

Running Head: A THREE-STAGE APPROACH FOR IDENTIFYING GENDER  
DIFFERENCES

A Three-Stage Approach for Identifying Gender Differences

on Large-Scale Science Assessments

Rebecca J. Gokiert

Jacqueline P. Leighton

Centre for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta, Canada

## Abstract

Recent study into students' performance on large-scale tests of academic achievement has revealed some tests of achievement, including those in science, as multidimensional (e.g., Ayala, Shavelson, Yin, & Shultz, 2002; Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow 1995; Leighton, Gokiert, & Cui, 2005; Nussbaum, Hamilton, & Snow 1997). According to the *Standards for Educational and Psychological Assessment* group differences can be attributed to the existence of multiple dimensions on a test (AERA, APA, & NCME, 1999). The following research employs a large-scale science assessment to illustrate the utility of a three-stage approach for investigating gender differences in science achievement. The three stages include analysis of dimensionality, differential item functioning (DIF), and think-aloud interviews. Preliminary results indicate that one of the tests examined displayed multidimensionality and two dimensions best describe the test. Furthermore, systematic gender differences were found within each of the dimensions found, where dimension one systematically favoured males and dimension 2 systematically favoured males. Preliminary themes extracted from interview data collected from grade 8 students solving 12 items that displayed large DIF provide some understanding of why the DIF is occurring and whether it is due to bias or impact.

## A Three-Stage Approach for Identifying Gender Differences on Large-Scale Science Assessments

Large-scale assessment has become a national and international method for monitoring student achievement and for ensuring that educational systems are working (Alberta Education, 2005; Hamilton, Stecher, & Klein, 2002; McGehee & Griffith, 2001). As society and governments place more emphasis on large-scale testing, it has become increasingly important to examine the quality of large-scale testing programs and, in particular, the validity of inferences drawn from large-scale tests. Validity, as defined by the *Standards for Educational and Psychological Assessment* (AERA, APA, & NCME, 1999, p. 9), is “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests.” Recognizing the importance of test validation, Haladyna and Downing (2004) argue that defining the construct, which is an explicit knowledge and understanding of the latent trait(s) or knowledge and skills being measured on the test, is the first step to be taken in order for appropriate test score interpretation to occur.

Construct validation is the process through which the suitability of interpreting test scores is examined in the context of the latent trait(s) or knowledge and skills measured by the test. It is often the case that developers of large-scale assessment tools fail to verify the types of skills that are measured by tests and fail to provide guidelines for how strengths and weaknesses in student performance should be interpreted (Messick, 1994; NRC, 2001). When developers of large-scale assessments fail to explicitly state and incorporate the skills that are measured by these tests, test score interpretation becomes problematic. In order to address concerns surrounding ill-defined constructs

researchers have begun to investigate empirically the underlying knowledge and skills measured by tests (NRC, 2001).

In the domain of science, achievement is often characterized by a number of skills, such as quantitative reasoning, scientific reasoning, and spatial-mechanical reasoning (e.g., Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Nussbaum, Hamilton, & Snow, 1997). If these dimensions represent distinct knowledge, skills, and attributes in scientific achievement, it is important that tests capture these domains and that test scores reflect an individual's performance across these areas. Ensuring the subject domain to be measured by the tests is in fact measured will yield student scores that can be validly interpreted in terms of the student's strengths and weaknesses in the different areas of science achievement.

The purpose of this study, therefore, was to examine the utility of three approaches-dimensionality, differential item functioning, and protocol analysis-in collecting evidence about the underlying knowledge and skills measured by the School Achievement Indicators Program (SAIP) Science Assessment administered in 2004 (Council of Ministers of Education, Canada [CMEC], 2000). These approaches were used to (1) determine the dimensional structure of the 2004 administration of the SAIP, (2) determine if the performance of male and female students differs systematically on the SAIP items, and (3) determine if interview data of male and female students can aid in the generation of hypotheses about the underlying knowledge and skills measured by the test.

*Dimensionality*

Much of the research on the construct validation of large-scale science assessments has focused on test dimensionality (Hamilton, et al., 1995; Nussbaum, et al., 1997). Test dimensionality is defined as the smallest number of “dimensions or statistical abilities required to fully describe all test-related differences among the examinees in the population” (Tate, 2002 p. 184). Knowledge of the latent dimensional structure can also provide more meaningful information about test scores, and can ultimately enhance the validity of the inferences made from the test scores (Ayala, Shavelson, Yin, & Shultz, 2002; Childs & Oppler, 2000; Frenette & Bertrand, 2000; Hamilton et al., 1995; Nussbaum et al., 1997). Dimensionality research can help answer questions that address how many latent traits are being measured by a test overall and whether reporting student performance with a single score is reasonable given the number of latent traits found to underlie the test.

Study into the complex nature of students’ cognitive skills and its interaction with measures of achievement has revealed that some tests of achievement, specifically in science, are multidimensional (e.g., Ayala et al., 2002; Hamilton, et al., 1995; Leighton, Gokiert, & Cui, 2005; Nussbaum, et al., 1997). Richard E. Snow and his colleagues established the multidimensional nature of science achievement using the NELS: 88 and later a compilation of NELS: 88, TIMSS, and NAEP items (Ayala, et al., 2002; Hamilton et al., 1995; Nussbaum et al., 1997). The dimensional structure that emerged after subjecting the NELS: 88 science test samples for the 8<sup>th</sup> and 10<sup>th</sup> grades to a full information factor analysis were four and three factors, respectively. The underlying knowledge and skills measured by the NELS: 88 included dimensions such as, spatial-

mechanical reasoning, basic knowledge and reasoning, chemistry knowledge, everyday science knowledge, and reasoning with knowledge. In a study examining the dimensional structure of the SAIP 1999 Science achievement test both traditional and contemporary tests of dimensionality were used (Leighton, et al., 2005). Results, using both factor analytic and nonparametric techniques, indicated that the SAIP science assessment is multidimensional. For grade 8 and grade 11 samples, between two to four factors were found to underlie the data.

Although the majority of dimensionality studies examining large-scale science assessments are exploratory in nature, results from exploratory analyses can be used as a data-driven method, both to investigate whether a science assessment is measuring a multidimensional construct, and to guide the development of hypotheses about scientific reasoning. When conducting confirmatory analyses, test specifications can act as a springboard for examining the dimensional structure of science content and skills. However, test specifications do not capture subtle psychological processes and therefore, may not fit the data well in a confirmatory paradigm (Leighton et al., 2005; Norris, Leighton, & Phillips, 2004). That there is a lack of fit between test specifications and the data in the form of student responses in some cases is not surprising given that test specifications do not necessarily represent the cognitive processes students use to respond to test items (Norris et al., 2004). As a result, there has been a push towards the use of cognitive models to better guide large-scale achievement assessment development (Embretson, 1999; Haladyna & Downing, 2004; NRC, 2001; Norris et al., 2004; Snow & Lohman, 1989). The National Research Council (2001) suggests that more meaningful inferences could be made about student knowledge and skills if they were tied to explicit

theories of cognition and learning. Theories of scientific reasoning exist; however, these theories are conceptual in nature, have not been used in test development, and are rarely applied to real test data. The majority of dimensionality studies examine test data *after* the test has been administered, and attempts to match test score interpretation to existing theories of scientific reasoning occur *after* the data have been collected (Leighton, Gierl, & Hunka, 2004; Leighton et al., 2005; NRC, 2001). Typically, a theory would be expected to guide test development and then be used to interpret test scores. However, often tests are designed without a theoretical model in mind (Lane, 2004; Leighton et al., 2005). Retrofitting data to existing theories of scientific reasoning, although well intentioned, may result in hypotheses about test score interpretation at best.

#### *Performance Differences*

When attempting to describe the dimensional structure of a test it is also important to identify and understand implicitly the dimensional structure of *individual items*. Roussos and Stout (1996) define *dimension* of an item as “any substantive characteristic of an item that can affect the probability of a correct response on the item” (p. 356). When an item is found to measure multiple dimensions, this will result in differential item functioning (DIF) for groups of students. DIF occurs when two groups of examinees with equal ability as indicated by observed test performance do not have the same probability of answering the item correctly. It has been suggested that items that display DIF measure a primary dimension (the dimension the item is intended to measure) along with at least one secondary dimension, which was not intended to be measured by the item (e.g., Messick, 1989; Roussos & Stout, 1996). The secondary dimension(s) that is measured by the item can be representative of the construct or

irrelevant to the construct being measured. Construct-irrelevant variance has been described as a form of systematic error that can affect the probability of an examinee answering an item correctly. When a construct-irrelevant (or nuisance) dimension is present on a test, examinees with lower ability on the nuisance dimension will likely score lower on the test than other examinees who are of equal ability on the dimension of interest, but who have higher ability on the nuisance dimension.

Once items have been identified as statistically significant for differential item functioning (DIF), the next step is to determine whether this difference is due to bias or impact. The *Standards* (1999) prescribe that tests must be free from bias to be considered fair. Willingham and Cole (1997) suggest that “fair test design should provide examinees comparable opportunity, as far as possible, to demonstrate knowledge and skills they have acquired that are relevant to the purpose of the test” (p. 10). Bias, in the context of assessment, occurs when items on a test systematically advantage or disadvantage one group over another even when the groups have the same ability. This bias may result in the inconsistent selection and classification of students, which can have potential consequences if the nature of the selection and classification is high stakes (Moss, 1998). Conversely, an item displays impact if the identified DIF is due to genuine knowledge or experience differences or both.

There are many available procedures for identifying items that differentially function for subgroups; however, the capacity to determine whether the DIF is due to bias or impact is still under-developed (e.g., Camilli & Shepard, 1994; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Rogers, & Klinger, 1999). DIF analyses are a routine part of large-scale assessment testing programs; however, less common are studies to

understand the potential sources of DIF (Gierl et al., 2001). Considering the adverse impact that multiple dimensions in test items can have on the measurement of the desired construct, the validity of inferences that are drawn about student performance needs to be systematically examined using multiple methods. Sources of DIF in large-scale assessments have been explored; these include differences in item format (multiple choice vs. open-ended/constructed response), gender, translation, culture, and background experience (e.g., Ercikan, Law, Arim, Domene, Lacroix, & Gagnon, 2004; Gierl, et al., 1999; Henderson, 1999).

*Gender differences in science assessment.* Gender differences in large-scale assessment are considered by many to be the most carefully examined aspect of test fairness (Ryan & DeMark, 2002). Maccoby and Jacklin's (1974) pioneering work on gender differences shaped the current research trend toward examining the accuracy of claims that males and females differ on verbal ability, quantitative ability, and spatial ability. Hedges and Nowell (1995) synthesized the results from several gender difference studies, which used nationally representative samples on large-scale assessments. Overall, their analyses suggested that gender differences are small for most areas of achievement, with the exception of writing achievement, science achievement, and stereotypically male related occupations (Hedges & Nowell, 1995). Although some findings suggest that average gender differences in science are decreasing (Linn & Hyde, 1989), Hedges and Nowell (1995) found that across the 32-year period they examined, gender differences were relatively stable. Research on gender differences reveal trends in content and skill areas, in which males and females differ (Beller & Gafni, 1996; Halpern 1997; Hamilton, 1998; Hedges & Nowell, 1995; Linn & Peterson, 1985). These trends

are especially apparent in spatial ability items, which reveal large male advantages along with physical science and earth and space science items. When considering the dimensional structure found previously in the NELS: 88 (Ayala, et al., 2002; Hamilton, et al., 1995; Nussbaum, et al., 1997), a large male advantage was found and the difference was attributed to the performance on the spatial mechanical reasoning (SM) dimension, which consists primarily of items that could be classified as physical science items. Beller and Gafni (1996) analyzed the 1991 International Assessment of Educational Progress (IEAP) and found a significant male advantage on physical science and earth and space science items. A similar pattern of male advantage was found for fourth grade students on the Third International Math and Science Study (TIMSS); while males outperformed females on physical and earth science items, little difference was found between males and females for life and nature of science, or for environmental issues (Hamilton, 1998). Although gender differences are frequently found on spatial ability measures, it is unclear how spatial ability specifically relates to the construct of science achievement. Furthermore, it is unclear why this advantage exists. There exists some research to suggest that males are more attracted to both extracurricular activities and courses that establish and enhance spatial abilities (Hamilton, 1998); however, this hypothesis has not been fully investigated or empirically tested.

*Gender format differences.* The likelihood that male and female performance diverges on assessment format (e.g., multiple choice [MC] and constructed response [CR]) has resulted in several studies examining this possible form of test bias (e.g., Klein, Jovanovic, Stetcher, McCaffrey, Shavelson, Haertel, Solano-Flores, & Comfort, 1997; Resnick & Resnick, 1992). The general trends have indicated that males tend to perform

better than females on MC tasks in science whereas females perform better on CR tasks in science (Resnick & Resnick, 1992). The possibility that MC and CR tasks measure different cognitive skills may explain, in part, why males and females perform differently across these tasks. If MC and CR are measuring different aspects of achievement, an interaction between item format and gender might be expected. The literature to follow has attempted to illuminate some of the reasons for gender differences on item format.

A tentative explanation of gender differences on CR tasks is that females experience performance advantages when scores depend on language usage; therefore, resulting in a female advantage on CR tasks (Henderson, 1999; Klein et al., 1997; Stumpf & Stanley, 1996). Klein et al. (1997) demonstrated that females generally performed better than males on hands-on science tasks that required attention to detail and reading. On the other hand, males outperformed females on items that required inferences and prediction. A comprehensive review of gender and fair assessment conducted by Willingham and Cole (1997) led to the conclusion that although females had the tendency to perform better on CR formats than on MC formats, this effect was not consistent, as many studies also demonstrated that females can also perform well on MC items. It was also found that gender format differences in mathematics, language, and literature did not occur as frequently as gender format differences in science test items. Beller and Gafni (2000) suggested that the superior verbal abilities of women may be better illuminated in CR items. They further suggested that writing ability may also play a role in the differential performance of women on CR items. In addition, they suggest that males may perform better on MC items as they take more risks in responding (as evidenced by guessing).

In an attempt to study gender format differences on the NELS: 88 science test and generate hypotheses about why the DIF was occurring, Hamilton (1999) used complimentary methodological approaches. Through the use of statistical DIF analyses and small-scale interviews gender differences were found to occur on items that required visualization requirements and items that required knowledge and skills obtained outside of the educational setting. If males possess stronger visualization skills and are more apt to use them in solving science items, this could explain boys outperforming girls in spatial reasoning. To fully appreciate how the multifaceted associations between format, content, and cognitive processes affect the performance of different groups of students, the investigation of possible contributing item features needs to be examined systematically (Hamilton, 1999). Small-scale interview studies offer one method, which can shed light on the gender differences associated with test item performance (Hamilton, 1999; Ercikan et al., 2004).

#### *Uncovering Higher-Level Thinking Skills*

Interview data, such as think aloud reports, can help yield hypotheses about gender differences and the underlying knowledge and skills measured by tests (Hamilton, et al., 1997; Ercikan et al., 2004). The interest in student performance, which goes beyond simple right and wrong response patterns has “increased the demand for data that trace cognitive processes” (Russo, Johnson, & Stephens, 1989). Think-aloud verbal protocols in which students are asked to verbally report their thoughts as they work through specified tasks have proven useful in examining the underlying cognitive skills that students employ in problem-solving (Ercikan et al., 2004; Ericsson & Simon, 1993; Hamilton et al., 1997; Norris et al., 2004). Think aloud methods offer one way to uncover

the substantive nature of dimensions at both the test level and item level (Hamilton et al., 1997; Leighton, 2004; NRC, 2001). The National Research Council (2001) suggests that the validity of inferences drawn from test performance can be improved when information is gathered about the specific knowledge and skills students actually use during test performance. The common approach in determining the knowledge and skills measured by tests is to consult with content experts, test developers and psychometricians. An inherent limitation to this approach is that content experts typically possess very different problem solving skills than students. Therefore, the hypotheses they generate or inferences they make about student performance may be misinformed (Leighton, 2004; Norris, et al., 2004). Protocol analysis offers an innovative way to support statistical investigations by allowing researchers to examine the actual scientific reasoning skills that students employ (Baxter & Glaser, 1998; Ercikan, et al., 2004; Ericsson & Simon, 1993; Hamilton, et al., 1997; Hamilton, 1998; Leighton, 2004; Norris, et al., 2004) as they solve science tasks.

Baxter and Glaser (1998) suggested a theoretical approach for evaluating the construct being measured by examining how the relationship between comprehensive verbal protocols, observation of student performance, and scoring criteria are evidenced in science assessments. Hamilton et al. (1997) used a small-scale interview study to aid in the interpretation of factors from the NELS: 88 science study. From this study, the researchers concluded that small-scale interviews could be used to enhance and support dimension interpretation in order to define the construct more clearly. Moreover, the interviews proved helpful in interpreting items that possessed inconsistent factor loadings. Other recent studies that utilized this method for investigating the latent traits

measured by tests have yielded information about the constructs that are measured by tests and potential explanations for student performance differences (e.g., Ercikan et al., 2004; Hamilton, et al., 1997). If the goal is to make valid inferences about student performance, it is imperative to examine the underlying knowledge and skills students bring to bear on tests of achievement.

### Methods and Results

The present study was conducted in three sequential steps. In the first step, a nonparametric technique was employed, Dimensionality Test (DIMTEST; Froelich, 2000; Froelich & Habing 2001; Nandakumar & Stout, 1993), to investigate the dimensionality of the English sample on the SAIP Science Assessment. If the test was found to measure more than one dimension as evidenced by DIMTEST, exploratory NOHARM (Fraser, 1988; McDonald, 1967, 1997, 1999) was used to determine the dimensional structure of the sample and the items that made up each dimension. In the second step, the SAIP was investigated for differential item functioning (DIF) using the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993a). The focal group consisted of the female sample and the reference group comprised the male sample that wrote the SAIP Science Assessment. In the third step, verbal reports based on a sample of SAIP questions identified as displaying gender DIF were collected for male and female grade 8 students using think-aloud methods (Ericsson & Simon, 1993; Leighton & Gokiert, 2005).

*Data*

Data from 8,373 grade 8 (4,224 males and 4,149 females) English speaking students who wrote the 2004 School Achievement Indicators Program (SAIP) Science Assessment were used to investigate the dimensional structure of the SAIP test, and gender differential item functioning (The Council of Ministers of Education, Canada [CMEC], 2000). CMEC uses the SAIP Science Assessment as a report card of Canadian students' knowledge and problem solving in science. The SAIP Science Assessment utilizes a two-stage testing procedure where students write an initial 12-item routing test (Test A) designed to assign them to a second stage test. The items in the SAIP Science Assessment are equally distributed across five ability levels (1 through 5). The 12-item routing test consists of items targeted to level three, which is of moderate difficulty. Students who receive a score less than 8 out of 12 are routed to the second-stage B test, while students receiving a score of 8 or greater are routed to the second-stage C test. Test B consists of 71 items targeted to ability levels 1 through 3, while test C consists of 65 items targeted to ability levels 3 through 5. Although the AB and AC tests can be considered distinct tests, items 1 through 12 (routing test) and items 65 through 83 are identical on both tests, which increases comparability. The first- and second-stage tests consist of both multiple-choice (MC) and constructed-response (CR) items dichotomously scored targeted to all ability levels. The test items included in the routing test and subsequent B and C tests address three broad science content domains: (1) knowledge and concepts of science, including biology, chemistry, earth, and physics (2) nature of science, and (3) relationship of science to technology and societal issues. Based on a student's performance on the B or C test items, they are assigned to one of 6

performance levels ranging from 0 to 5. Those students with a performance score below 3 demonstrate lower level science achievement in the measured domains, while those students scoring 3 or higher demonstrate relative strengths in the measured science domains.

For ease of analyses the tests were combined to form the AB and AC test. Although the test is administered to grade 8 and grade 11 students, this paper presents the analyses of the grade 8 sample only and is illustrated in Table 1.

### Stage 1: Dimensionality

#### *Dimensionality Method*

The most recent version of Dimensionality Test (DIMTEST; Froelich, 2000; Froelich & Habing 2001; Nandakumar & Stout, 1993) and exploratory Normal Ogive by Harmonic Analysis Robust Method (NOHARM; Fraser, 1988; McDonald, 1967, 1997, 1999) were used to investigate whether the SAIP Science Assessment is multidimensional.

DIMTEST is a nonparametric procedure that tests the null hypothesis that the set of items analyzed can be described in terms of a single dimension ( $H_0: d=1$  vs.  $H_1: d>1$ , where  $d$  is the number of dimensions). DIMTEST can be used within an exploratory or confirmatory paradigm. However, given the lack of structure with which to organize the SAIP Science data, exploratory DIMTEST was employed. For a more detailed description of DIMTEST see Froelich, (2000) and Froelich & Habing (2001).

The NOHARM program is used to estimate the multidimensional latent trait model of nonlinear factor analysis; the probability of a correct response is the dependent variable while the independent variable is the student ability. This is demonstrated

through the use of the cumulative distribution of a normal curve, called the normal ogive. NOHARM was used in exploratory mode; however, NOHARM is a confirmatory approach in that the suspected number of factors is specified in the NOHARM program. The fit indices, Tanaka's (1993) unweighted least squares goodness of fit index, and the root mean square residual (RMSR), are compared for different factor solutions. The main advantage of using NOHARM is that it can estimate the parameters for a large number of items because matrix inversion is not required (Gierl, Tan, & Wang, 2005). Linear factor analysis and nonlinear methods such as NOHARM produce very comparable results when working with SAIP data (Leighton, et al., 2005). For a more detailed description of NOHARM see Fraser (1988).

#### *Dimensionality Results*

The results from exploratory DIMTEST resulted in rejecting the null hypothesis of unidimensionality for the AB-13 test, but not for the AC-13 test. Given that  $d > 1$  for the AB-13 test as evidenced by DIMTEST, exploratory factor analysis of the tetrachoric correlations was conducted using NOHARM. Because the AC-13 test on the other hand, was unidimensional in nature ( $d = 1$ ) as evidenced by DIMTEST, no further dimensionality analyses were conducted.

*AB-13 test.* The NOHARM program was run in exploratory mode, for a 2-, 3-, 4-, and 5-dimensional model, given that there was no a priori model with which to fit the data. To evaluate goodness of fit, NOHARM reports Tanaka's (1993) unweighted least squares goodness of fit index, and the root mean square residual (RMSR). Presently, there are no published guidelines for Tanaka's index beyond the interpretation that a value close to one equals good model fit. A RMSR value equal to or less than four times

the reciprocal of the square root of the sample size indicates good model fit (Fraser, 1988). Given the sample size was 4,402 for the AB-13 test, a RMSR value of 0.0603 would be indicative of good model fit. Furthermore, the Chi-square change (Gierl & Rogers, 1996) was used to determine which dimensional model provided the best model data fit. A 2-dimensional model represented the most parsimonious fit to the AB-13 test with little to no change in Tanaka's index or the RSMR, and the Chi-square change was not significant, when all four models were compared. Given this result, it was concluded that the AB-13 test consists of two dimensions within an exploratory framework. The promax-rotated solution was used and if an item possessed a loading  $\geq 0.30$  it was considered to load on the dimension. Of the 83 items on the AB-13 test, 19 items loaded on dimension one, 37 items loaded on dimension two and 27 items did not load on either dimension.

### Stage 2: Differential Item Functioning (DIF)

#### *DIF Method*

The Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993a) is a statistical procedure that can be used to detect and estimate differential item functioning (DIF) on a given test. According to the Shealy and Stout model (1993a), an item is likely to display DIF if the item measures a secondary dimension along with the primary dimension the item was intended to measure. Camilli and Shepard (1994) and Roussos and Stout (1996) suggest that when an item displays DIF it is because the item is measuring the primary dimension assessed by the full set of items and a secondary dimension, creating multidimensionality within a test. The secondary dimension may measure construct relevant information and be termed an auxiliary dimension, or measure

construct-irrelevant variance and be termed a nuisance dimension. The purpose of SIBTEST, therefore, is to determine the difference between the probabilities of two groups, with equal ability on a latent trait, selecting a correct response. The reference group is typically considered the advantaged group while the focal group is typically considered the disadvantaged group. For example, in this study the reference group included males and the focal group included females. The amount of DIF present in a suspect item is denoted as  $\hat{\beta}_{UNI}$ , a parameter estimate with a standard normal distribution with mean of 0 and standard deviation of 1. If a statistically significant value of  $\hat{\beta}_{UNI}$  is found to be positive, this indicates DIF against the focal group; conversely, a negative value of  $\hat{\beta}_{UNI}$  indicates DIF against the reference group. Guidelines for determining the degree of DIF present in test items have been provided by the Educational Testing Services (ETS; Zwick & Ercikan, 1989) and adopted by Roussos and Stout (1996, p. 218, p. 220). A comprehensive description of the technical aspects of SIBTEST can be found in Shealy and Stout (1993a).

For the present study, both the AB-13 and AC-13 test were subjected to an item DIF analysis using SIBTEST.

#### *DIF Results*

*AB-13 results.* A single-item analysis was used to determine DIF on the AB-13 test. Ten of the 83 AB-13 test items displayed moderate DIF ( $0.059 \leq \hat{\beta}_{UNI} < 0.088$ ) and 14 items were identified as possessing large DIF ( $\hat{\beta}_{UNI} \geq 0.088$ ). Of the 14 items that possessed large DIF, 5 loaded on dimension one and favoured males, 8 loaded on

dimension two and favoured females, and the final DIF item that favoured males did not load on either dimension.

*AC-13 results.* A single-item analysis was used to determine DIF on the AC-13 test. Nine of the 77 AC-13 test items displayed moderate DIF ( $0.059 \leq \hat{\beta}_{UNI} < 0.088$ ), while 7 items were identified as possessing large DIF ( $\hat{\beta}_{UNI} \geq 0.088$ ). Of the 7 large DIF items, 3 items favoured males and 4 items favoured females.

Based on the DIF analyses conducted on the AB-13 and AC-13 test, items that possessed statistically significant large DIF were reviewed for possible use in the interview portion of the study. Due to the difficulty in conducting think-aloud interviews with students during class time, it was determined that 12 items was a reasonable amount of items for students to complete during a 45-minute interview. The following criteria were used to select items for the interview study: balanced item format representation (multiple choice and constructed response); equal gender DIF representation (female and male DIF items); items at differing difficulty levels (level one through five); and items that overlapped between both tests (AB-13 and AC-13). After examining the DIF items in consideration of the four criteria, 12 items were ultimately selected for use in the interview portion of the study. Table 2 illustrates the items that were selected.

### Stage 3: Protocol Analysis

#### *Protocol Analysis Method*

A structured interview was used to probe students' cognitive processing and meta-cognitive knowledge as they solved the 12 selected test items (Ericsson & Simon, 1993).

*Participants.* A total of 16 grade 8 students (9 males and 7 females) were recruited from a suburban junior high school. During a 45-minute interview, students were asked to think out loud concurrently as they solved each of the selected test items. Following each item administration students were asked retrospective questions in an attempt to probe their meta-cognitive knowledge of problem solving. All interviews were audio recorded and transcribed to maintain the accuracy of verbal reports. The concurrent portion of the interview was presented to students as follows:

*Thank you for agreeing to participate in this study. Please know that your participation is completely voluntary and you are free to go at any time. In this study, I am trying to find out what students your age think about when solving science questions on tests. In order to do this I'm going to ask you to THINK ALOUD as you work on the problems that I give you. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you give an answer.*

*I would like you to talk aloud CONSTANTLY from the time I present each question until you have given your final answer to the question. I don't want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time I will remind you to talk. Do you understand what I want you to do? I will tape record our session because I want to get an accurate record of your think aloud reports. Please know that all the information you share today with me will be kept confidential and anonymous. Do you have any questions?*

*Please tell me what you are thinking as you answer this question. Please remember to say everything that is going through your mind.*

Following the concurrent portion of each test item, students were asked the following retrospective questions:

1. *Now tell me all that you can remember about how you solved this question*
2. *Did you find any parts of this question confusing? If so,*
  - a. *What parts did you find confusing?*
  - b. *Why were they confusing?*
3. *Did you find any parts of the question helpful in answering the question? If so,*

- a. *What parts did you find helpful?*
- b. *How did they help you answer the question?*

After the students had completed all 12 items they were asked the following two general questions about their problem solving:

1. *Do you remember using any strategies to solve the items? If so,*
  - a. *What strategies and on what items*
2. *Do you remember using visualization to solve any of the items? If so,*
  - a. *Describe the visualization to me and tell me on what items you used it.*

### *Protocol Analysis Results*

Preliminary findings of themes that emerged within analyzed protocols (concurrent reports) from male and female grade 8 students are presented. Due to test security, the test items used in the protocol analysis cannot be presented in any detail in the paper. The following represents a breakdown of the content areas, as specified by CMEC, addressed within the six items found to favour females: three items measured the nature of science; one item addressed science and technology in society; one item related to knowledge of chemistry; and one item measured knowledge about physics. Males were favoured on six test items: three items measured knowledge of earth; two items measured knowledge about physics; and one item measured content related to science and technology in society. As is evident based on this breakdown, there appears to be some overlap between the content areas of items found to favour either gender. Further analysis of question type and content may offer a deeper explanation as to why DIF may be occurring.

The following three themes, which highlighted male and female cognitive processing and performance differences, emerged from the concurrent interview data (1)

reading comprehension, (2) the use of visualization and background knowledge in solving items, and (3) strategy use in solving items.

The first theme, reading comprehension, was present in all of the six DIF items that favoured females. A heavier reading component, including background stories related to the item and item stems, was required on these items. Furthermore, at least three items required attention to detail in comprehending what the question was asking of students. For example, within the item stem words such as “least”, “not necessarily” and “most likely” were italicized to emphasize their importance in answering the question. To illustrate this with an example, one question required students to select the “factor which had the *least* effect” on a given structure. Despite the fact that the key word was italicized, males had the tendency to skip this word as evidenced by the concurrent reports, and assume the question was asking for the “most” or “largest” effect, resulting in an incorrect response. During the concurrent interviews it was apparent that females focused in on the word *least* by commenting that it was helpful in answering the question correctly. During the retrospective interview, at least three of the males realized they had misinterpreted the question recognizing the word *least* and changed their answer.

The second theme, visualization and background knowledge was evidenced on five of the six DIF items that favoured males. Of the five items, four items included a picture or diagram that could be used to answer the question. During the concurrent portion of the interview, those students that answered the item correct – both male and female – reported the use of visualization while solving the question. Furthermore, those students that performed well on these items mentioned their background or out of school experiences as the primary reason for the answer they selected. For example, one item

dealt with campfires and portable stoves and those students that had engaged in camping had no difficulties answering this question and would describe their experiences and how those lead them to the answer they chose. Another item which required both visualization and background experience, dealt with drawing the trajectory of a hockey puck as it goes from one player, to the boards, and back to another player. Students stated that visualizing the angle of incidence and angle of reflection aided in answering the question. Furthermore, they commented on the usefulness of the hockey rink diagram in answering the question. Students also mentioned a number of outside activities including billiards, hockey, and indoor soccer, as the main source of their knowledge for this item. Based on the DIF results, this item favoured males. This is in line with findings from the protocol analysis where all males successfully completed the item where only 5 of the 7 grade 8 females did the same. Those females that were successful reported the use of visualization and/or personal experience in helping to answer the question.

The final theme involved systematic differences in strategy use by male and female students when solving the 12 items. When describing how they determined item responses, both groups of students engaged in, and described the use of, process of elimination – the act of discarding item options based on their knowledge and experiences. When males reviewed each alternative they were very confident in acceptance or rejection of an alternative as a possible answer, they would respond “yes” or “no” to possible alternatives. Furthermore, they appeared to engage in reasoning about each acceptance. For instance, the majority of male students provided a reason for why they thought an alternative was a reasonable answer. On the other hand, the female students had the tendency to provide reasons for why they eliminated an item alternative

as opposed to why they selected their final answer. Although the female students appeared confident in their decisions to reject an alternative, they were more hesitant when selecting the final solution to the problem. When determining the answers for constructed-response items, which do not allow for direct process of elimination of alternatives, females would go so far as to continue to provide probable alternatives, and reasons why they were not selecting a particular answer. In one case, a female student prepared three to four alternatives for each constructed response question and would describe why the alternatives she was eliminating did not answer the question. In essence, she created a multiple-choice question.

#### Discussion

This study introduced a three-stage approach to understanding gender differences on a large-scale science assessment. The results of the study indicate that these three approaches, when used together, demonstrate a promising way of identifying and understanding gender differences. Study into the underlying knowledge and skills that are measured by large-scale tests of achievement can be accomplished through data driven procedures for assessing the dimensionality of the test. The dimensionality results from this study, when utilizing DIMTEST and NOHARM, indicated that the SAIP AB-13 test is multidimensional and two dimensions can best explain student performance in science. The first dimension included 19 items that measure content related to biology, chemistry, earth, physics, and science and technology in society. Although items fell across all of these content areas, the majority of the items within the first dimension related to content in the areas of biology, earth, and physics.

As with items found to load on the first dimension, items within the second dimension similarly addressed content across broad science domains. What differed on the second dimension was the inclusion of an additional content area, that being the nature of science. The majority of items that loaded on the second dimension related to content in chemistry, nature of science, and science and technology in society. Defining these two dimensions in terms of the content, cognitive processing, and skills required for performance, although not the focus of the present paper, is a necessary and future step that will be taken in describing the construct of science achievement as measured by the SAIP. Furthermore, systematic gender differences were found on each dimension; dimension one possessed large DIF, which systematically favoured males, while dimension two included large DIF items that systematically favoured females. It was determined through the use of DIMTEST that the SAIP AC-13 test was unidimensional and that one dimension was most appropriate in defining student performance. Given the unidimensional nature of the AC-13 test, systematic performance differences for males and females cannot be described in terms of clusters of items.

Based on the DIF results, 12 items were selected for inclusion in the protocol analysis study. The 12 selected items possessed an equal representation of item format (MC and CR), equal representation of gender DIF (males and females), equal distribution of difficulty level (difficulty ranging from 1 to 5), and equal distribution of items across both tests (AB-13 and AC-13). Three preliminary salient themes were identified after reviewing the interview data from 16 grade 8 students (9 males and 7 females) that engaged in concurrent and retrospective questioning while solving the 12 items.

The first theme addresses reading comprehension. It was found that males had the tendency to rush through items, often missing essential details within a question, necessary for a correct response. This finding suggests that in some instances a construct-irrelevant feature, in this case reading comprehension, may be impacting whether an item is answered correctly. If the construct of interest is scientific knowledge and skills and not reading comprehension, this may suggest directions for the modification of items that possess a heavy reading component.

Visualization was the second theme identified and could be considered a construct relevant and important skill that is necessary when attempting certain science items. This is consistent with previous findings where visualization, a strategy more frequently used by male students, was related to spatial-mechanical reasoning and/or physics content (Hamilton, 1997). Items that necessitate the use of visualization, but that also introduce construct-irrelevant knowledge, for example, knowledge about sports, may result in DIF for reasons unrelated to the use of visualization. That is, when an item includes both a visualization component and a construct-irrelevant topic (e.g., hockey) that may favour one gender over the other, it becomes difficult to determine if the DIF is occurring as a result of a construct relevant or irrelevant feature. If visualization and spatial-mechanical reasoning are important cognitive skills in science and are likely to continue to play a prominent role within science tests of achievement, it will be important to teach all students the effectiveness of using visualization when approaching and solving these types of questions. As evidenced in the findings of this study, items with a visualization component were often related to content that is learned outside of school, resulting in a disadvantage for those students that could not rely on personal experiences to answer the

question correctly. Ultimately the developers of tests, such as the SAIP, should take into consideration the effect that construct-irrelevant features can have on student outcomes and modify items accordingly.

The final theme illuminated an interesting difference between how the males and females within the interview study approached multiple choice and constructed response items. Specific to multiple-choice items, males provided less extensive reasons as to why item alternatives were rejected, and instead focused on providing explanations for why an item alternative was selected. In contrast, females spent little time explaining why an alternative was selected, and more time describing the process of elimination and why item alternatives were rejected. Whether this reflects an actual preference for a response strategy and whether it is related to performance differences may be an area of future exploration.

This paper has presented preliminary findings of the utility of a three-stage approach to understanding gender differences on a large-scale test of science achievement. That one of the tests examined in this study revealed multidimensionality further suggests that a total test score does not always adequately reflect the construct of science achievement. This particular test also revealed systematic gender differences on the two dimensions found. Examining the content of items that loaded on either dimension did not reveal obvious content differences, aside from knowledge of the nature of science, which was found more readily on the dimension favouring females. In order to fully describe the underlying knowledge and skills measured across dimensions, it will be necessary to examine each item for item format, more subtle content differences, and reasoning skills. Interview data may offer some direction as to how items should be

examined. Preliminary themes extracted from interview data (concurrent reports from grade 8 students) suggest that bias within some test items may, to some extent, account for performance differences on those items found to display DIF. Whether the preliminary themes can be used to explain performance differences on the 12 items included in the protocol analysis and on all remaining DIF items within the AB-13 and the AC-13 tests is one of the next steps in analyzing this data. Related to this is exploring how interview data might be used to tease apart whether the DIF is a result of bias or impact ultimately enhancing the fairness and validity of test score interpretation (Beller & Gafni, 1996; Halpern 1997; Hamilton, 1998; Hedges & Nowell, 1995; Linn & Peterson, 1985).

## References

- Alberta Education (2005). *Public information*. Retrieved February 18, 2005, from <http://www.education.gov.ab.ca>.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Ayala, C.C., Shavelson, R.J., Yin, Y., & Shultz, S.E. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment: Issues and Practice*, 8(2), 101-121.
- Baxter, G.P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17(3), 37-45.
- Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology*, 88, 365-377.
- Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, 42 (1/2), 1-21.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.
- Childs, R.A., & Oppler, S.H. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high-stakes testing program. *Educational and Psychological Measurement*, 60(6), 939-955.

- Council Ministers of Education, Canada (2000). *Public report on science assessment: SAIP School Achievement Indicators Program 1999*. Retrieved August 12, 2004, from <http://www.cmec.ca/saip/science2/science2.en.stm>.
- Embretson, S.E. (1999). Cognitive psychology applied to testing. In F.T. Durso (Ed.), *Handbook of applied cognition* (pp.629-660). Chichester, England: John Wiley & Sons.
- Ercikan, Law, Arim, Domene, Lacroix, & Gagnon (2004). *Identifying Sources of DIF Using Think-Aloud Protocols: Comparing Thought Processes of Examinees Taking Tests in English versus in French*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.
- Ericsson, K.A., & Simon, H.A. (1993). *Protocol analyses: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.
- Fraser, C. (1988). *NOHARM: an IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England.
- Frenette, E. & Bertrand, R. (2000, April). *Assessing dimensionality with TESTFACT and DIMTEST using large-scale assessment data sets*. Paper presented at the annual meeting of the American Educational Research Association (AERA). New Orleans, LA.
- Froelich, A.G. (2000). *Assessing unidimensionality of test items and some asymptotics of parametric item response theory*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.

- Froelich, A.G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Gierl, M.J., & Rogers, W.T. (1996). A confirmatory factor analysis of the test anxiety inventory using Canadian high school students. *Educational and Psychological Measurement, 56*, 315-324.
- Gierl, M.J., Rogers, W.T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.
- Gierl, M.J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT: Research report No. 2005-11*. Retrieved March 3, 2006, from <http://www.collegeboard.com>.
- Gierl, M.J., Bizanz, J., & Bizanz, G.L., Boughton, K. A., & Khaliq, S.N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36.
- Haladyna, T.M. & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.
- Halpern, D.F. (1997). Sex differences in intelligence. *American Psychologist, 52*(10), 1091-1102.

- Hamilton L.S. (1998) Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis, 20*, 179-195.
- Hamilton, L.S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(3), 211-235.
- Hamilton, L., Nussbaum, E.M., Kupermintz, H., Kerkhoven, J.I.M., & Snow, R.E. (1995). Enhancing the validity and usefulness of large scale educational assessments: II. NELS:88 science achievement. *American Education Research Journal, 32*, 555-581.
- Hamilton, L.S., Nussbaum, E.M., & Snow, R.E. (1997). Interview procedure for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.
- Hamilton, L., Stecher, B., & Klein, S. (Eds.). (2002). *Making sense of test-based accountability in education*. Santa Monica, CA:RAND.
- Hedges, L.V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41-45.
- Henderson, D. (1999). *Investigation of DIF across item format*. Unpublished Doctoral Dissertation, University of Alberta at Edmonton, Alberta.
- Klein, S.P., Jovanovic, J., Stetcher, B.M., McCaffrey, D., Shavelson, R.J., Haertel, E., Solano-Flores, G., & Comfort, K. (1997). Gender and racial/ethnic differences in

- performance assessments in science. *Educational Evaluation and Policy Analysis*, 19(2), 83-97.
- Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*,
- Leighton, J. P. (2004). Avoiding Misconceptions, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice*, Winter, 1-10.
- Leighton, J.P., Gierl, M.J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, 41, 205-236.
- Leighton, J.P., & Gokiert, R.J. (April, 2005). *The cognitive effects of test item features: Identifying construct irrelevant variance and informing item generation*. Paper presented at the National Council on Measurement in Educational Annual Meeting, Montreal, Canada.
- Leighton, J.P., Gokiert, R.J., & Cui, Y. (2005). *Investigating the Statistical and Cognitive Dimensions of Large-Scale Science Assessments*. Paper presented at the American Educational Research Association Meeting, Montreal, Canada.
- Linn, M.C., & Hyde, J.S. (1989). Gender, mathematics, and science. *Educational Researcher*, 18, (8), 17-19, 22-27.
- Linn, M.C., & Peterson, A.C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 56, 1479-1498.
- Maccoby, E.E., & Jacklin, C.N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

- McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, No. 15.
- McDonald, R.P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York: Springer.
- McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- McGehee, J.J., & Griffith, L.K. (2001). Large-scale assessments combined with curriculum alignment: agents of change. *Theory into Practice*, 40(2), 137-144.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3<sup>rd</sup> Ed., pp. 13-103). New York: American Council on Education, Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher*, 23 (2), 13-23.
- Moss, P.A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6-12.
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, 18, 41-68.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. J.W. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Washington, DC: National Academy Press.
- Norris, S.P., Leighton, J.P., & Phillips, L.M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education*, 2, 283-308.
- Nussbaum, E.M., Hamilton, L.S., & Snow, R.E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: IV. NELS:88 science

- achievement to 12<sup>th</sup> grade. *American Educational Research Journal*, 34(1), 151-173.
- Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Changing Assessments: Alternative views of aptitude, achievement and instruction*. Norwell, Massachusetts: Kluwer Academic Publishers.
- Roussos, L.A., & Stout, W.F. (1996). Simulation studies of the effects of small sample size and studies item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33, 215-230.
- Ryan, J.M., & Demark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Shealy, R., & Stout, W.F. (1993a). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, 58, 159-194.
- Shealy, R., & Stout, W.F. (1993b). An item response theory model for test bias and differential test functioning. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 197-239). Hillsdale, NJ: Erlbaum.
- Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational Measurement* (3<sup>rd</sup> Ed., pp. 263-331). New York: American Council on Education, Macmillan.

- Stumpf, H., & Stanley, J.C. (1996). Gender-related differences on the College Board's advanced placement and achievement tests, 1982-1992. *Journal of Educational Psychology, 88*, 353-364.
- Tanaka, J.S. (1993). Multifaceted concepts of fit in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.
- Tate, R. (2002). Test dimensionality. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Tindal, G. & Haladyna, T.M. (Eds.) (2002). *Large-Scale assessment programs for all students: Validity, technical adequacy, and implementation*. Mahway, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Willingham, W.W., & Cole, N.S. (1997). *Gender and fair assessment*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Zwick, W.R., & Ercikan, K. (1989). Analyses of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 53-66.

Table 1

Sample of grade 8 students that completed the AB and AC tests

	AB-13 Test	AC-13 Test
Combined	4,402	3,971
Males	2,133	2,091
Females	2,269	1,880

Table 2

Items selected for protocol analysis

Item	DIF Favouring	Test	Item	
			Format	Content
3	Female	AB and AC	MC	Knowledge of Physics
9	Male	AB and AC	MC	Knowledge of Earth
12	Male	AB and AC	MC	Knowledge of Earth
30	Female	AB	CR	Nature of Science
42	Female	AB	CR	Knowledge of Chemistry
46	Male	AB	CR	Science, technology, and society
56	Male	AB	MC	Knowledge of Earth
79	Male	AB and AC	CR	Knowledge of Physics
82	Female	AB and AC	MC	Nature of Science
95	Male	AC	MC	Knowledge of Physics
106	Female	AC	MC	Nature of Science
122	Female	AC	MC	Science, technology, and society

