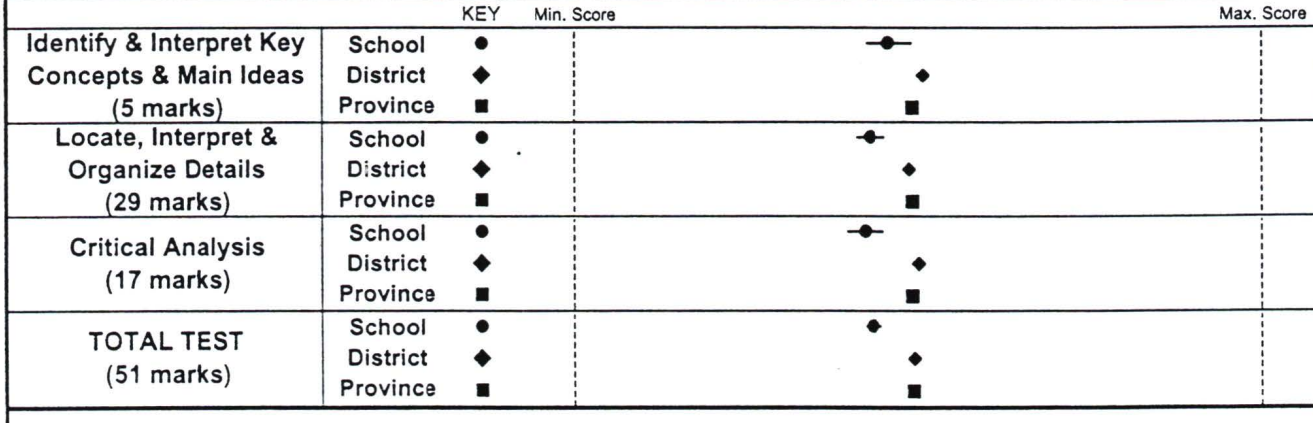
● - not significantly different from district or province ↑ - significantly higher than district or province

↓ - significantly lower than district or province n/a - not applicable

Overall Distribution of Results **



Average (Mean Score) Results by Subscale Categories and Total***



The average (mean score) performance of students in this school on the Total Test is statistically:

⇦ lower than the average performance of students in other schools in the district; and

⇦ lower than the average performance of all students in the province.

Discussions about the educational importance of these results for your school should be based on additional information and guiding questions in *Interpreting and Communicating BC FSA Results 2001.*

*Below each bar graph are the numbers and proportions of students returning sufficient information to be placed within a performance level. Participating students were those who attempted or completed the component regardless of whether or not their responses were valid. See page 2 for further details.

**Each line and each box represents 25% of the students. See page 3 for further details.

***Results for each subscale category have been converted to a common scale. See page 3 for further details.

Appendix H Continued

GRADE 4 : READING COMPREHENSION

Foundation Skills Assessment 2001
Results for Particular Groups of Students

Not Yet Within Expectations | Meets Expectations | Exceeds Expectations

	Total Enrolment	Participated # and %	Excused # and %	% Meeting or Exceeding
ALL STUDENTS				
Number* 15 / 24 / 2	44	44	0	63%
Proportion* 37% / 59% / 5%		100%	0%	
Distribution**				
Min. Score ... Max. Score				
MALE				
Number* 10 / 17 / 1	28	29	0	64%
Proportion* 36% / 61% / 4%		104%	0%	
Distribution**				
Min. Score ... Max. Score				
FEMALE				
Number* 5 / 7 / 1	16	15	0	62%
Proportion* 38% / 54% / 8%		94%	0%	
Distribution**				
Min. Score ... Max. Score				
ABORIGINAL/ FIRST NATIONS				
Number* 3 / 5 / 0	7	9	0	62%
Proportion* 38% / 63% / 0%		129%	0%	
Distribution**				
Min. Score ... Max. Score				

ENGLISH AS A SECOND LANGUAGE: CURRENTLY ENROLLED

There are fewer than five students in this particular group who participated in the assessment.

FRENCH IMMERSION

There are fewer than five students in this particular group who participated in the assessment.

*Below each bar graph are the numbers and proportions of students returning sufficient information to be placed within a performance level. Participating students were those who attempted or completed the component regardless of whether or not their responses were valid. See page 2 for further details.

**Each line and each box represents 25% of the students. See page 3 for further details.

Appendix I

2001 Student Report Sample

Individual Student Results

Grade

Foundation Skills Assessment 2001
INDIVIDUAL STUDENT REPORT

Personal Education Number :

2000/2001 School: _____
Student Name: _____ 2001/2002 School: _____

Reading Comprehension	This student's performance on the reading comprehension component fell within the "Exceeds Expectations" category.			
	Individual Results	**Not Yet Within Expectations**	**Meets Expectations**	**Exceeds Expectations**
				■
Writing	This student's performance on the writing component fell within the "Meets Expectations" category.			
	Individual Results	**Not Yet Within Expectations**	**Meets Expectations**	**Exceeds Expectations**
			■	
Numeracy	This student's performance on the numeracy component fell somewhere between the "Meets Expectations" and "Exceeds Expectations" categories.			
	Individual Results	**Not Yet Within Expectations**	**Meets Expectations**	**Exceeds Expectations**
			■	■

Interpreting Individual Student Results: A Guide for Students, Parents and Guardians

The report above summarizes this student's achievement on the BC Foundation Skills Assessment (FSA).

FSA is only one measure of student learning. It is intended to complement regular classroom assessment by teachers. *Results of this assessment should be interpreted carefully, considering everything else you know about the student's achievement and progress (including report card marks, classroom assignments, tests, quizzes, presentations, etc.).*

What is the Foundation Skills Assessment?

The Foundation Skills Assessment (FSA) is a set of three annual provincial tests. FSA measures selected reading comprehension, writing, and numeracy skills. Numeracy is the application of mathematics in daily activities (e.g., money and financial transactions, measurement, and statistics and probability). *These skills are important for the development of other reading, writing, and numeracy skills, and for future learning success.*

The skills tested are linked to the provincial curriculum. The learning outcomes in the provincial curriculum can be found in the Integrated Resource Packages, available in schools and at the following web site: www.bced.gov.bc.ca/irp

The tests were administered in spring 2001 and took about 4.5 hours to complete. With limited exceptions, the tests were taken by all students in Grades 4, 7, and 10 in BC.

A small number of students may have been excused from one or more tests, following provincial guidelines.

How will FSA results be used?

The main purpose of FSA is to help the province, school districts, and schools evaluate how well reading, writing and numeracy are being addressed and make plans for improvement.

A secondary purpose is to give teachers, students and families an additional, external source of information about a student's performance on important foundation skills. FSA results, together with other information collected by teachers, will help focus home-school discussions about how to improve student learning.

Note: FSA results do not count towards students' report card marks.

How were the tests structured?

The test for reading comprehension consisted of multiple-choice and written response questions based on a variety of reading selections.

The test for writing asked students to complete two different tasks.

The test for numeracy consisted of multiple-choice and open-ended questions.

Appendix J

Item Difficulty, Item Discrimination and Distractor Analysis

ITEMAN (tm) for 32-bit Windows, Version 3.6

Conventional Item and Test Analysis Program

Scale: 1
Type of Scale DICHOT
N of Items 35
N of Examinees 45387

Item Statistics					Alternative Statistics					
Seq. No.	Scale Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Endorsing Total	Low	High	Biser.	Point Key
1	MC1	.68	.52	.47	A	.03	.07	.00	-.24	
					B	.01	.02	.00	-.13	
					C	.68	.39	.92	.47	*
					D	.28	.52	.08	-.38	
					Other	.00	.00	.00	-.04	
2	MC2	.95	.12	.28	A	.01	.02	.00	-.12	
					B	.95	.88	.99	.28	*
					C	.01	.02	.00	-.12	
					D	.03	.08	.00	-.22	
					Other	.00	.00	.00	-.04	
3	MC3	.93	.15	.33	A	.03	.08	.00	-.20	
					B	.93	.84	.99	.33	*
					C	.02	.05	.00	-.20	
					D	.01	.03	.00	-.17	
					Other	.00	.00	.00	-.05	
4	MC4	.71	.40	.36	A	.10	.15	.03	-.15	
					B	.12	.21	.04	-.22	
					C	.07	.12	.02	-.17	
					D	.71	.50	.90	.36	*
					Other	.01	.00	.00	-.07	
5	MC5	.75	.36	.36	A	.17	.24	.07	-.17	
					B	.06	.15	.01	-.28	
					C	.02	.05	.00	-.16	
					D	.75	.55	.92	.36	*
					Other	.00	.00	.00	-.05	
6	MC6	.77	.46	.47	A	.05	.11	.01	-.22	
					B	.02	.06	.00	-.19	
					C	.15	.32	.03	-.34	
					D	.77	.51	.96	.47	*
					Other	.00	.00	.00	-.05	

Appendix J Continued

Item Statistics					Alternative Statistics					
Seq. No.	Scale Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Endorsing Total	Low	High	Biser.	Point Key
7	MC7	.77	.49	.51	A	.07	.18	.01	-.31	
					B	.11	.23	.01	-.28	
					C	.77	.48	.97	.51	*
					D	.05	.10	.01	-.20	
					Other	.00	.00	.00	-.07	
8	MC8	.86	.34	.46	A	.86	.65	.99	.46	*
					B	.03	.08	.00	-.22	
					C	.04	.10	.00	-.26	
					D	.06	.16	.00	-.27	
					Other	.00	.00	.00	-.06	
9	MC10	.91	.21	.37	A	.01	.02	.00	-.13	
					B	.01	.03	.00	-.14	
					C	.91	.77	.98	.37	*
					D	.05	.11	.01	-.24	
					Other	.03	.00	.00	-.19	
10	MC11	.84	.34	.42	A	.01	.04	.00	-.21	
					B	.11	.23	.02	-.28	
					C	.03	.08	.00	-.21	
					D	.84	.64	.97	.42	*
					Other	.00	.00	.00	-.09	
11	MC12	.93	.19	.40	A	.93	.81	1.00	.40	*
					B	.02	.05	.00	-.23	
					C	.01	.03	.00	-.17	
					D	.04	.10	.00	-.26	
					Other	.00	.00	.00	-.11	
12	MC13	.92	.21	.42	A	.92	.78	.99	.42	*
					B	.03	.07	.00	-.23	
					C	.03	.09	.00	-.27	
					D	.01	.04	.00	-.19	
					Other	.00	.00	.00	-.11	
13	MC14	.81	.43	.50	A	.14	.32	.02	-.37	
					B	.03	.08	.00	-.21	
					C	.01	.04	.00	-.17	
					D	.81	.54	.97	.50	*
					Other	.01	.00	.00	-.12	
14	MC15	.69	.51	.47	A	.04	.11	.01	-.24	
					B	.09	.20	.02	-.28	
					C	.69	.40	.91	.47	*
					D	.16	.25	.06	-.18	
					Other	.01	.00	.00	-.16	

Appendix J Continued

Item Statistics					Alternative Statistics				
Seq. No.	Scale Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Endorsing Total	Low	High	Point Biser.	Key
15	MC16	.76	.54	.55	A	.10	.24	.01	-.32	
					B	.76	.43	.97	.55	*
					C	.02	.06	.00	-.18	
					D	.10	.23	.01	-.31	
					Other	.01	.00	.00	-.16	
16	MC17	.72	.42	.40	A	.05	.08	.02	-.10	
					B	.72	.49	.91	.40	*
					C	.12	.24	.03	-.27	
					D	.10	.16	.04	-.16	
					Other	.02	.00	.00	-.17	
17	MC18	.75	.46	.45	A	.15	.24	.04	-.19	
					B	.03	.07	.00	-.21	
					C	.06	.15	.00	-.29	
					D	.75	.49	.95	.45	*
					Other	.02	.00	.00	-.18	
18	MC20	.77	.35	.36	A	.17	.31	.05	-.29	
					B	.03	.05	.01	-.13	
					C	.04	.06	.02	-.11	
					D	.77	.57	.92	.36	*
					Other	.00	.00	.00	-.02	
19	MC21	.86	.30	.37	A	.86	.68	.98	.37	*
					B	.06	.13	.01	-.24	
					C	.06	.14	.01	-.23	
					D	.02	.04	.00	-.12	
					Other	.00	.00	.00	-.04	
20	MC22	.73	.45	.41	A	.04	.08	.01	-.14	
					B	.73	.49	.94	.41	*
					C	.08	.14	.02	-.17	
					D	.14	.29	.02	-.30	
					Other	.00	.00	.00	-.05	
21	MC23	.57	.19	.18	A	.07	.15	.02	-.24	
					B	.57	.48	.67	.18	*
					C	.04	.08	.01	-.17	
					D	.31	.28	.30	.02	
					Other	.00	.00	.00	-.04	
22	MC24	.88	.28	.40	A	.02	.06	.00	-.18	
					B	.02	.05	.00	-.20	
					C	.88	.71	.98	.40	*
					D	.08	.18	.01	-.28	
					Other	.00	.00	.00	-.06	

Appendix J Continued

Item Statistics					Alternative Statistics					
Seq. No.	Scale Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Endorsing Total	Low	High	Biser.	Point Key
23	MC25	.70	.41	.38	A	.06	.12	.02	-.17	
					B	.70	.48	.89	.38	*
					C	.05	.11	.02	-.22	
					D	.18	.28	.08	-.21	
					Other	.00	.00	.00	-.08	
24	MC26	.82	.38	.44	A	.10	.20	.02	-.25	
					B	.02	.06	.00	-.22	
					C	.82	.59	.97	.44	*
					D	.05	.13	.01	-.26	
					Other	.00	.00	.00	-.08	
25	MC27	.88	.23	.35	A	.03	.09	.00	-.27	
					B	.88	.74	.97	.35	*
					C	.05	.09	.01	-.17	
					D	.04	.06	.02	-.13	
					Other	.00	.00	.00	-.09	
26	MC28	.86	.38	.52	A	.03	.10	.00	-.27	
					B	.06	.15	.01	-.28	
					C	.05	.14	.00	-.29	
					D	.86	.61	.99	.52	*
					Other	.00	.00	.00	-.10	
27	MC30	.76	.27	.29	A	.76	.61	.88	.29	*
					B	.02	.05	.01	-.16	
					C	.04	.11	.01	-.24	
					D	.16	.21	.11	-.11	
					Other	.01	.00	.00	-.13	
28	MC31	.62	.54	.45	A	.12	.19	.04	-.17	
					B	.06	.13	.01	-.23	
					C	.62	.34	.88	.45	*
					D	.19	.31	.06	-.24	
					Other	.01	.00	.00	-.14	
29	MC32	.58	.45	.38	A	.12	.12	.10	-.02	
					B	.07	.18	.01	-.29	
					C	.58	.35	.80	.38	*
					D	.22	.33	.10	-.21	
					Other	.01	.00	.00	-.15	
30	MC33	.68	.50	.44	A	.06	.12	.01	-.21	
					B	.68	.41	.91	.44	*
					C	.12	.23	.03	-.25	
					D	.13	.20	.05	-.17	
					Other	.01	.00	.00	-.15	

Appendix J Continued

Item Statistics					Alternative Statistics					
Seq. No.	Scale Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Endorsing Total	Low	High	Biser.	Point Key
31	MC34	.86	.28	.38	A	.10	.19	.03	-.25	
					B	.02	.05	.00	-.17	
					C	.01	.04	.00	-.17	
					D	.86	.69	.97	.38	*
					Other	.01	.00	.00	-.15	
32	MC35	.90	.24	.40	A	.03	.08	.00	-.24	
					B	.90	.74	.99	.40	*
					C	.03	.06	.01	-.17	
					D	.02	.06	.00	-.18	
					Other	.02	.00	.00	-.19	
33	MC36	.52	.68	.53	A	.04	.09	.01	-.19	
					B	.52	.18	.86	.53	*
					C	.12	.23	.03	-.24	
					D	.29	.44	.10	-.26	
					Other	.02	.00	.00	-.19	
34	MC37	.38	.53	.41	A	.30	.41	.16	-.20	
					B	.09	.13	.05	-.13	
					C	.38	.15	.68	.41	*
					D	.21	.25	.10	-.11	
					Other	.02	.00	.00	-.18	
35	MC38	.73	.51	.49	A	.14	.29	.03	-.30	
					B	.06	.12	.02	-.18	
					C	.04	.09	.01	-.20	
					D	.73	.43	.94	.49	*
					Other	.03	.00	.00	-.19	

Appendix J Continued

	Item Statistics				Alternative Statistics			
Seq. No.	Scale Item	Item Mean	Item Var.	Item-Scale Correlation	N per Item	Alter-native	Proportion Endorsing	Key
36	OE1	3.397	0.918	.64	41118	1	.05	+
						2	.20	
						3	.07	
						4	.69	
						Other	.08	
37	OE2	2.342	1.149	.66	32492	1	.26	+
						2	.34	
						3	.20	
						4	.20	
						Other	.36	
38	OE3	2.524	1.213	.64	34547	1	.17	+
						2	.45	
						3	.07	
						4	.31	
						Other	.28	
39	OE4	2.883	0.998	.63	38982	1	.12	+
						2	.21	
						3	.34	
						4	.33	
						Other	.14	

Scale:	1	2
N of Items	35	4
N of Examinees	45387	44318
Mean	26.841	2.767
Variance	33.025	0.498
Std. Dev.	5.747	0.706
Skew	-0.917	-0.345
Kurtosis	0.378	-0.280
Minimum	6.000	1.000
Maximum	35.000	4.000
Median	28.000	2.750
Alpha	0.852	0.618
SEM	2.212	0.436
Mean P	0.767	N/A
Mean Item-Tot.	0.412	0.644
Mean Biserial	0.604	N/A
Max Score (Low)	24	N/A
N (Low Group)	13196	N/A
Min Score (High)	31	N/A
N (High Group)	14341	N/A

Appendix J Continued

SCALE # 1 Score Distribution Table

Number Correct	Freq-uency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
5	0	0	1	0	+
6	33	33	1	0	|
7	43	76	1	0	|
8	84	160	1	0	|
9	131	291	1	0	|
10	190	481	1	0	+
11	237	718	2	1	|#
12	292	1010	2	1	|#
13	390	1400	3	1	|#
14	418	1818	4	1	|#
15	546	2364	5	1	+#
16	618	2982	7	1	|#
17	715	3697	8	2	|##
18	866	4563	10	2	|##
19	986	5549	12	2	|##
20	1146	6695	15	3	+###
21	1354	8049	18	3	|###
22	1508	9557	21	3	|###
23	1708	11265	25	4	|####
24	1931	13196	29	4	|####
25	2319	15515	34	5	+#####
26	2534	18049	40	6	|######
27	2817	20866	46	6	|######
28	3082	23948	53	7	|#######
29	3299	27247	60	7	|#######
30	3799	31046	68	8	+########
31	3837	34883	77	8	|########
32	3694	38577	85	8	|########
33	3348	41925	92	7	|#######
34	2427	44352	98	5	|#####
35	1035	45387	99	2	+##

```
|----+----+----+----+----+
     5   10   15   20   25
Percentage of Examinees
```

Elapsed Time: 6.760 seconds

VITA

Surname: Walt Given Names: Nancy Jane

Place of Birth: Victoria, British Columbia, Canada

Education Institutions Attended:

University of Victoria	1999 to 2003
University of Victoria	1980 to 1984

Degrees Awarded:

B.A.	University of Victoria	1984

Title of Thesis/Dissertation:

Gathering Evidence for Construct Validity: A Case for Large-Scale Educational Assessments

Author 

Nancy Jane Walt

August 4, 2003

CS STATISTICS - WEEKLY SHEET FOR
Special Cataloguing, Authority Work and General Maintenance

Name: [redacted]

FY2003-2004 Month: July Week of: July 26 - 30

Type of Transaction	Week Day									
	Monday		*Tuesday*		*Wednesday*		*Thursday*		*Friday*	
	titles	vols	titles	vols	titles	vols	titles	vols	titles	vols
Uvic theses catalogued (per CP18)			3	5			1	2		
microfiche copy			1							
M.Ed projects catalogued (per CP19)										
Working papers catalogued (per CP16)										
Special videos catalogued (per CP15)										
Volumes added (print / cd / videos)										
Reels added										
Fiche added										
Copies added (all locations)										
Discards/Deletes			6	6	1	1				
Re-instates										
Replacements										
Location change/transfers (all)										
Location change/transfers by location										
Total transfers from Main										
Total transfers to Main										
Total transfers from BCS										
Total transfers to BCS										
Total transfers. from Ref										
Total transfers to Ref										
Total transfers from SC										
Total transfers to SC										
Total transfer. from Curr										
Total transfers to Curr										
Transfers from Reading Rooms										
Re-labeling			9				10			
Reclassification only										
Recataloguing										
http update only (field 856)										
Authority records created										
Authority records updated										

VAC transfers/REF 91/103 AG OCLC analyzed serials COF for "The" AG OCLC AG OCLC