

Running Head: Differential Bundle Functioning

**Differential Bundle Functioning on Social Studies High
School Certification Exams**

Keith A. Boughton

Tess E. Dawber

Laurie-Ann M. Hellsten

University of Alberta

Paper Presented at the Annual Meeting of the
American Educational Research Association (AERA)

Seattle, Washington, USA

April 10-14, 2001

Differential Bundle Functioning on Social Studies High School
Certification Exams

Most provinces and territories throughout Canada have introduced large-scale learning assessment programs into their educational systems. These standardized achievement tests are used for accountability, program evaluation, program improvement, and student evaluation. Student performance is evaluated on well-defined learning objectives put forth by the educational ministries. The Alberta Ministry of Education tests 11 subject areas at the grade 12 level, including English, Social Studies, Mathematics, Chemistry, Biology, and Physics, based on the provincial curriculum. In the final year of high school, the examinations are high stakes since the students' final marks in a course are a combination of the teacher-assigned mark and the diploma exam, each accounting for 50 percent of the overall mark (Lafleur & Ireland, 1999). The outcome of the final marks determines whether students pass or fail the course, whether students are granted or denied entrance to post-secondary institutions, and whether students are awarded scholarships. Given the importance of these exams, test developers need to ensure their examinations are valid and fair for all examinees.

Validity is of central concern to those employing standardized tests for evaluative purposes. Validity is defined as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.... Inevitably, then, validity is an evolving property and validation is a continuing process. Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide most current use of the test and current research to advance understanding of what the test scores mean" (Messick, 1993, p. 13).

Procedures have been developed over the last 15 years to help ensure that tests are fair for all intended groups and purposes. Most test developers screen for items that may exhibit differential performance across groups. Statistical procedures, however, are only capable of identifying items that function differentially across groups. The procedures cannot indicate whether the item unfairly favors one group over another (i.e., item bias) or whether group

differences reflect the ability intended to be measured (i.e., item impact). Items flagged using these methods have been termed “differential item functioning” (DIF) to denote the lack of specificity (Camilli & Shepard, 1994).

When two groups, such as native/non-native, male/female, or African-American/Caucasian, have the same ability it is expected they will receive similar scores on a particular item or group of items. However, one group may consistently receive lower scores because of insufficient knowledge to answer the item(s) or because something other than the subject area is being measured. Therefore, we would expect members of the comparison groups to obtain the same probability of correctly responding to an item or group of items when their ability in the subject area is the same. When this does not occur, there is bias against one group. Inferences generated from the test may be questioned (Angoff, 1993).

Differential Item Functioning vs. Differential Bundle Functioning

Differential performance may be assessed for individual items, known as differential item functioning, or for groups of items, known as differential bundle functioning (DBF). Considerable attention has been paid to DIF over the last 15 years. It is not well understood why many items display DIF. Although DIF can be detected statistically by a computer program called Simultaneous Item Bias TEST (SIBTEST; Stout & Roussos, 1999), there is a need for substantive interpretation of the results to determine whether the item displays bias or impact. If the item is biased, the item must be removed or revised. If the item demonstrates impact, further investigation is necessary to explore why one group achieves a higher score for an item. Herein lies the problem of single item DIF—when reviewing a single item, numerous hypotheses can be generated to explain why an item functions differentially (Angoff, 1993; Bond, 1993; Camilli & Shepard, 1994; O’Neil & McPeck, 1993). The use of content specialists to explain the causes of DIF or to predict which items will display DIF has yielded poor results (Angoff, 1993; Bond, 1993; Camilli & Shepard, 1994; Englehard, Hanche, & Rutledge, 1990; O’Neil & McPeck, 1993; Plake, 1980; Rengel, 1986; Sandoval & Mille, 1980). All that remains after the discovery of single item DIF is a trail of speculation that may be misleading or uninformative.

DBF is a natural extension of DIF. The building blocks of exams are often small bundles of items. These bundles provide an opportunity for testing DIF amplification (Douglas, Roussos, & Stout, 1996). Nandakumar (1993) described how DIF amplification is fundamental to the underlying premise for studying fairness at the bundle level. She suggested that DIF may be present in small quantities that may go statistically undetected in the single item approach, but may be detected at the bundle level. She studied SIBTEST's role in the detection of DIF amplification and found greater statistical power nested within the bundle approach compared to the single item DIF approach. DIF amplification is caused by items acting in concert, each contributing to an overall unacceptable level of DBF. Amplification is an important reason why test fairness should be studied at the bundle level rather than at the item level.

In an effort to obtain meaningful and interpretable results, Douglas et al. (1996) advocate a confirmatory approach such that the substantive analysis drives the statistical analysis. Stout and Roussos (1999) state that:

In order to achieve a maximally statistically effective and substantively informative DIF/DBF analysis, the essential step in augmenting, indeed sometimes even replacing, a standard one at a time DIF analysis is to select bundles judged to be substantively homogeneously and/or statically dimensionally homogenous and then to analyze each selected bundle for DBF. When a homogeneous bundle is found to display DBF, it is often possible to reliably provide a substantive explanation for why the DBF has occurred (p. 3).

Gierl, Bisanz, Bisanz, Boughton, and Khaliq (2001) showed that the table of specifications provides a conceptually meaningful way to bundle items. Figure 1 illustrates a DIF analysis performed on a multiple-choice science test. The difference between the item and bundle level approaches is clearly demonstrated. At the item level, the open circles, representing moderate to large effect sizes, are scattered across the content areas. At the bundle level, one can more easily define the commonalities for a set of items chosen on a theoretical organizing principle by observing whether the majority of items favour males or females. At the bundle level, Figure 1 reveals that the content area

of Air and Aerodynamics favors males and the content area of Observation and Inference favors females.

Multidimensional-Based Approach to DIF

Shealy and Stout's (1993) multidimensional model asserts that an item has the potential to display DIF if it is measuring a secondary ability in addition to the target ability the test was designed to measure. Secondary abilities are conceptualized as two types: auxiliary abilities, which are part of the construct intended to be measured, and nuisance abilities, which are outside of the construct intended to be measured. DIF attributable to an auxiliary dimension is referred to as benign whereas DIF arising from a nuisance dimension is referred to as adverse. Test bias occurs when the groups possess nuisance abilities in varying amounts. Therefore, when a test has items that function differentially across groups, the test may possess multidimensionality because it measures a nuisance dimension in addition to the intended dimension. An example of adverse DIF as it relates to DIF amplification is presented below.

Imagine a reading comprehension test containing a paragraph-based bundle of items concerning American professional football. If females are the focal group and males are the reference group, one might suspect that most of the items would display adverse DIF in favor of males. However, the amount of DIF present in any single item could be quite small and thus difficult to detect statistically. Nonetheless, over several items, small amounts of DIF can add up to an unacceptable amount of DIF at the bundle level—in other words, an unacceptable level of DBF (Douglas et al., 1996, p. 468).

A computer program that uses a multidimensional-based approach for the detection of DIF and DBF is SIBTEST (Stout & Roussos, 1999). The model only works if there is a portion of the test that measures the intended construct without bias. A crucial yet unresolved issue concerns the valid subtest with which the differential functioning of items or bundles will be compared. The assumption is made that the majority of items will be unbiased. To the extent that this holds true, a valid subtest will be found to match examinees at the different ability levels from which the DIF/DBF items can be compared. It is assumed for long tests that examinees with the same number-correct score on the valid subtest are of equal ability and thus comparable. Examinees from the reference (i.e., group considered to be favored by the test) and focal (i.e.,

group considered to be disadvantaged by the test) groups with similar valid subtest abilities are compared across all suspect bundles, the subtest under review for bias or impact (Nandakumar, 1993).

The statistical procedure employed by SIBTEST starts with the null hypothesis

$$H_0: \mathbf{b}_U = 0,$$

and the alternative hypothesis

$$H_1: \mathbf{b}_U > 0,$$

where the β_u (called beta-uni) denotes how much DBF there is against the focal group (i.e., the group scoring lower on the bundles). The actual test statistic for the above null hypothesis is

$$B = \frac{\hat{\mathbf{b}}_U}{\hat{\mathbf{s}}(\hat{\mathbf{b}}_U)},$$

where

$$\hat{\mathbf{b}}_U = \sum_0^n \hat{p}_k (\bar{Y}_{Rk} - \bar{Y}_{Fk}).$$

\hat{p}_k is the proportion of the focal group attaining a certain subtest score, $k = 0, 1, \dots, n$. \bar{Y}_{Rk} and \bar{Y}_{Fk} are the adjusted means using a regression correction of the suspect subtest for examinees with a valid subtest score of $k = 0, 1, \dots, n$, for the reference and focal groups respectively. The estimated standard error for $\hat{\mathbf{b}}_U$ is given by this expression

$$\hat{\mathbf{s}}(\hat{\mathbf{b}}_U) = \left[\sum_{k=0}^n \hat{p}_k^2 \left(\frac{1}{J_{Rk}} \hat{\mathbf{s}}^2(Y|k, R) + \frac{1}{J_{Fk}} \hat{\mathbf{s}}^2(Y|k, F) \right) \right]^{1/2},$$

where $\hat{\mathbf{s}}^2(Y|k, g)$ is the sample variance of the studied subtest for examinees in group g , which stands for either the reference or focal group. The k represents each score level on the valid subtest across which both groups will be compared. J_{Rk} and J_{Fk} are the sample sizes for the reference and focal groups respectively with a score of k on the valid subtest. $\hat{\mathbf{b}}_U$ can be used as an estimate of the amount of DBF, where positive values favor the focal group and negative values favor the reference group.

Literature Review

Systematic gender differences in the Alberta Social Studies diploma examinations have been observed over the past 9 school years. Males outperform females on the exams by a margin of 3.0 to 4.7 percentage points (see Table 1). However, females are equal to or slightly better than males, demonstrating a 0.1 to 5.2 percentage point advantage, in school awarded marks (see Table 2). The Student Evaluation Branch at Alberta Learning initiated a Special Study into the gender differences on the Social Studies diploma examinations after a 6-percentage point difference was observed between the average scores of males and females on the multiple-choice (MC) section during the school year 1991-1992. The Special Report indicated that males outperformed females on the *Political and Economic Systems* and the *Global Interaction in the Twentieth Century*, but the differences were more pronounced, almost double, for the *Global Interaction in the Twentieth Century* section. The latter is informally referred to as 'war.' The gender difference was consistently found across exam sittings and years, reaching more than double the size of gender differences on the MC components of the other diploma exams. The essay component of the Social Studies diploma examination, however, does not yield the same gender differences observed in the MC section. On average, female students scored slightly higher than male students on all subscales of the essay in 4 of the 6 exams prepared from 1990 to 1992 (Alberta Education, Special Study, 1992).

The finding that males outperform females on the MC section is not unique. There is an abundance of research to support the claim that the MC format favors males (Aiken, 1987; Barrs, 1990; Bolger & Kellaghan, 1990; Kirkland, 1971; Rosser, 1989; Stobart, Wood, & Quinlan, 1992). In addition, Sadker, Sadker, and Klein (1986) found that a contradiction existed between males and females at the secondary school level where males outperform females on standardized tests and females outperform males on report card grades.

With regard to the social studies curriculum content, research has shown that certain topic areas are considered more masculine, such as military, technology, and the environment, while other topic areas are considered more feminine, such as family planning, domestic issues, education, training, and social and ethnic issues (Eagley & Carli, 1981). The 1992 Special Report

examined the interests of 897 students who participated in Social Studies diploma field tests. The students' indicated their interest level on the topics covered in the curriculum. Females expressed greater interest than males in the communitarian topics, including communism, Sweden, League of Nations, and Peace Movements. Males expressed greater interest than females in the confrontational topics, including battles, strategies, tactics and turning points in the Second World War, and other conflicts such as civil wars and small disputes. In short, most of the grade 12 Social Studies curriculum may be more interesting to males than females.

The 1992 Special Study collected information on students' extracurricular activities relating to social studies topics. Results indicated that males engaged in more activities relating to war or historical topics (i.e., read newspapers, watched documentaries/movies, read magazines/books on such topics), although males and females devoted approximately the same amount of time to news programs, documentaries or movies about social issues, and books or magazine articles on current events or social issues. Similarly, Campbell (1994) studied the extracurricular activities of students in the Alberta Social Studies course. He found that males were more involved in extra curricular activities on war-related topics and historical events. Only when the topic was social issues did female involvement exceed male involvement.

Moreover, a review of four Social Studies diploma exams (January and June administrations of 1984 and 1994) by Christison (1995) found few female references and many male references. The male references were often made to males holding positions of power and influence. According to Christison (1995, 1997), the exams and the corresponding grade 12 Social Studies textbooks had an intensely male sense. Christison (1997) stated that for female Social Studies students, the feeling and experience of "being the other" is prevalent. These solitary experiences may transpire because of minimal material in the Social Studies curriculum with which females can identify (Christison, 1997). Sadker et al. (1986) found that lack of female representation in literature and history textbooks is common. However, the most important consequence of the lack of female representation on the Social Studies diploma exams may be that students perform better on tests when they identify with the topics being covered by the test items (Neill & Medina, 1989).

In summary, research investigating the gender differences in Social Studies suggests the MC format appears to favour males, males are more interested in the topics covered by the curriculum and devote more extra curricular time to these interests, and male role models are primarily presented in the materials used. As a result, males are able to identify more closely with the course material and the test items.

The Construction of a Theory

Our theoretical rationale for developing hypotheses about differential performance on the Social Studies exams was obtained from the work of Walter (1996) and Walter and Young (1997). Walter (1996) undertook a content analysis of the MC component of the Social Studies diploma exams from 1991 to 1993 based on the techniques of Berg (1989). She reviewed all the exam questions and categorized each question according to the topic. The purpose of her study was to address the gender differences from a qualitative, feminist perspective. Question content included government strategies of power and control, economic tactics of regulation and control, coalition for world power and control, and the failures thereof. Walter (1996) concluded that the predominant themes of almost all questions were power relations and control, and threats of violence, whether it was overtly or covertly stated.

Walter (1996) also observed that a conflict model of human experience prevailed over a nurturing model. Words conveying conflict and aggression were more prevalent than words conveying connection, harmony and nurturance. To illustrate the underlying theme of aggression, words such as control, differences, threat, challenge, oppose, and enforce were used widely in the Global Interaction section. The Political and Economic Systems questions contained concepts such as winning, losing, and fighting to control forces. Moreover, words normally associated with nurturing, such as support, collective, security, and development, were often used in the context of a conflict model. For example, the use of “support” was used to communicate choosing sides in a conflict.

In conclusion, Walter suggested that the content of Social Studies diploma examinations exclusively presented the male perspective; focused on power relations and control; promoted a

conflict model over a nurturing model through word choice and constructions that convey the struggle for power and domination; and featured male role models.

Walter's (1996) content analysis also included a reconceptualization of the test specifications. The following broad topic areas appeared: *Economics, Politics, History, Peace Initiatives and Internationalism*, and *Control Tactics and Strategies*. The first two categories correspond to *Political and Economic Systems* and the latter three categories correspond to *Global Interaction in the Twentieth Century* defined by Alberta Education.

Our interest in the content analysis of the Social Studies Diploma exams by Walter (1996) originated from a desire to develop a theoretical rationale for developing hypotheses about differential performance on the exams. Despite intuitive appeal of the results of Walter's content analysis, five fundamental concerns are identified.

The first concern involves Walter's (1996) reconceptualization of the test specifications into five broad topic areas. Walter (1996) and Walter and Young (1997) provide only a minimum description of the development process of the five content areas. For example, Walter and Young (1997) report that Walter (1996) "drew on the extensive feminist critique of traditional social studies and history curricula" (p. 83) to develop the content topic areas. Despite the claim by Walter and Young (1997) that the content analysis was exploratory in nature, no direct connection was established between the feminist theoretical rationale for the gender differences and the rationale for the reconceptualization of the content areas.

Second, although Walter (1996) provides the test specifications for her regrouping of the items, the lack of information about the development process is problematic. It is very difficult for an independent coder to reproduce the coding system without reviewing each item in turn and carefully inspecting the types of items situated in each content area. In addition, no inter-rater reliability analysis is reported by Walter to demonstrate that an independent rater could replicate the coding system.

A third concern relates to the division of items into a single category. Walter (1996) placed some questions into two categories. A total of 45 items across the six exams were double-coded, with a range of 3 to 15 items per exam ($M = 7.5$). The most frequently overlapping categories

were Economics Theory and Political Theory (8 items were double coded), followed by History and Control Tactics (7 overlapping items), and Economics Theory and Politics Applied (6 overlapping items). The following pairs of categories had 4 items each double coded: Economics Applied and Politics Applied, History and Politics Applied, Control Tactics and Politics Applied, and Peace Initiatives and Control Tactics. Given the frequency of double coded items, the categories are not mutually exclusive. In particular, questions relating to Politics Applied may intersect with several other categories. The Social Studies curriculum encompasses the evolution of people, situations, events, and theories that develop amongst the backdrop of historical, political, and economic contexts. However, for a coding system to be applied, some guidelines need to be provided to assign questions to one category only. Walter (1996) included some rules regarding how to handle questions that pertain to more than one content area. For example, she explained that Peace Initiatives questions may intersect with other categories, such as History, Politics, and Control Strategies, but when the focus of the question related to an attempt to achieve and maintain harmony, Peace Initiatives takes precedence. In addition, questions grouped as Control Tactics must concentrate on the control strategy more than on any other feature (i.e., ideology, history, economics, or politics). Even with some guidelines, Walter (1996) had numerous items double-coded, suggesting that such judgments are not easily made and more guidelines are necessary.

The fourth problem relates to the research hypotheses of Walter and Young (1997) that are associated with the content analysis conducted by Walter (1996). They do not explicitly state their hypotheses about which content area was expected to favour which gender. After careful inspection of their paper and verbal clarification (personal communication with Beth Young, March, 2000), the hypotheses were twofold. First, males were expected to outperform females in the areas of *Economics*, *Politics*, *History*, and *Control Tactics*. Second, females were expected to outperform males in the area of *Peace Initiatives and Internationalism*.

The fifth and most significant problem with the content analysis is that Walter (1996) does not provide empirical support for the content analysis. Although Walter (1996) and Walter and

Young (1997) have a conceptual basis to evaluate the Social Studies Diploma exam items, they do not have any empirical results to support their claims.

Purpose of the Study

The purpose of the study is to demonstrate statistical hypothesis testing with respect to the newer DBF approach within the Social Studies context. DBF research is based on the premise that group differences may be more easily be identified using a theoretical organizing principle. The adaptation of a theoretical organizing principle may prove useful in explaining why items function differentially between males and females. Thus, the importance of the current study is to demonstrate a new statistical approach that may provide evidence to support gender differences in the Social Studies Diploma exams. For our hypotheses, we expect males to outperform females in the content areas of *Economics, Politics, History, and Control Tactics*, and females to outperform males in the area of *Peace Initiatives and Internationalism* based on the hypotheses put forward by Walter (1996) and Walter and Young (1997).

Method

The Alberta Social Studies Grade 12 diploma examinations were studied using the January and June administrations from 1991 and 1992. The exams consist of 70 MC items and one essay question, weighted 70 percent and 30 percent respectively. The MC section tests understanding, terminology, recall of concepts, and synthesis and analysis. Source materials, such as maps, graphs, charts, political cartoons, and excerpts from political speeches and economics texts are included whereby students must analyze and synthesize understanding. The Alberta Ministry of Education classifies the MC questions by knowledge and skills required to answer the question and by topic area. The blueprint for knowledge and skills consists of three areas: Comprehension of Information and Ideas; Interpretation and Analysis of Information and Ideas; and Synthesis and Evaluation of Information and Ideas. There are approximately equal numbers of questions in these three categories. Questions also are equally divided into the two content areas: Political and Economic Systems, and Global Interaction in the Twentieth Century. The two general areas are further divided into five categories. Political and Economic Systems include questions on democracy, dictatorship, mixed economy, private enterprise economy, and

public enterprise economy. Global Interaction in the Twentieth Century includes questions on nationalism, balance of power, confrontation, co-operation, and internationalism.

Walter's (1996) test specifications consisted of five content areas: 1) Economic questions related to economic theory or the application of theory; 2) Political questions related to political theory or the application of theory; 3) History questions related to world wars or to post war incidents that relied on knowledge of people, situations, or events; 4) Peace Initiatives and Internationalism related to attempts by communities to achieve and maintain harmony. Questions pertaining to United Nations, League of Nations, or peace conferences were categorized here; 5) Control Tactics and Strategies related to control strategy rather than the specific details of ideology, or historical, economic, or political fact. Questions pertaining to pre-Second World War times and questions about atomic weapons deployment or economic strategies were categorized here.

Walter (1996) differentiated between Economic Theory and Application, and Political Theory and Application. A clear distinction existed between the theory and application questions for Economics and Politics. The distinction was based on whether students were required to use fact to answer the question (i.e., theory question) or whether students were required to think more abstractly and/or manipulate the material (i.e., application question). Questions pertaining to source materials were classified as application questions. History questions related to superficial knowledge rather than manipulation of information, therefore no distinction was drawn between theory and application for history. In total, seven categories were examined.

The second author conducted inter-rater reliability on one of the four exams (January, 1991). Bakeman and Gottman (1986) argue that percentage agreement for categorical coding is an inappropriate means of assessing reliability because a number of factors may affect the results. Most notably, some agreement would occur by chance alone. Cohen's kappa corrects for this. Kappas of 0.4 to 0.6 are regarded as fair, 0.6 to 0.75 as good, and over 0.75 as excellent (Fleiss, 1981). The inter-rater reliability yielded a Cohen's kappa value of 0.74, indicating an independent coder could replicate the coding system.

Statistically-Based Analyses

The computer program SIBTEST (Stout & Roussos, 1999) was used to determine which item bundles displayed statistically significant DBF. Before any of the bundles could be tested, a strict screening process was implemented. First, all of the DIF items were flagged using SIBTEST for single item DIF analyses. Specifically, each item was tested using the remaining items as the valid subtest. Plots of the beta-uni statistical indices were prepared for the predetermined bundles and were reviewed for groups of items that cluster on the male or female side. The DIF items were classified into three groups according to effect size: A-level (beta-uni statistic < 0.059 , negligible DIF), B-level (beta-uni statistic ≥ 0.059 and < 0.088 , moderate DIF), and C-level (beta-uni statistic ≥ 0.088 , large DIF) (Roussos & Stout, 1996). Next, DBF analyses were performed on the bundles. The valid subtest consisted of the non-bundled items, excluding B- or C-level DIF items obtained from the preliminary single item DIF analyses. Although the bundles may contain B- and C-level DIF items, the valid subtest did not. Hence, the number of items in the valid subtest differed across the categories, depending on the number of items in the bundle. Tables 3 and 4 show the number of items bundled for each category, as well as the number of items composing the valid subtest for the 1991 and 1992 exams respectively. For the 1991 exams, 13 B- and C-level items of the January exam were removed from the valid subtest (i.e., 7 items favored females and 6 items favored males), and 7 B- and C-level items of the June exam were removed from the valid subtest (i.e., 1 item favored females and 6 items favored males). For the 1992 exams, 17 B- and C-level items of the January exam were removed from the valid subtest (i.e., 7 items favored females and 10 items favored males), and 10 B- and C-level items of the June exam were removed from the valid subtest (i.e., 3 items favored females and 7 items favored males).

When a bundle favoured one gender, a reliability check was done by reviewing the next administration within the same year (1991). If the same bundle was significantly flagged across both administrations, a prediction study was employed for the bundles across both administrations in 1992. The prediction aspect was key to our study. If we accurately predicted a bundle of items to favour one group, supporting evidence is provided for the Walter and Young

(1997) theory. However, if the bundles could not withstand our screening and prediction process, the theory may be questioned.

Results

Descriptive statistics for the 1991 and 1992 Social Studies diploma examinations are presented in Tables 3 and 4, respectively. The mean total test scores demonstrate that males performed better than females across the four administrations.

The first stage of the bundle analyses tested the bundles across the January and June administrations of 1991. If a content area did not favor the theoretically chosen group, the direction of the bundle analyses was altered accordingly. Positive beta-unis favored males whereas negative beta-unis favored females. The results are presented in Table 5. The bundles for *Economic Theory* favored females, yielding significant beta-unis of -0.221 for the January exam and -0.114 for the June exam ($p < .001$). The bundles for *History* favored males. Significant beta-unis of 0.544 for the January exam and 0.350 for the June exam were found ($p < .001$). A third bundle, *Control Tactics*, was observed to significantly favor males. When tested across the January and June administrations, significant beta-unis of 0.280 and 0.217 respectively were found ($p < .001$). Significant beta-unis resulted for the content areas of *Economics Applied*, *Political Theory*, *Politics Applied*, and *Peace Initiatives and Internationalism* (-0.164, -0.389, -0.179, and 0.106 respectively) for the January exam only. Beta-uni results for these four areas did not reach significance for the June exam and thus were excluded from the prediction study. In summary, gender differences were consistent across the two administrations for the content areas of *Economic Theory*, *History*, and *Control Tactics*.

The second stage of the bundle analyses involved employing the findings from the 1991 exams to predict the 1992 results. The results are presented in Table 6. The content area of *Economic Theory* favored females across the two 1992 administrations. Significant beta-unis of -0.244 and -0.207, respectively, were found ($p < .001$). The content area of *History* favored males across the two administrations in 1992. The *History* bundle yielded significant beta-unis of 0.279 for the January and 0.541 for the June exams ($p < .001$). For the third bundle, *Control Tactics*, significant beta-unis of 0.209 and 0.293 were found for the January and June exams respectively

($p < .001$). In summary, the *Economic Theory* bundle was found to favor females, and *History* and *Control Tactics* bundles were found to favor males across four administrations of the Social Studies diploma exam

Discussion and Conclusions

The purpose of the study was to demonstrate statistical hypothesis testing with respect to this newer DBF approach within the Social Studies context. The Social Studies diploma exams provided an opportunity to demonstrate the use of DBF, since they have consistently shown gender differences favoring males on the MC component. DBF research is based on the premise that the reasons for group differences on a set of items may more easily be identified using a theoretical organizing principle. The adaptation of a theoretical organizing principle may prove useful in explaining why items function differentially between males and females. The work of Walter (1996) and Walter and Young (1997) provided a theoretical framework for which to classify or bundle the items, which we could test empirically. Their hypotheses specifically included males outperforming females in the areas of *Economics*, *Politics*, *History*, and *Control Tactics*, and females outperforming males in the area of *Peace Initiatives and Internationalism*. The findings suggested that two bundles, *Control Tactics* and *History*, favored males. The bundle of *Economic Theory* rather than *Peace Initiatives and Internationalism* favored females. The results provide some support for the theoretical organizing principle developed by Walter (1996) and Walter and Young (1997) since their theory correctly predicted gender differences favoring males in two of the four content areas, but did not predict the content area favoring females.

On closer inspection, *Peace Initiatives and Internationalism* contained questions that not only alluded to international peace efforts, but other content areas as well. Items were classified on the basis of the inclusion of United Nations, peace conferences, League of Nations, and other international peace efforts. On an overt level, questions were more closely linked to peace than any other content domain. On a covert level, questions coded under *Peace Initiatives and Internationalism* also contained elements of *Control Strategies* and *History*. However, to make the categories mutually exclusive, the mention of peace efforts overrode the mention of other content (Walter, 1996). The inclusion of these other elements may make some of the items in this

category more relevant to males' interests than females' interests. In fact, the results from the January 1991 administration of the diploma exam demonstrated that *Peace Initiatives and Internationalism* significantly favored males. Although the beta-uni value was in the right direction, the results from the June administration indicated that this bundle neither favored males nor females.

The finding that males performed better on the content areas of *Control Tactics* and *History* was predicted. Questions on *History* test knowledge of events, policies, and leaders. Questions on *Control Tactics* address countries' and leaders' efforts to control events, situations, or forces even though the strategies may occur simultaneously with or because of an historical event. Walter (1996) noted that the underlying themes of power relations and control are prevalent in obvious and subtle ways throughout the exam. In addition, males tend to be more interested in history and war-related topics (Alberta Education, Special Study, 1992; Eagley & Carli, 1981), and devote more of their extra curricular activities to pursuing these interests (Alberta Education, Special Study, 1992; Campbell, 1994). Consistent with the finding that males outperform females

principle, then we should be able to predict in a meaningful way which topic areas will display differential group performance in future administrations.

References

- Alberta Education (1992). Special Study: Social Studies 30 Diploma Examination Gender Differences. Internal Report, Student Evaluation Branch. No publication data available.
- Aiken, L. R. (1987). Testing with multiple-choice items. Journal of Research and Development in Education, 20(4), 44-58.
- Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.
- Bakeman, R., & Gottman, J. M. (1986). Observing interaction: An introduction to sequential analysis. New York: Cambridge University Press.
- Barrs, M. (1990). The Primary Language Record: Reflection on issues in evaluation. Language Arts, 67(3), 244-253.
- Berg, B. (1989). Qualitative research methods. Toronto: Allyn and Bacon.
- Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27(2), 165-174.
- Bond, L. (1993). Comments on the O'Neill and McPeck paper. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 277-279). Hillsdale, NJ: Lawrence Erlbaum.
- Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.
- Campbell, J. (1994). Studies related to gender issues in Social Studies. Calgary Board of Education, Calgary, Alberta.
- Christison, W. M. (1995). Sex bias in Social Studies textbooks. Unpublished manuscript.
- Christison, W. M. (1997). Social 30 Performance Differences. Unpublished Doctoral Dissertation.
- Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. Journal of Educational Measurement, 33, 465-484.

Eagley, A. H., & Carli, L. L. (1981). Six researchers and sex-typed communications as determinants of sex differences in influencability: A meta-analysis of social influence studies. Psychological Bulletin, 90, 1-20. ** check title

Englehard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied Measurement in Education, 3, 347-360.

Fleiss, J. L. (1981). Statistical methods for rates and proportions. New York: Wiley.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K. A., & Khaliq, S. N. (2001, April). Using differential bundle functioning to identify and interpret gender differences on science achievement tests. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Kirkland, M. C. (1971). The effects of tests on students and schools. Review of Educational Research, 41, 4, 303-350.

Lafleur, C. & Ireland, D. (1999). Canadian and provincial approaches to learning assessments and educational performance indicators. Technical report submitted to Commonwealth Caribbean Program, Americas Branch: The Canadian International Development Agency.

Messick, S. (1993). Validity. In R. L. Linn (Eds.), Educational Measurement (3rd ed., pp. 13-103). New York: National Council on Measurement in Education. ORYX Press.

Murphy, P. (1988). Gender and assessment. Curriculum, 9, 165-174.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. Journal of Educational Measurement, 30, 293-311.

Neill, D. M. & Medina, N. J. (1989, May). Standardized testing: Harmful to educational health. Phi Delta Kappan, 688-697.

O'Neil, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the validation process. Educational and Psychological Measurement, 40, 397-404.

Rengel, E. (1986, August). Agreement between statistical and judgmental item bias methods. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Rosser, P. (1989). Gender and testing. Report for the National Commission on Testing and Public Policy.

Roussos, L., & Stout, W. (1995). DIF from the multidimensional perspective. Champaign, Illinois: University of Illinois, Department of Statistics.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, 20, 355-371.

Sadker, M., Sadker, D., & Klein, S. (1986). Abolishing misperceptions about sex equity in education. Theory into Practice, XXV(4), 219-226.

Sandoval, J., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249-253.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, 58, 159-194.

Stobart, G., Wood, J., Quinlan, M. (1992). Gender bias in examinations: How equal are the opportunities? British Educational Research Journal, 3, 261-276.

Stout, W., & Roussos, L. (1999). Dimensionality-based DIF/DBF Package [computer program]. William Stout Institute for Measurement: University of Illinois.

Walter, C. (1996). Gender bias in social studies examinations. Master's thesis, University of Alberta, Edmonton, Alberta.

Walter, C., & Young, B. (1997). Gender Bias in Alberta Social Studies 30 Examinations: Cause and Effect. Canadian Social Studies, 31, 83 – 89.

Table 1

Social Studies Diploma Total Examination Marks By Gender Averaged Across January and June Administrations

<u>Year</u>	Males			Females		
	<u>Mean</u>	<u>SD</u>	<u>N</u>	<u>Mean</u>	<u>SD</u>	<u>N</u>
1989-1990	65.10	14.20	8,987	60.40	15.00	9,885
1990-1991	66.70	13.60	9,543	62.70	14.30	10,625
1991-1992	64.80	13.90	9,870	61.00	14.50	10,934
1992-1993	64.30	13.90	9,951	60.60	14.40	11,147
1993-1994	65.80	13.40	10,090	62.00	14.10	11,261
1994-1995	66.10	13.80	9,339	62.60	14.30	10,406
1995-1996	67.10	14.00	9,340	63.10	14.90	10,306
1996-1997	65.50	13.90	9,051	62.10	14.30	10,158
1997-1998	66.00	13.80	9,123	63.00	14.70	9,972
Overall Mean	65.70			61.93		

Source: Alberta Education Diploma Examinations Program Annual Reports.

Table 2

Grade 12 Social Studies School Awarded Marks By Gender Averaged Across the First and Second Semesters

<u>Year</u>	Males			Females		
	<u>Mean</u>	<u>SD</u>	<u>N</u>	<u>Mean</u>	<u>SD</u>	<u>N</u>
1991-1992	67.30	12.20	9,870	67.40	12.50	10,934
1992-1993	58.50	11.20	9,951	63.70	11.10	11,147
1993-1994	67.90	12.30	10,090	67.80	14.10	11,261
1994-1995	68.10	12.50	9,339	68.40	12.30	10,406
1995-1996	68.80	12.30	9,340	68.70	12.20	10,306
1996-1997	69.10	12.20	9,051	69.00	12.10	10,158
1997-1998	69.20	12.20	9,123	69.70	11.90	9,972
Overall Mean	66.90			67.75		

Source: Alberta Education Diploma Examinations Program Annual Reports.

Table 3

Descriptive Statistics for the MC Component of the 1991 Social Studies Diploma Examinations

Characteristic	January		June	
	Males	Females	Males	Females
No. of Examinees	4,252	4,608	5,592	6,430
No. of Items	70	70	70	70
Mean	47.94	43.49	49.68	45.26
Standard Deviation	10.15	10.91	11.31	11.79
Skewness	-0.38	-0.19	-0.57	-0.23
Kurtosis	-0.40	-0.65	-0.29	-0.75

Table 4

Descriptive Statistics for the MC Component of the 1992 Social Studies Diploma Examinations

Characteristic	January		June	
	Males	Females	Males	Females
No. of Examinees	4,229	4,659	5,992	6,711
No. of Items	70	70	70	70
Mean	48.62	44.37	48.04	43.47
Standard Deviation	10.98	11.81	11.05	11.44
Skewness	-0.40	-0.15	-0.41	-0.08
Kurtosis	-0.47	-0.72	-0.42	-0.75

Table 5

Differential Bundle Functioning Results for the 1991 Social Studies Diploma Examinations

Bundle	No. of items	No. of Items in Valid Subtest	\hat{b}_U	Favors
<u>January</u>				
Economic Theory	8	51	-0.221*	Females
Economics Applied	7	52	-0.164*	Females
Political Theory	14	45	-0.389*	Females
Politics Applied	6	53	-0.179*	Females
History	19	41	0.544*	Males
Peace Initiatives and Internationalism	7	50	0.106*	Males
Control Tactics	9	50	0.280*	Males
<u>June</u>				
Economic Theory	8	55	-0.114*	Females
Economics Applied	7	56	-0.064	-
Political Theory	14	49	-0.026	-
Politics Applied	10	54	-0.057	-
History	13	51	0.350*	Males
Peace Initiatives and Internationalism	11	53	0.092	-
Control Tactics	6	60	0.217*	Males

*p<.001

Note. For each bundle, the matching subtest consisted of the remaining items with the exception of items displaying B- and C-level DIF.

Table 6

Differential Bundle Functioning Results for the 1992 Social Studies Diploma Examinations

Bundle	No. of Items	No. of Valid Subtest Items	\hat{b}_U	Favors
<u>January</u>				
Economic Theory	10	44	-0.244*	Females
History	15	43	0.279*	Males
Control Tactics	5	50	0.209*	Males
<u>June</u>				
Economic Theory	10	51	-0.207*	Females
History	18	48	0.541*	Males
Control Tactics	5	56	0.293*	Males

*p<.001

Note. For each bundle, the matching subtest consisted of the remaining items with the exception of items displaying B- and C-level DIF.

Figure

Figure 1. A plot of the MC science items from the 1997 grade 6 Provincial Achievement Test based on the content specifications

Grade 6 1997 Science

