AN EMPIRICAL STUDY OF THE CONSISTENCY OF DIFFERENTIAL ITEM FUNCTIONING DETECTION

by Paulette C. Brown

Thesis submitted to

the School of Graduate Studies and Research

in partial fulfillment of the requirements for the

degree of Master of Arts in Education

ISBN 0-315-93560-X

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ABSTRACT

Total test scores of examinees on any given standardized test are used to provide reliable and objective information regarding the overall performance of the test takers. When the probability of successfully responding to a test item is not the same for examinees at the same ability levels, but from different groups, the item functions differentially in favour of one group over the other group. This type of problem, defined as differential item functioning (DIF), creates a disadvantage for members of certain subgroups of test takers. Test items need to be accurate and valid measures for all groups because test results may be used to make significant decisions which may have an impact on the future opportunities available to test takers. Thus, DIF is an issue of concern in the field of educational measurement.

The purpose of this study was to investigate how well the Mantel-Haenszel (MH) and logistic regression (LR) procedures perform in the identification of items that function differentially across gender groups and regional groups. Research questions to be answered by this study were concerned with three issues: (1) the detection rates for DIF items and items which did not exhibit DIF, (2) the agreement for the MH and LR methods in the detection of DIF items, and (3) the effectiveness of these indices across sample size and over replications.

The test data used in this study came from the Ministry of Education in British Columbia. The grade four provincial reading assessment test (Form M), administered in May, 1988, consisted of a 36 item test which contained a 20 item Literal Comprehension Subtest and a 16 item Inferential Comprehension Subtest. The data were organized into three types of comparison groups: gender groups, gender groups within regions, and regional comparison groups. Thus, a total of 16 comparison groups were analyzed across several sample sizes and over replications for both subtests.

Subtest items flagged by the associated significance tests of the MH chi-square and LR

uniform and nonuniform methods were reviewed and classified as DIF items according to pre-determined standards. Based on the selection criteria for DIF items, both subtests appeared to contain DIF items.

In general, findings indicated that the DIF detection rates across all conditions were low for all three indices. The LR uniform procedure performed slightly better than the MH chi-square procedure in the detection of DIF. All three procedures incorrectly identified approximately 6% of the unbiased items for both subtests. Agreement found between the MH chi-square and LR uniform procedures was similar for both subtests: agreement for the classification of items flagged for DIF was low (5%) while agreement for the classification of items flagged for not exhibiting DIF was very high (92% and 93%); disagreement for the classification of items was very low (3%). A great deal of variability was found in the detection rates of all three indices across sample sizes. Detection rates for large sample sizes were generally high, while detection rates for small sample sizes were low. Thus the statistical indices were found to be unstable in the detection of DIF across sample sizes. The use of samples of 750 examinees per group or more was recommended to obtain fairly stable DIF estimates. The findings of this study support the use of both the MH and LR procedures in the detection of DIF; however, if only one procedure could be selected for application, the LR procedure would be the procedure recommended. In general, the LR procedure is a more powerful tool because it can detect both uniform and nonuniform DIF.

Further research investigating the reliability of the standards used to select DIF items and a replication of this study using different standardized tests, different test lengths, and larger sample sizes would be beneficial.

ACKNOWLEDGEMENTS

Personal achievement is affected by the circle of people around you. I have been surrounded by very special people who have contributed immeasurably to my growth.

I would like to thank Dr. Marvin Boss, my thesis advisor, for being one of those special people. His encouragement, guidance, expertise, and dedication kept me motivated and focused on the completion of this thesis.

I will be forever grateful to the members of my family for being special people. Without the love and support of my spouse, Peter, the completion of this program would not have been possible. Loving thanks to my son, Adam, who willingly cooperated and maturely understood the endless changes in schedules so that deadlines could be met. Without the wisdom and guidance provided by my mother and father, I would not have ventured down this path. They made me realize that dreams can come true if you believe they can.

Finally, I would like to thank Jim Gaskill, Ross Norrington, and Bill Postl, from the Ministry of Education in British Columbia, for their kind cooperation in providing the data for my research.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Testing has been and continues to be an integral part of North American educational systems; as such, the results of tests have a major impact on decisions made regarding future opportunities accessible to students. Therefore, it is extremely important to ensure that tests used to measure student abilities provide equal possibilities for success for diverse cultural groups. When the probability of successfully responding to a test item is not the same for individuals of equal ability who belong to different groups, the item is said to be biased in favour of one group over the other group. Usually the two different groups are referred to as the reference group (RG) and the focal group (FG). The RG represents the group which is used as a standard against which the performance of the FG is compared and the FG represents the group of primary interest (Holland & Thayer, 1986). While the definition of item bias is dependent in part upon the techniques that are used to detect differentially performing items (Baghi & Ferrara, 1990), for purposes of this study, "an item is biased if two individuals with equal ability but from different groups do not have the same probability of success on the item" (Shepard, Camilli, & Averill, 1981, 319). More recently, item bias has been referred to as differential item functioning (DIF) (Holland & Thayer, 1986).

According to Mellenberg (1982), DIF can be further defined as uniform and nonuniform. Uniform DIF exists when there is no interaction between ability level and group membership and the difference in the probability of correctly answering the item is greater for one group than the other group uniformly across all ability levels. Nonuniform DIF exists when there is interaction between ability level and group membership; the difference in the probability of answering the item correctly for the two groups is not the same at all ability levels.

Research focusing on DIF in items used in standardized testing has been and continues to be an issue of concern in the field of educational measurement (Clauser et al., 1991). Researchers

continue to build on and refine existing DIF methodologies in order to devise procedures which will reliably detect DIF so that the cause of DIF can be identified, removed, and items free of DIF can be developed (Scheuneman, 1987).

Important aspects which further clarify the concept of DIF have been delineated by various researchers. Shepard et al. (1981) indicate that DIF for an item cannot be detected in isolation; it is a contextual property that is discovered when an item does not fit the pattern established by other items on the test. Items exhibiting DIF are those that do not function properly and are relatively more difficult for members of a particular group. Techniques using total test score to determine ability cannot detect pervasive DIF, that is, DIF that is exactly the same for all the test items. Rudner et al. (1979) suggest that DIF is a matter of degree and reference to an item being biased or unbiased simply refers to whether the bias in an item is considered to be relatively minimal or substantial. Finally, Shepard et al. (1985) issue a word of caution regarding the use of DIF flags. They suggest that statistical flags are sometimes inaccurate and should be used as indicators that some other factor, other than measured ability, may be influencing performance on the item. Therefore, flagged items should not be automatically discarded, they should be further investigated to determine the source of the difficulty.

DIF detection research is a complex field of study; problems and limitations exist for all DIF detection procedures and the practical application of these approaches in the context of the real world is an additional concern. Hambleton and Cook (1983) indicated that the following assumptions had to be met when using models to investigate DIF: Test items should measure only one trait or ability (the assumption of unidimensionality); It should be stated whether guessing on a test is assumed to take place or not; An examinee's performance on one item should not affect his or her performance on other items (the assumption of local independence). In addition, it is assumed that examinees will answer all test items and that non-speeded tests will be administered. Violations of these assumptions can produce misleading and inaccurate results. In some cases,

results cannot be interpreted. Thus the complexity of issues associated with the phenomenon of DIF and the tools used to measure DIF still remain a challenging responsibility for all who work in the field. Researchers are still refining existing models and developing new ones so that items exhibiting DIF can be detected with confidence.

DIF detection instruments are also of interest to practitioners. Decisions made on the basis of standardized test results may have a significant impact on test takers, educators, test developers, and employers. Therefore, test items which function differentially for groups of equal ability giving one group an advantage over another must be detected prior to the administration of a test. It is absolutely essential that reliable DIF detection tools, which are cost effective and which can be implemented with statistical ease, be available to practitioners. Thus, it is important to test DIF detection instruments using real data because these instruments will be applied in the real world by practitioners.

The purpose of this study is to examine the performance of different DIF detection procedures in the identification of items that do and do not function differentially. More specifically, this study provides answers to the following research questions: How well do the MH and LR procedures perform in the detection of items which exhibit DIF and the items which do not exhibit DIF? Do the MH chi-square (MHC) and LR uniform (LRU) methods show agreement in the identification of items exhibiting DIF? How effective are each of these procedures over sample sizes and replications?

Results of this study may contribute to the existing body of knowledge relating to DIF research by providing empirical results indicating how well the MH and LR procedures perform in the detection of items exhibiting DIF, when using real data. Furthermore, the results may provide information which enable one to have a better understanding of how these indices perform over sample sizes and replications.

In the following chapter, a review of the literature is presented, a description of the most

promising and popular DIF detection procedures and indices is provided, and a review of selected DIF detection studies found in the literature is presented.

CHAPTER II

REVIEW OF THE LITERATURE

Overview of DIF Detection Techniques

DIF detection techniques of interest to this study included those which met the following practical application criteria: (1) Can be used with sample sizes of 1,000 examinees or less; (2) A known test of significance is available to classify items exhibiting DIF and no DIF; (3) Practical implementation is not overly expensive in terms of monetary costs or time; (4) The techniques are not overly complex statistically and can be understood by a lay public.

During the late seventies and early eighties, a variety of DIF detection techniques were compared to determine how effective they were in the detection of items exhibiting DIF and the degree of concordance among the methods. (Merz & Rudner, 1978; Ironson & Subkoviak, 1979; Rudner et al., 1979, 1980; Diamond, 1981; Shepard et al., 1981, 1983, 1985; Berk, 1982; Ironson & Craig, 1982; Subkoviak et al., 1984). Based on the studies reviewed, empirical support was found for the use of item response theory (IRT) models and indices, chi-square procedures, and the transformed item difficulty procedure. The transformed item difficulty approach, also referred to as Angoff's delta, has been criticized because the delta plot may show evidence of DIF spuriously if the item discrimination parameters are not equal or if there is variation in the ability distributions of the groups examined (Shepard et al., 1981; Crocker & Algina, 1986).

While the IRT models and indices are theoretically preferred and empirically supported by the literature for the purpose of DIF detection (Hambleton & Cook, 1977; Ironson & Subkoviak, 1979; Shepard et al., 1981, 1983, 1985; Crocker & Algina, 1986; Hambleton & Rogers, 1988; Baghi & Ferrara, 1989, 1990; Camilli & Smith, 1990), these procedures are costly to implement in terms of computer time, they are statistically complex to analyze and explain, they require sample sizes of 1,000 examinees or more to reach stable parameter estimates, and some of the indices such as

the area method or sum of squares do not have associated tests of significance (Hambleton & Rogers, 1988; Scheuneman, 1990; Ryan, 1991). Therefore, based on the criteria established above, the IRT models and indices are not used in this study.

In the mid to late eighties and early nineties, researchers were still looking for procedures that could be used with smaller sample sizes, that were less costly, and that could be readily explained to and used by practitioners. The literature reviewed revealed a number of articles relating to the MH technique (Holland & Thayer, 1988, 1986; Hambleton & Rogers, 1988; Zwick & Erclkan, 1989; Gutierrez, 1989; Ryan, 1990; DeMauro, 1990; Clauser et al., 1991; Donoghue & Allen, 1991; Fisk, 1991; Gafni, 1991; Mazor et al., 1991). Many researchers considered this chi-square approach to be the best substitute for IRT-based methods (Hambleton & Rogers, 1988; Baghi & Ferrara 1989, 1990; Shermis & St. George, 1990; Camilli & Smith, 1990; Ellis, 1990). The MH procedures are useful because they require smaller sample sizes than the IRT methods, the test statistics are inexpensive to compute, the procedures are easier to understand and implement, and the MHC has a known test of significance. In addition, empirical evidence has been found in the DIF literature which indicates that this method is an adequate substitute for the IRT based models for DIF detection (Hambleton & Rogers, 1988; Mazor et al., 1991; Clauser et al., 1991). A major limitation found with the MH procedure is the inability of the procedure to detect nonuniform DIF.

More recently, a potentially promising DIF detection procedure which is able to identify both uniform and nonuniform DIF and which has been described as an extension of the MH procedure is the LR technique proposed by Rogers and Swaminathan (1990). While this technique is a logistic model, according to Swaminathan and Rogers (1990), it is more cost-effective to implement than IRT-based methods. In addition, this procedure has an associated test of significance. Very few articles relating to LR have appeared in the literature; therefore, further research testing the utility of this technique in detecting items which exhibit DIF seems to be needed.

Thus, this study focuses on the MH and LR DIF detection procedures.

Mantel-Haenszel Procedures

The MH technique uses a chi-square test of significance. According to Holland and Thayer (1986), the MH approach was proposed by Mantel and Haenszel for the study of matched groups in the medical field in 1959. This procedure was adapted by Holland (1985) for DIF identification and further refined by Holland and Thayer (1988). Based on the total test scores, score intervals are formed for RG and FG examinees. A two-by-two contingency table (see Table 1) is used for each score category (j) in order to evaluate the performance of examinees from the RG and FG at the same ability level.

Table I: Two by Two Contingency Table for Matched Reference and Focal Groups at Score Category j (Holland & Thayer, 1986, p. 3)

SCORE ON STUDIED ITEM			
GROUP	CORRECT SCORE = 1	INCORRECT SCORE = 0	TOTAL
RG	A_j	B_j	n_{Rj}
FG	C_j	D_j	n_{Fj}
TOTAL	m_{1j}	m_{0j}	T_j

where A_j and C_j denote the number of examinees in RG and FG who answered correctly;

B_j and D_j denote the number of examinees in RG and FG who answered incorrectly;

m_{1j} and m_{0j} denote the total number of examinees who answered the item correctly and incorrectly, respectively;

n_{Rj} and n_{Fj} denote the number of examinees in the RG and FG respectively;

T_j denotes the total number of examinees in the jth score group (ability level).

A sampling model is required to express the hypothesis for the data in Table 1. According to Holland & Thayer (1986), this is done by acting as though the values of the marginal totals (n_{Rj} and n_{Fj}) for the RG and FG are fixed and thought to be drawn as random samples from the population .

MH Chi-Square

The null hypothesis for the MHC index corresponds to the null hypothesis given as $H_0: P_{Rj} = P_{Fj} \quad j=1,...,k$

where P_{Rj} and P_{Fj} denote the probabilities of success on the item of the RG and FG, respectively, at the jth score category;

j is the ability level.

The MHC statistic is distributed with one degree of freedom and has the form:

$$MHC = \frac{(|\Sigma_j A_j - \Sigma_j E(A_j)| - \frac{1}{2})^2}{\Sigma_j Var(A_j)}$$

where $E_j(A_j) = n_{Rj} m_{1j} / T_j$

and $$Var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

Values of the elements in the equations are taken from a number of contingency tables similar to Table 1. If the MHC statistic is significant, then the item is said to be performing differentially for one of the groups (Baghi & Ferrara, 1989). Gutierrez (1989) indicated that the MHC only flags DIF when the DIF is against or in favour of the same group along the ability scale. DIF is assumed to be uniform; the MH procedure does not readily detect nonuniform DIF.

According to Holland and Thayer (1986), the MHC statistic provides the uniformly most powerful unbiased test of H_0 versus H_1.

MH Alpha

The null and alternative hypotheses for MH alpha (MHα) are as follows:

H_0: $P_{Rj} / Q_{Rj} = P_{Fj} / Q_{Fj}$, $j = 1,..., K$

and H_1: $P_{Rj} / Q_{Rj} = \alpha\, P_{Fj} / Q_{Fj}$ $j = 1,..., K$

where Q_{Rj} and Q_{Fj} denote the probabilities of failure on the items of the RG and FG, respectively, at the jth score category.

For the alternative hypothesis $\alpha \neq 1$.

The estimated MHα is denoted by the following formula:

$$MH\hat{\alpha} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j}.$$

The MHα value is the average factor by which the odds are greater that an examinee from the RG would succeed on the item as compared to an examinee with the same ability from the FG. When the MHα has a value equal to one, no difference is found between the two groups. When the MHα value is less than one the focal group performs better on average than the reference group and for values greater than one the reference group performs better on average than the focal group (Holland & Thayer, 1988).

MH Delta

The MH delta (MHΔ) is used to indicate the amount of DIF for each test item. The MHΔ is transformed from MHα by the following equation:

$$MH\Delta = \frac{-4}{1.7} \ln(MH\alpha) = -2.35 \ln(MH\alpha)$$

Delta values denote the difference in difficulty of an item for the examinees of the reference

group compared to examinees of the focal group within the same ability levels. A MHΔ value of zero is interpreted as an absence of DIF; a negative value indicates that the item is easier for the reference group while a positive value indicates that the item is easier for the focal group. This index does not have a test of significance. MHΔ can also be expressed as MHZ by means of a logarithmic transformation of the MHα: MHZ = -1/1.7 *l*n(MHα). Both MHΔ and MHZ have an expected value of zero and an unknown standard deviation; thus the value of zero represents the absence of DIF.

Holland and Thayer (1986) recommend that the MH statistics be computed in two steps: (1) MH statistics are calculated for all test items and based on the total scores, score groups are selected. Using the MHC test of significance, items found significant exhibit DIF. The MHΔ and MHZ statistics indicate the amount and direction of the difference in item difficulty for the comparison groups; (2) items exhibiting DIF are omitted from the test and the total test scores of the remaining items plus the studied item are computed again. Items identified as significant exhibit DIF. In this study, step two was not conducted because of the short length of the subtests used; by eliminating items exhibiting DIF to purify the matching criterion, the reliability of the test would be questionable.

Logistic Regression Procedures

Rogers and Swaminathan (1990) indicate that the LR model for DIF detection is given by:

$$P(u_{ij} = 1|\theta_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}\theta_{ij})}}{[1 + e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}]}, \quad i = 1, \ldots, n_j, j = 1, 2.$$

where

P predicts the probability of a correct response to item i for group j

U_{ij} is the item score of examinee i in group j; a value of one represents a correct

response and the value of zero represents an incorrect response;

θ_{ij} is the ability of examinee i observed in group j;

β_{oj} is the intercept parameter for group j;

β_{1j} is the slope parameter for group j.

Separate equations are stated for the two different groups of interest to detect DIF. If the LR equation values are the same for the two groups, the LR curves for the two groups are equal and no DIF is present in the item. This is indicated when $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. Uniform DIF may be inferred when the slope parameters of the two groups are equal and the intercept parameters for the two groups differ: $\beta_{11} = \beta_{12}$ and $\beta_{01} \neq \beta_{02}$. Nonuniform DIF may be inferred when the slope parameters for the two groups differ: $\beta_{11} \neq \beta_{12}$. In the refinement of the LR procedure, Swaminathan and Rogers (1990) treated ability as a continuous variable instead of a discrete variable so that both uniform and nonuniform DIF could be identified.

While Swaminathan and Rogers (1990) present a chi-square statistic with two degrees of freedom to test the null and the alternative hypotheses simultaneously, it is also possible to test each of the two logistic regression equations (uniform and nonuniform DIF) separately with two chi-square tests each with one degree of freedom. When the chi-square statistic for either of the two equations equals or exceeds the critical chi-square values, the hypothesis of no DIF is rejected.

Rogers and Swaminathan (1990), pointed out that the major advantage of the LR technique was its completeness. While most DIF detection techniques focused only on the identification of uniform DIF, the LR procedures could identify both uniform and nonuniform DIF. Since the estimation of parameters in the LR procedure is iterative, the researchers indicated that computations would cost three to four times more than those of the MH procedure. They further noted that this would still cost a lot less than IRT procedures.

Studies Examining Detection Rates and the Effects of Sample Size of the MH and LR Procedures

A review of studies which examined the detection rates and/or the effects of sample size of the MH or the LR procedures was conducted. Studies found most relevant included those in which the true state of the items is known. Rogers and Swaminathan (1990), compared the performance of the MH and LR procedures in the detection of uniform and nonuniform DIF. The true state of items was known in this study because items exhibiting uniform and nonuniform DIF were simulated and items with no DIF were generated. Of particular interest to this study, were the effects of sample sizes of 250/250 and 500/500 and test lengths of 40 and 80 items on the power of the MH and LR procedures in the detection of DIF. The results indicated that sample size had a strong effect on the detection rates for both procedures. Detection rates increased by 15% when the sample size increased from 250 to 500 examinees. With samples of 500 and 250 examinees per group, items with uniform DIF were identified with 83% and 68% accuracy, respectively, using MH and 80% and 66% accuracy, respectively, using LR. With samples of 500 and 250 examinees per group, nonuniform DIF was identified with 46% and 35% accuracy, respectively, using MH and 69% and 50% accuracy, respectively, using LR. Test length did not appear to affect detection rate for either procedure. The authors concluded that in the detection of items with uniform DIF both the LR and MH procedures were equally powerful and in the detection of items with nonuniform DIF, the LR procedure was more powerful than the MH procedure.

In another study, Swaminathan and Rogers (1990) examined the effects of sample sizes of 250/250 and 500/500 and test lengths of 40, 60 and 80 items on the power of the MH and LR procedures in DIF detection. Uniform and nonuniform DIF were induced in 20% of the items for each test. Twenty replications were computed for the sample of 500 examinees and the 80 item test to test the power of the MH and LR procedures. The authors found that items with uniform DIF were detected with 100% accuracy for both procedures with samples of 500 examinees per group

across all test lengths. With samples of 250 examinees per group, items with uniform DIF were detected with 75% accuracy for both procedures across all test lengths. The MH procedure was not able to detect nonuniform DIF under any condition. Using the LR procedure, with samples of 500 examinees per group, items with nonuniform DIF were detected with 50% accuracy with a 40 item test, 83% accuracy with a 60 item test and 75% accuracy with an 80 item test. With samples of 250 examinees per group, items with nonuniform DIF were detected with 50% accuracy across all test lengths. The false positive rates, that is, identifying items which have no DIF as exhibiting DIF, was 1% for the MH procedure and 4-6% for the LR procedure over all conditions.

Mazor et al. (1991) investigated the size of sample required to detect items exhibiting DIF using the MH procedure. Items exhibiting DIF were simulated; thus, the true state of the items was known in advance. In this study, tests with 75 items and sample sizes of 2000, 1000, 500, 200 and 100 examinees per group were examined for groups of equal and unequal ability distributions. Two replications of the results were conducted for samples of 200 and 100. For groups with equal and unequal ability distributions, the MH procedure identified items exhibiting DIF with 74% and 64% accuracy, respectively, in samples of 2000 examinees per group, 61% and 58% accuracy, respectively, in samples of 1000 examinees per group, 38% and 31% accuracy, respectively, in samples of 500 examinees per group, 28% and 24% accuracy, respectively, in samples of 200 per group, and 18% and 9% accuracy, respectively, in samples of 100 examinees per group. Thus, using samples of 500 or more with groups of equal ability distributions, the MH procedure produced more accurate results.

One relevant issue investigated in a study conducted by Ryan (1991), focused on the stability of the MHC and MHΔ procedures in detecting DIF across different sample sizes for Black and White test takers. Of the four samples examined for this purpose, RG (White test takers) samples ranged from 1,221 to 1,263 examinees per group and FG (Black test takers) samples ranged from 141 to 179 examinees per group. A baseline study of a large sample, consisting of

5,015 White examinees and 670 Black examinees, was used as the criterion against which the stability of the four small samples was assessed. Pearson product-moment correlation for the MHΔ values and Spearman rank-order correlations for the MHC values were low. Thus, it was suggested that larger sample sizes than those used in the study were required to obtain stable estimates using the MH procedure.

MHC was one of the procedures used by Shermis & St. George (1990), in a study which was conducted to examine the detection of DIF for items for the 50 item Progressive Achievement Test for Mathematics (PAT Math) for 63 Maori and 726 Pakeha students in New Zealand. Due to small sample sizes, the total test scores were divided into three strata of performance: high, medium, and low. The MH chi-square statistic detected four DIF items for the 50 item test. While, this approach was found to be the most conservative one used in the detection of DIF, high agreement appeared to be found for conclusions drawn by the chi-square approach and other methods used.

Camilli and Smith (1990) compared the MHC test with another procedure to determine DIF detection rates for a 30 item mathematics subtest contained within a New Jersey basic skills placement exam. The sample consisted of 1,085 white students and 300 black students. Using both real data and simulated data generated from a three-parameter IRT model, findings indicated that the MHC statistic was able to detect small amounts of DIF in small samples moderately well, and this index was also good at detecting moderate DIF when the discrimination parameter was high.

One of the relevant issues examined in a study by Baghi and Ferrara (1990), concerned the stability of the MHC statistic in the detection of DIF across different sample sizes. Test data used in the study was taken from the Maryland Test of Citizenship Skills which contained 45 multiple choice items. DIF items were defined using a three-parameter IRT technique. Samples of 1,000, 750, 500, and 200 examinees per group were used to assess the stability of the MHC statistic.

Results for this study indicated that the MHC index detected fewer DIF items as the sample sizes decreased; the index detected seven DIF items for the sample of 1,000 and was unable to detect any DIF items for the sample of 200. Researchers concluded that the MHC could be substituted for the three-parameter approach for sample sizes of 750 or more.

Hambleton and Rogers (1988) compared the MH procedure to an IRT method to determine how well the MH statistic would perform in the detection of DIF. Test data (75 test items) and examinee samples of 1,000 Anglo American and Native American examinees per group were taken from the 1982 New Mexico High School Proficiency Exam which assessed life skills for five different areas. The true state of the item was not known; therefore, items detected by the three-parameter model were used as the standard for DIF detection. Findings of this study indicated that the MHC procedure was able to detect DIF with 80% consistency over two separate comparisons. Furthermore, it was suggested that the MH method could be used for DIF detection studies as long as additional measures were taken to detect nonuniform DIF.

Based on the literature reviewed, it appears that very little work has been done on the consistency of the MH and LR DIF detection techniques over different sample sizes. Further research is required to provide more information regarding optimal sample sizes required to obtain consistent DIF detection results when using the MH and LR procedures. While evidence indicates that larger sample sizes will produce more stable results, the question of how large still remains unanswered. Thus, the issue of sample size in DIF detection over replications is examined in this study.

The methodology for this study is described in the next chapter.

CHAPTER III

METHODOLOGY

The methodology used to investigate the detection rates, agreement, and stability of the MHC, LRU, and LR nonuniform (LRN) procedures is presented in this chapter. This chapter is divided into three sections which include a description of the following: the measuring instrument used to obtain measures of performance, the comparison groups formed for observation, and procedures used for the analysis of the data.

The search for an instrument was originally based on three major criteria which a standardized provincial test with demonstrated reliability; test data which included a large aboriginal population; test data which was available by item. Using a standardized provincial test was desirable because it would provide results from a large number of respondents and it would have been previously tested for reliability. Using an aboriginal population was desirable because the comparison of two different cultural groups would be possible. By definition, DIF detection procedures indicate whether an item favours one group over another at the same ability level. This is important because all ethnic groups in Canada should be able to write a test with the knowledge that test items provide equal advantage to all test takers. Contact was made with educational and aboriginal representatives in Ontario, Manitoba, Saskatchewan, the North West Territories, and British Columbia. It was found that standardized provincial tests which were available did not include designators for different ethnic groups, therefore, it was not possible to meet the second criterion. While this was a great disappointment, the Ministry of Education in British Columbia had test data which was designated by region and met the first and third criteria stated previously. This seemed to be the most reasonable compromise in meeting the second criterion.

Measuring Instrument

Provincial assessments within the province of British Columbia have been an annual exercise since 1976 when the Provincial Learning Assessment Program was implemented. The Reading and Written Expression assessments for 1988 were used to test approximately 100,000 British Columbia students in grades 4, 7, and 10 in the areas of reading comprehension, attitudes toward reading, and background characteristics. Test data for this study include item responses from the Form M Achievement Survey. Form M includes a total of 36 multiple-choice items, with four response options, designed to assess the reading comprehension skills of grade four students; twenty test items assess literal comprehension and 16 test items assess inferential comprehension. For literal comprehension items, students were required to read a passage and answer direct questions about it. For the inferential comprehension items, students were required to read a passage and perform more complex operations to answer questions, such as reorganizing information, drawing conclusions, or making predictions, etc.

Stimulus materials for which test items were written included narrative passages (16 items), poetry (5 items), and informative passages (15 items). Test items were developed by teams of resource teachers at the appropriate grade levels. Test items were reviewed and screened for appropriateness, difficulty-level, and interest-level by a number of review panels at various locations in the province. Items were then field tested using groups of 100 students each from grade 5. Grade four students were not included in the field test because all grade four students were required to write the test. The final test for Form M for grade 4 students was tested for internal consistency using the Cronbach Alpha. The reliability coefficient for the total test which consisted of 36 items was 0.88; the reliability coefficients for the 20 literal comprehension items and the 16 inferential comprehension items were 0.82 and .76, respectively. Therefore, it is assumed that the length of the two subtests is adequate because the reliability coefficients for both subtests were acceptable. During development, Form M was not screened statistically for DIF.

In this study, statistics were computed and assessed separately for the two subtests in order to account for the two different kinds of comprehension skills being tested. In addition, the content of the subtests was assessed by review panels throughout the province. Therefore, it is assumed that the two subtests meet the conditions of unidimensionality. The Literal Comprehension Subtest included items 1 to 4, 8 to 10, 16 to 18, 22 to 25, and 29 to 34. The Inferential Comprehension Subtest included items 5 to 7, 11 to 15, 19 to 21, 26 to 28, 35, and 36.

Comparison Groups

Student populations were taken from a dataset containing the responses of 33,809 grade four public school students who completed assessment tests administered throughout the province of British Columbia in May, 1988. Since student respondents came from many diverse locations and ethnic backgrounds, it was desirable to control some of these differences by incorporating them into this study. Thus, different comparison groups were formed to help isolate and explain differences among the student population. The student population was examined on the basis of two cultural factors: gender and regional location. The gender distribution for this population was approximately equal with 17,081 (51%) male students and 16,416 (49%) female students. The Student Assessment Branch of the Ministry of Education for the Province of British Columbia formed five regions for the 76 school districts of British Columbia. These regional categories were created on the basis of the geographic distribution of the student population, ethnicity and other cultural factors. Thus, the regional factor was assessed using these 5 regions: Vancouver Island, the Coast Region, the Lower Mainland Region, the B.C. South Region, and the B.C. North Region. (See Appendices A and B for maps showing the British Columbia school regions and the British Columbia school districts.)

The Lower Mainland Region is very different from the other regions because it covers the smallest geographic area and includes the largest grade four student population (16,429 students),

the second largest number of school districts (16), the greatest number of schools (457), the greatest number of communities (8) with populations over 26,000, and the greatest diversity in ethnic group composition (Japanese, Chinese, East Indians, aboriginal groups, and others). The number of grade four students found in the schools ranged from 21 to 100.

The B.C. South Region covers the second largest geographic area and includes the second largest grade four student population (6,731 students), the largest number of school districts (28), and the second largest number of schools (249). The majority of school districts (18) are found in communities with populations of less than 5,000, and the number of grade four students in the schools ranged from 1 to 60.

The Vancouver Island Region covers the second smallest geographic area and includes the third largest grade four student population (5,658 students), eleven school districts, and the third largest number of schools (180). The majority of school districts (6) are found in communities with populations less than 5,000 while the remaining five districts are found in communities with populations of 6,000 to 100,000. The number of grade four students found in the schools ranged from 1 to 80, and this region is known to be quite prosperous.

The B.C. North Region covers the largest geographic area and includes the second smallest grade four student population (3,172 students), the smallest number of school districts (9), and the second smallest number of schools (134). The majority of school districts (5) are found in communities with populations less than 5,000. The number of grade four students found in the schools ranged from 1 to 60, and this region included the most isolated communities.

The Coast Region is also very distinct from the other regions because it covers the third largest geographic area and includes the smallest grade four student population (1,819 students), the second smallest number of school districts (11), and the smallest number of schools (80). It is the only region which does not have a community with a population larger than 26,000. It includes the smallest number of grade four students in the schools ranging from 1 to 40, and

according to the staff in the Student Assessment Branch in Victoria, British Columbia, this region includes the largest number of aboriginal students.

Thus, it appears that the larger urban regions of the Lower Mainland and Vancouver Island might be considered similar because they have a larger number of grade four students in their schools, these regions include a greater number of cities with populations greater than 26,000, and they have very diverse ethnic populations. The smaller rural regions of the Coast and B.C. North might be also be considered similar because they have fewer schools and smaller numbers of grade four students in their schools, these regions include smaller towns, and they probably have a fewer number of different ethnic groups and a greater number of aboriginal students attending their schools. Differences in the comparison of urban and rural regions might well be expected.

Different comparison groups were used in this study to create a variety of conditions under which the detection rates and effects of sample size for the MH and LR indices could be studied. All comparison groups were created by randomly selecting cases by gender and/or region across six levels of sample size: 1000/1000, 750/750, 500/500, 300/300, 200/200, and 100/100 examinees per group. These sample sizes were selected for investigation because they covered a broad range in order to determine how well the DIF indices would perform. In addition, a greater number of sample sizes were selected because this was thought to reflect the kinds of sample sizes available to practitioners. The 16 comparison groups used in this study are displayed in Table 2, by region, sample size and number of replications.

Procedures

Procedures are presented for the computation of the statistics for the selected methodologies, the standards for the classification of DIF items, and the descriptive statistics.

Table 2

Comparison groups by region, sample size, and over replications.

COMPARISON GROUPS & POPULATIONS	SAMPLE SIZE REPLICATIONS	
	One	Five
1. Gender Groups of Total Population Female (16,416) Male (17,081)		1000, 750, 500, 300 200, 100
2. Gender Groups for Lower Mainland Female (8,038) Male (8,255)		1000, 750, 500, 300 200, 100
3. Gender Groups for B.C. South Female (3,229) Male (3,411)	1000, 750	500, 300, 200, 100
4. Gender Groups for Vancouver Island Female (2,736) Male (2,872)	1000, 750, 500	300, 200, 100
5. Gender Groups for B.C. North Female (1,508) Male (1,641)	1000, 750, 500, 300	200, 100
6. Gender Groups for the Coast Female (905) Male (902)	750, 500, 300	200, 100
7. Lower Mainland & B.C. South (16,429) (6,731)	1000, 750	500, 300, 200, 100
8. Lower Mainland & Vancouver Island (5,658)	1000, 750, 500	300, 200, 100
9. Lower Mainland & B.C. North (3,172)	1000, 750, 500	300, 200, 100
10. Lower Mainland & the Coast (1,819)	1000, 750, 500, 300	200, 100
11. B.C. South & Vancouver Island	1000, 750	500, 300, 200, 100
12. B.C. South & B.C. North	1000, 750, 500	300, 200, 100
13. B.C. South & the Coast	1000, 750, 500, 300	200, 100
14. Vancouver Island & B.C. North	1000, 750, 500	300, 200, 100
15. Vancouver Island & the Coast	1000, 750, 500, 300	200, 100
16. B.C. North & the Coast	1000, 750, 500, 300	200, 100

Note. In some cases the gender group populations do not add up to the total regional populations because the question regarding gender was not completed by the students.

Computation of Statistics

Computations of the MH and LR statistics were performed using programs developed by Ackerman (1987) and Spray (1989), respectively. These programs were used because they were widely available, they were reasonably efficient to use in terms of computer time, and they have been used by other research organizations (the American College Testing Program and the Educational Testing Service). No response to a test item was interpreted as an incorrect answer. The MH, LRU, and LRN chi-square statistics were used to detect items which exhibited DIF at the .05 level of significance. Thus the critical value of 3.841 was used as the cutoff score for flagging DIF items. This level of significance was selected to improve the probability of DIF detection at smaller sample sizes. The MHZ values were used to determine the direction of DIF.

Standards

Items flagged as statistically significant by the MHC, LRU, and LRN procedures were reviewed and classified as DIF items if they met one of the following standards.

Standard One: For comparison groups which were computed over five replications for samples of 1,000, DIF items were defined as those which were flagged three, four or five times out of the five replications.

For example, in the comparison of gender groups for the total population, listed in Table 2, the sample size of 1,000 was used as the standard to determine which items, flagged by the indices, exhibited DIF. If any of the three indices flagged the item three (60%), four (80%), or five (100%) times, over the five replications, the item exhibited DIF.

Standard Two: For comparison groups which were replicated only once at sample sizes of 1,000, 750, and 500, DIF items were defined as those which were flagged by the same index two out of three times over the three sample sizes with one item flagged in the sample of 1,000. DIF items were also defined as those which were flagged by both the MHC and LRU indices for

samples of 1,000.

For example, in the comparison of gender groups for B.C. North, samples of 1,000, 750, 500, and 300 were replicated only once due to the small size of the population. Therefore, items which were flagged by the same statistical procedure at sample size 1,000 and one of the other sample sizes of 750 or 500, were identified as exhibiting DIF.

Standard Three: For comparison groups which were replicated only once at samples of 1,000 and 750, DIF items were defined as those items which were flagged by both the MHC and LRU indices at samples of 1,000 or those items which were flagged by the same index in both sample sizes.

For example, in the comparison of gender groups for B.C. South, samples of 1,000 and 750 were replicated once. Therefore, items which were flagged in both samples of 1,000 and 750 by one procedure, were found to exhibit DIF. Items which were flagged by both the MHC and LRU procedures at the sample size of 1,000, also were found to exhibit DIF.

Descriptive Statistics

Frequencies were tabulated for items which exhibited DIF and for items which did not exhibit DIF, for false positive (FP) items (those which have been flagged as DIF items when they are not biased), and for items which showed agreement and disagreement for the MHC and LRU indices. Descriptive statistics were also tabulated for the test scores of all the comparison groups.

In the next chapter, the results and a discussion of the findings are presented.

CHAPTER IV

RESULTS AND DISCUSSION

The results of this study are presented in two sections. Section one includes the results for the 20 item Literal Comprehension Subtest and section two includes the results for the 16 item Inferential Comprehension Subtest. In both sections, results from the analyses of the MH and LR detection of DIF and false positive items are displayed for gender groups, gender groups within regions, and within regional comparison groups. Next, a summary of the MHC, LRU, and LRN DIF detection rates over all conditions is given. Then the agreement of the MHC and LRU indices in the identification of DIF items is provided. Finally, MH Z values are displayed in order to observe the consistency with which the direction of DIF was found in the results for all groups.

Following the presentation of the results for both subtests, the findings are discussed in relation to the research questions posed in Chapter One.

Section One - Results for the Literal Comprehension Subtest

Results for the Literal Comprehension Subtest are presented in three sections: detection rates for gender groups from the total population, and detection rates for regional comparison groups. This is followed by a summary of the overall DIF identification, the consistency of direction of the MHZ values for DIF items, and agreement of the MHC and LRU indices.

Detection Rates for Gender Groups from the Total Population

The results displaying the detection of DIF and FP items by the MHC, LRU, and LRN procedures for the comparison of gender groups from the total population across sample sizes and over replications are found in Table 3. A detailed explanation of the information in this table is provided to facilitate familiarity with the organization of the results. This is important because all other tables showing the same information for different comparison groups for both the

Table 3

DIF detection rates and FP for the Literal Comprehension Subtest for gender groups across sample sizes and over five replications at the .05 level of significance for the total population.

Sample Size	Percent of DIF Items Identified						Mean Percent		Mean Percent of FP		
	Item 9		Item 30		Item 32						
	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	LRN
1000/1000	80	80	100	100	60	60					
750/750	40	40	60	60	40	40	47	47	8	11	11
500/500	0	20	40	40	20	20	20	27	18	18	1
300/300	20	20	40	40	20	20	27	27	9	12	4
200/200	0	0	0	0	0	0	0	0	5	6	2
100/100	0	0	20	0	0	0	7	0	2	6	8
TOTAL	12	16	32	28	16	16	20	20	8	11	5

Literal Comprehension and Inferential Comprehension Subtests are set up in the same way. Furthermore, all the statistical results displayed in all the tables in this study were found at the .05 level of significance.

An examination of Table 3 indicates that results for each of the three statistical indices are organized under three main headings: (1) the Percent of DIF Items Identified, which includes the results for each specific item found to exhibit DIF for the six levels of sample size; (2) the Mean Percent, which includes the averaged sum of the totals for the DIF items for each index across each sample size; (3) the Mean Percent of FP, which includes the averaged sum of the totals for the items which were falsely identified as DIF when they were not biased, across each sample size. It is important to note that results which were computed over one replication are denoted by an asterisk (*) located in front of the sample size; if there isn't an asterisk in front of the sample size,

the results for the sample size have been replicated five times. Thus, results for all sample sizes in Table 3 have been computed over five replications. Next, the sample size(s) used to establish the standards for the selection of DIF items is highlighted by a big bold box. In this particular case, the sample of 1,000 examinees per group, was used to define DIF items; thus, items which were flagged three (60%), four (80%), or five (100%) times out of five, exhibited DIF. Therefore, items 9, 30, and 32 were found to exhibit DIF because they were flagged 80%, 100%, and 60% of the time, respectively, by the MHC and LRU indices. The LRN index did not detect any DIF items; therefore, results for this index are not found listed under the Percent of DIF Items Identified.

DIF detection rates for the MHC and LRU indices for samples of 750, 500, 300, 200 and 100 are listed below the sample size(s) used as the standard(s). While the overall mean percent of DIF correctly detected by both the MHC and LRU procedures was the same, it can be seen that a greater percentage of DIF was identified by the LRU index for item 9 and the MHC index detected a greater percentage of DIF for item 30. The highest detection rates for both indices were found for item 30 which is consistent with the degree of certainty shown for DIF identification of the item found in the standard.

Table 3 also displays the mean percent of the FP rates for each procedure. The highest FP rates were detected by the LRU index and the lowest by the LRN index. While the MHC and LRU indices detected the same percentage of DIF items, the LRU index had a 3% higher FP rate than the MHC index. This is important because the amount of Type 1 error, that is, items incorrectly identified as DIF items when they are not biased, with which the index is functioning can be compared to the amount of DIF being detected. How much error is acceptable will depend on the purpose of the study, the impact it will have in the circumstances, and the researcher.

The detection rates for both the MHC and LRU statistics decreased with decreasing sample sizes; thus, detection rates for both indices were best for samples of 750 examinees per group and for samples of 200 and 100 examinees per group, both indices performed poorly because they were

unable to detect DIF.

Detection Rates for Gender Groups Within Regions

The results for the MH and LR detection rates for gender groups within regions for the Literal Comprehension Subtest are found in Tables 4 to 8.

Lower Mainland

MHC, LRU, and LRN detection rates for female and male comparison groups within the Lower Mainland are displayed in Table 4. The sample of 1,000 examinees per group was used as the standard to select DIF items: for this data set items 4 and 30 were found to exhibit DIF for both the MHC and LRU indices. The LRN index did not identify any DIF items.

DIF detection rates for the MHC and LRU indices differed, with the LRU index detecting a greater percentage of DIF for both items. The highest and lowest detection rates for both indices were found for items 30 and 4, respectively, which is consistent with the degree of certainty shown for DIF identification of the items found in the standard. Except for the MHC results for item 4, the detection rates for both the procedures showed variability and decreased with decreasing sample sizes. Detection rates for both indices were best for samples of 750 and 500 examinees per group and for samples of 200 and 100 examinees per group, both indices were very poor indicators of DIF.

The FP rate was highest for the LRU index and lowest for the MHC index. While the LRU procedure was able to detect a greater percentage of DIF, this index had a slightly higher error rate than the MHC procedure. It seems that in this instance, one would have to decide which was more important, the technique which had the better detection rate and a higher error rate, or, the procedure with the lower detection rate and a lower error rate.

Table 4

DIF detection rates and FP for the Literal Comprehension Subtest for gender groups from the Lower Mainland.

Sample Size	Percent of DIF Items Identified				Mean Percent		Mean Percent of FP		
	Item 4		Item 30						
	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	LRN
1000/1000	60	60	80	80					
750/750	20	40	100	100	60	70	8	12	12
500/500	20	20	60	60	40	40	10	11	5
300/300	20	20	20	40	20	30	11	11	6
200/200	0	0	20	20	10	10	2	3	3
100/100	0	20	20	20	10	20	0	2	7
TOTAL	12	20	44	48	28	34	6	8	7

B.C. South

MHC, LRU, and LRN detection rates for female and male comparison groups within B.C. South are shown in Table 5. The samples of 1,000 and 750 examinees per group were used as the standard to select DIF items: for this population, items 9, 30, and 34 were found to exhibit DIF for both the MHC and LRU indices; item 3 exhibited DIF for the LRU index; item 29 exhibited DIF for the LRN index.

DIF detection rates for the three items identified by both the MHC and LRU indices were very similar. For these items, the highest and lowest detection rates for both indices were found for items 30 and 9, respectively. Detection rates by the LRU and LRN were very low for items 3 and 29. In general, the detection rates for both the procedures decreased with decreasing sample sizes. Detection rates for the MHC and LRU indices were best for samples of 500 and 300

Table 5

DIF detection rates and FP for the Literal Comprehension Subtest for gender groups from B.C. South.

Sample Size	Percent of DIF Items Identified								Mean Percent			Mean Percent of FP		
	Item 3	Item 9		Item 29	Item 30		Item 34							
	LR U	MH C	LR U	LR N	MH C	LR U	MH C	LR U	MH C	LR U	LR N	MH C	LR U	LR N
*1000/1000	100	100	100	100	100	100	100	100						
*750/750	100	0	0	100	0	0	0	0						
500/500	20	60	60	20	80	80	40	60	60	55	20	6	5	5
300/300	0	20	20	20	60	60	40	40	40	30	20	6	11	6
200/200	40	0	0	0	20	20	0	0	7	15	0	6	4	5
100/100	0	0	0	0	0	0	0	0	0	0	0	5	5	2
TOTAL	15	20	20	10	40	40	20	25	27	25	10	6	6	5

*One replication only.

examinees per group and for samples of 200 and 100 examinees per group, both indices were poor indicators of DIF.

FP rates were the same for the MHC and LRU indices and slightly lower for the LRN index.

Vancouver Island

MHC, LRU, and LRN detection rates for female and male comparison groups within Vancouver Island are shown in Table 6. The samples of 1,000, 750 and 500 examinees per group were used as the standards to select DIF items: items 4, 9, and 30 were found to exhibit DIF for both the MHC and LRU indices and item 16 exhibited DIF for the LRN index.

DIF detection rates for both the MHC and LRU indices were highest for items 30 and 4 and

Table 6

DIF detection rates and FP for the Literal Comprehension Subtest for gender groups from Vancouver Island.

Sample Size	Percent of DIF Items Identified							Mean Percent			Mean Percent of FP		
	Item 4		Item 9		Item 16	Item 30							
	MHC	LRU	MHC	LRU	LRN	MHC	LRU	MHC	LRU	LRN	MHC	LRU	LRN
*1000/1000	100	100	100	100	100	100	100						
*750/750	0	100	100	100	100	0	0						
*500/500	100	100	0	0	0	100	100						
300/300	40	40	0	0	0	60	80	33	40	0	3	8	10
200/200	20	40	0	0	0	60	80	27	40	0	3	3	5
100/100	0	20	20	20	0	20	20	13	20	0	5	6	5
TOTAL	20	33	7	7	0	47	60	24	33	0	4	6	7

*One replication only.

lowest for item 9; The LRU index detected more DIF than the MHC index in items 4 and 30. The LRN was unable to detect DIF for item 16. The detection rates for both procedures were better than expected for samples of 300 and 200 examinees per group. For the sample of 100 examinees per group, both indices were poor indicators of DIF.

The FP rate was highest for the LRN index, and lowest for the MHC index. Again in this data set, the DIF detection procedure with the higher detection rate had the higher error rate.

B.C North

MHC, LRU, and LRN detection rates for female and male comparison groups within B.C. North are displayed in Table 7. The samples of 1,000, 750, and 500 examinees per group were

Table 7

DIF detection rates and FP for the Literal Comprehension Subtest for gender groups from B.C. North.

Sample Size	Percent of DIF Items Identified									Mean Percent			Mean Percent of FP		
	Item 1	Item 4	Item 9		Item 30		Item 32	Item 34							
	LRN	MHC	MHC	LRU	MHC	LRU	LRU	MHC	LRU	MHC	LRU	LRN	MHC	LRU	LRN
*1000/1000	100	100	100	100	100	100	100	100	100						
*750/750	100	100	100	100	100	100	100	100	100						
*500/500	0	0	0		0	0	100	100	100						
*300/300	0	0	100	100	0	0	100	0	0	25	50	0	13	6	0
200/200	0	20	40	40	40	60	40	0	0	25	35	0	8	9	5
100/100	100	0	0	0	0	20	0	0	0	0	5	100	5	8	5
TOTAL	45	9	27	27	18	36	27	0	0	14	23	45	7	8	4

Note. To reflect the appropriate percentages, the totals are calculated for the total number of replications at each sample size.

*One replication only.

used as the standard to select DIF items: for this population, items 9, 30, and 34 were found to exhibit DIF for both the MHC and LRU indices; item 4 exhibited DIF for the MHC index; item 32 exhibited DIF for the LRU index; item 1 exhibited DIF for the LRN index.

The lowest DIF detection rates for both the MHC and LRU indices were found for item 34, and the highest DIF detection rates for the MHC and LRU indices were found for items 9 and 30, respectively. The LRU index performed slightly better than the MHC index in the identification of DIF items for this population. The MHC and LRU procedures performed the best in detecting DIF with

Table 8

DIF detection rates and FP for the Literal Comprehension Subtest for gender groups from the Coast.

Sample Size	Percent of DIF Items Identified			Mean Percent		Mean Percent of FP		
	Item 18	Item 30						
	LRU	MHC	LRU	MHC	LRU	MHC	LRU	LRN
*750/750	100	100	100					
*500/500	100	0	100					
*300/300	100	100	100					
200/200	0	60	60	60	30	7	6	3
100/100	0	40	40	40	20	1	3	7
TOTAL	0	50	50	50	25	4	5	5

*One replication only.

samples of 300. They were poor indicators of DIF in samples of 100.

The FP rates were highest for the LRU procedure, slightly lower for the MHC procedure and lowest for the LRN procedure.

Coast

MHC, LRU, and LRN detection rates for female and male comparison groups within the Coast are shown in Table 8. Due to the limited population size, the largest sample for this group was 750 examinees per group. The samples of 750, 500, and 300 examinees per group were used as the standards to select DIF items: for this population, both the MHC and LRU indices detected DIF for item 30 and the LRU index detected DIF for item 18. The LRN index did not detect any items exhibiting DIF.

DIF detection rates were highest for the MHC and LRU indices for item 30. The LRU index was not able to detect DIF for item 18. The MHC procedure performed better than the LRU procedure in the detection of DIF. Detection rates were best for the sample of 200 examinees per group. The FP rates for the three procedures were very similar and relatively small.

Detection Rates for Regional Comparison Groups

The results for the MH and LR detection rates for the comparison of regional groups for the Literal Comprehension Subtest are displayed in Tables 9 to 18.

Lower Mainland and B.C. South

Results for the groups from the Lower Mainland and B.C. South are shown in Table 9. The MHC and LRU indices detected DIF in items 4, 25, and 32. The LRN index did not detect DIF for items in this data set.

While detection rates were highest for the LRU index, in general, detection rates were very low. In this case, one might find that the standard used to detect DIF is too lax, because the LRU index only detected 15% of the DIF in item 4 and did not detect DIF in the other two items. The MHC index only detected DIF for one replication in the sample of 300. FP detection rates were approximately the same for all three indices.

Lower Mainland and Vancouver Island

The MHC, LRU, and LRN procedures did not detect DIF items for regional groups from the Lower Mainland and Vancouver Island. The FP rates for the three indices were relatively small: both the LRU and LRN indices had the highest rates of 4% and the MHC had the lowest rate of 3%.

Lower Mainland and B.C. North

The MHC, LRU, and LRN procedures did not detect DIF for regional groups from the Lower Mainland and B.C. North. FP rates for the three indices were as follows: the LRU index had the highest rate of 7%, the LRN index had a 5% rate and the MHC index had the lowest rate of 4%.

Table 9

DIF detection rates and FP for the Literal Comprehension Subtest for groups from the Lower Mainland and B.C. South.

Sample Size	Percent of DIF Items Identified						Mean Percent		Mean Percent of FP		
	Item 4		Item 25		Item 32						
	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	LRN
*1000/1000	100	100	100	100	100	100					
*750/750	0	0	0	0	0	0					
500/500	0	20	0	0	0	0	0	7	7	8	5
300/300	20	20	0	0	0	0	7	7	5	5	5
200/200	0	0	0	0	0	0	0	0	2	2	2
100/100	0	20	0	0	0	0	0	7	5	1	6
TOTAL	5	15	0	0	0	0	2	5	5	4	5

* One replication only.

B.C. South and Vancouver Island

The MHC, LRU, and LRN procedures did not detect DIF for regional groups from B.C. South and Vancouver Island. The FP rates for this population were low with the LRU (5%) and LRN (5%) indices having the highest rates and the MHC (3%) the lowest.

B.C. South and B.C. North

The MHC, LRU, and LRN indices did not detect DIF for regional groups from B.C. South and B.C. North. The FP rates were highest for the LRU index (6%) and lowest for the MHC index (4%).

Vancouver Island and the Coast

The MHC, LRU, and LRN indices did not detect DIF for regional groups from Vancouver Island and the Coast. The LRN procedure showed a high FP rate of 8% while both the

Table 10

DIF detection rates and FP for the Literal Comprehension Subtest for groups from the Lower Mainland and the Coast.

Sample Size	Mean Percent	Mean Percent of FP		
	Item 25			
	LRN	MHC	LRU	LRN
*1000/1000	100			
*750/750	100			
*500/500	0			
*300/300	100	0	0	0
200/200	0	10	15	3
100/100	0	4	11	5
TOTAL	9	6	12	4

Note. To reflect the appropriate percentages, the totals are calculated for the total number of replications at each sample size.

*One replication only.

MHC and LRU indices showed rates of 4%.

Lower Mainland and the Coast

Detection rates for groups from the Lower Mainland and the Coast are presented in Table 10. The LRN procedure identified DIF for item 25; DIF was not detected by the MHC or LRU indices.

In this particular case, DIF was only detected for the sample of 300. The FP rate was highest for the LRU and lowest for the MHC. Again, the FP for both the LRU and LRN seemed higher than usual.

Table 11

DIF detection rates and FP for the Literal Comprehension Subtest for groups from B.C. South and the Coast.

Sample Size	Mean Percent		Mean Percent of FP		
	Item 4				
	MHC	LRU	MHC	LRU	LRN
*1000/1000	100	100			
*750/750	0	0			
*500/500	100	100			
*300/300	0	0	5	5	5
200/200	0	0	2	3	4
100/100	20	0	6	9	7
TOTAL	9	0	4	6	5

Note. To reflect the appropriate percentages, the totals are calculated for the total number of replications at each sample size.

*One replication only.

B.C. South and the Coast

Detection rates for regional groups from B.C. South and the Coast are displayed in Table 11. As indicated in Table 11, DIF was detected for item 4 by the MHC and LRU procedures. DIF was not detected by the LRN procedure.

While, the MHC index performed better than the LRU index, only a very small amount of DIF was detected by this procedure. FP rates were highest for the LRU procedure and lowest for the MHC procedure.

Table 12

DIF detection rates and FP for the Literal Comprehension Subtest for groups from Vancouver Island and B.C. North.

Sample Size	Percent of DIF Items Identified						Mean Percent		Mean Percent of FP		
	Item 2		Item 8		Item 23						
	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	LRN
*1000/1000	100	100	100	100	100	100					
*750/750	0	0	100	100	0	0					
*500/500	100	100	0	0	100	100					
300/300	20	20	20	20	20	40	20	27	5	7	8
200/200	0	0	0	0	0	0	0	0	3	6	3
100/100	0	20	0	20	0	0	0	13	3	5	3
TOTAL	7	13	7	13	7	13	7	13	4	6	5

*One replication only.

Vancouver Island and B.C. North

Detection rates for the regional groups from Vancouver Island and B.C. North are presented in Table 12. As shown in Table 12, items 2, 8, and 23 were found to exhibit DIF by both the MHC and LRU procedures. DIF was not detected by the LRN procedure.

While the LRU procedure was the better indicator of DIF, detection rates for both indices were low. The best results were found for samples of 300 for both procedures. The LRU had the highest mean percent for the identification for FP and the MHC index had the lowest.

B.C. North and the Coast

The MHC, LRU, and LRN did not detect DIF for regional groups from B.C. North and the Coast. FP rates were low: the LRU procedure had the highest rate of 4%, the LRN had a

3% rate, and the MHC had the lowest rate of 2%.

Summary of Overall DIF Identification

A summary of the overall detection rates for the MH and LR procedures over all conditions is displayed in Table 13. These results indicate the mean percent of DIF that was detected in the Literal Comprehension Subtest. As indicated in Table 13, the LRU procedure performed slightly better than the MHC procedure in the detection of DIF, however, it also had a higher FP rate than the MHC procedure. The overall results clearly showed that sample size greatly effects the DIF detection rates of the MHC and LRU procedures. Detection rates for both indices were best in the samples of 750 in which the indices detected more than half of the DIF items. Both the MHC and LRU procedures were poor indicators of DIF with samples of 200 and 100. It should be noted that the findings for the LRU procedure are based on two more items than the MHC procedure; the results for the LRN procedure are based on four DIF items.

Finally, the FP rates for all three indices were slightly higher than expected with the LRU procedure having the highest percentage of FP. Swaminathan and Rogers (1990) found that the MH procedure consistently produced 1% FP and the LR procedure produced 4-6% FP at the .01 level of significance.

Consistency of Direction of MHZ Values for DIF Items

An examination of the DIF items exhibiting positive and negative MHZ values was necessary to determine if DIF was consistently found in the same direction for the items. The results of the assessment of the MHZ values showed that there was complete agreement in direction for each replication of each DIF item.

Table 13

Summary of the DIF detection rates and FP for the Literal Comprehension Subtest.

Sample Size	Number of Replications	Mean Percent of DIF			Mean Percent of FP		
		MHC	LRU	LRN	MHC	LRU	LRN
750/750	10	54	59		8	12	12
500/500	20	30	32	20	9	9	4
300/300	33	24	27	17	6	8	5
200/200	50	14	14	0	4	6	5
100/100	50	9	9	25	3	5	5
Total	163	26	28	16	6	8	6

Agreement of MHC and LRU Indices

A summary of the agreement and disagreement of the MHC and LRU indices is shown in Table 14. It can be seen that the MHC and LRU indices agreed 5% and 92% of the time when classifying items as DIF or NO DIF, respectively. The mean percent disagreement was very low. This means that the same decision would result over all replications 92% of the time. It is recognized that the DIF or NO DIF agreement does not mean that the decision was correct. It simply indicates that agreement between the two indices occurred. In other words, the results for agreement and disagreement included all items that both indices flagged or did not flag at the .05 level of significance.

Table 14

Agreement and disagreement of the MHC and LRU indices for the Literal Comprehension Subtest.

COMPARISON GROUPS	Mean Percent Agreement Regarding Classification of Items		Mean Percent Disagreement
	DIF	NO DIF	
Gender Groups for Total Population	11	86	3
Gender Groups for Lower Mainland	10	87	3
Gender Groups for B.C. South	8	90	2
Gender Groups for Vancouver Island	7	89	4
Gender Groups for B.C. North	7	88	5
Gender Groups for the Coast	5	93	2
Lower Mainland & B.C. South	4	94	2
Lower Mainland & Vancouver Island	3	95	2
Lower Mainland & B.C. North	4	92	4
Lower Mainland & the Coast	6	87	7
B.C. South & Vancouver Island	3	95	2
B.C. South & B.C. North	4	94	2
B.C. South & the Coast	4	94	2
Vancouver Island & B.C. North	4	92	4
Vancouver Island & the Coast	3	95	2
B.C. North & the Coast	1	96	3
MEAN PERCENT OF TOTAL	5	92	3

Section Two - Results for the Inferential Comprehension Subtest

Results for the Inferential Comprehension Subtest are presented in three sections: detection rates for gender groups from the total population, and detection rates for regional comparison groups. This is followed by a summary of the overall DIF identification, the consistency of direction of the MHZ values for DIF items, and agreement of the MHC and LRU indices.

Table 15

DIF detection rates and FP for the Inferential Comprehension Subtest for gender groups for the total population.

Sample Size	Mean Percent of DIF Items Identified			Mean Percent of FP		
	Item 13		Item 26			
	MHC	LRU	LRN	MHC	LRU	LRN
1000/1000	80	80	60			
750/750	0	0	20	7	5	8
500/500	40	40	20	4	5	9
300/300	0	0	0	5	8	3
200/200	0	20	20	3	3	4
100/100	0	0	0	4	12	12
TOTAL	8	12	12	5	7	7

Detection Rates for Gender Groups from the Total Population

Detection rates for the MHC, LRU, and LRN procedures for the comparison of gender groups from the total population for the Inferential Comprehension Subtest are displayed in Table 15. As indicated in Table 15, item 13 was found to function differentially for the MHC and LRU indices and items 26 was found to function differentially for the LRN index. The LRU procedure performed 4% better than the MHC procedure in the detection of DIF. The DIF detection rates were found to be very low for all three procedures. The highest rates of DIF were detected by both the MHC and LRU indices with samples of 500. The FP rates were also low for all three indices, however, both the LRU and LRN procedures had higher FP rates than the MHC procedure.

Table 16

DIF detection rates and FP for the Inferential Comprehension Subtest for gender groups from the Lower Mainland.

Sample Size	Mean Percent of DIF	Mean Percent of FP		
	Item 26			
	LRN	MHC	LRU	LRN
1000/1000	60			
750/750	0	13	16	9
500/500	20	6	6	5
300/300	0	4	4	4
200/200	20	3	5	11
100/100	0	1	3	9
TOTAL	8	5	7	8

Detection Rates for Gender Groups Within Regions

Detection rates for the MHC, LRU, and LRN procedures for the comparison of gender groups within the five regions for the Inferential Comprehension Subtest are displayed in Tables 16 to 20.

Lower Mainland

MHC, LRU, and LRN detection rates for gender groups within the Lower Mainland are shown in Table 16. The LRN procedure detected DIF for item 26 while the MHC and LRU procedures were not able to detect DIF for this population. The DIF detection rate for the LRN procedure was very low and DIF was only identified in samples of 500 and 200. FP rates were low with the LRN procedure having the highest rate and the MHC the lowest.

Table 17

DIF detection rates and FP for the Inferential Comprehension Subtest for gender groups from B.C. South.

Sample Size	Mean Percent of DIF			Mean Percent of FP		
	Item 20		Item 35			
	MHC	LRU	LRN	MHC	LRU	LRN
*1000/1000	100	100	100			
*750/750	0	0	100			
*500/500	100	100	0			
300/300	0	20	20	5	9	3
200/200	20	20	0	0	1	5
100/100	0	0	0	4	7	5
TOTAL	7	13	7	3	6	4

*One replication only.

B.C. South

Detection rates for the gender groups in B.C. South are presented in Table 17. It can be seen that both the MHC and LRU techniques detected DIF in item 20 and the LRN technique detected DIF in item 35. In general, detection rates were low with the LRU index having the highest rate. DIF was found in samples of 300 and 200 and DIF was not detected in the sample of 100. FP rates were low with the LRU index showing the highest rate and the MHC showing the lowest.

Vancouver Island

Detection rates for gender groups within Vancouver Island are displayed in Table 18. As indicated in Table 18, the MHC and LRU procedures detected DIF for items 13 and 36; the LRN procedure did not detect DIF for this population. Both the MHC and LRU procedures detected very

Table 18

DIF detection rates and FP for the Inferential Comprehension Subtest for gender groups from Vancouver Island.

Sample Size	Percent of DIF Items Identified				Mean Percent		Mean Percent of FP		
	Item 13		Item 36						
	MHC	LRU	MHC	LRU	MHC	LRU	MHC	LRU	LRN
*1000/1000	100	100	100	100					
*750/750	100	100	100	100					
*500/500	100	100	0	0					
300/300	40	40	40	40	40	40	10	9	5
200/200	40	60	20	20	30	40	4	7	3
100/100	40	40	0	20	20	30	3	4	10
TOTAL	40	47	20	27	30	37	7	7	6

*One replication only.

similar rates of DIF. At samples of 200 for item 13 and samples of 100 for item 36, the LRU index was a better indicator of DIF than MHC index. FP rates were similar for all three indices.

B.C. North

Detection rates for the MHC, LRU, and LRN procedures for gender groups within B.C. North are shown in Table 19. Results indicated that all indices performed very poorly because the DIF detected for items 27 and 20 in the standard were not identified by the procedures. FP rates varied with the LRU index having the highest and the LRN index having the lowest.

Table 19

DIF detection rates and FP for the Inferential Comprehension Subtest for gender groups from B.C. North.

Sample Size	Mean Percent of DIF			Mean Percent of FP		
	Item 27		Item 20			
	MHC	LRU	LRN	MHC	LRU	LRN
*1000/1000	100	100	100			
*750/750	100	100	100			
*500/500	0	0	0			
*300/300	0	0	0	13	13	7
200/200	0	0	0	5	7	3
100/100	0	0	0	3	7	4
TOTAL	0	0	0	5	7	3

Note. To reflect the appropriate percentages, the totals are calculated for the total number of replications at each sample size.

*One replication only.

Coast

Detection rates for gender groups from the Coast are displayed in Table 20. the LRU procedure detected DIF for item 6 and the MHC procedure detected DIF for item 35. Both indices detected DIF with samples of 200, but, did not detect DIF for samples of 100. The MHC procedure performed better than the LRU procedure in the detection of DIF for this population. The FP rates were highest for the LRU procedure and lowest for the MHC procedure.

Table 20

DIF detection rates and FP for the Inferential Comprehension Subtest for gender groups from the Coast.

Sample Size	Percent of DIF Items Identified		Mean Percent		Mean Percent of FP		
	Item 6	Item 35					
	LRU	MHC	MHC	LRU	MHC	LRU	LRN
*750/750	100	100					
*500/500	0	100					
*300/300	100	0					
200/200	20	40	40	20	4	9	5
100/100	0	0	0	0	4	7	8
TOTAL	10	20	20	10	4	8	7

*One replication only.

Detection Rates for Regional Comparison Groups

The results for the MH and LR procedures for the comparison of regional groups for the Inferential Comprehension Subtest are shown in Tables 21 to 24.

Lower Mainland and Vancouver Island

Detection rates for the regional groups from the Lower Mainland and Vancouver Island are displayed in Table 21. According to the standards, the MHC and LRU procedures detected DIF for item 19 and the LRN procedure detected DIF for item 21. As indicated by the results in Table 21, the MHC and LRU were very poor indicators of DIF because they did not detect DIF in item 19 for any sample size. In addition, the LRN procedure detected DIF for item 21, but, this was only for the samples of 300. The FP rate was highest for the MHC and LRU procedures and lowest for the LRN

Table 21

DIF detection rates and FP for the Inferential Comprehension Subtest for groups from the Lower Mainland and Vancouver Island.

Sample Size	Mean Percent of DIF			Mean Percent of FP		
	Item 19		Item 21			
	MHC	LRU	LRN	MHC	LRU	LRN
*1000/1000	100	100	100			
*750/750	100	100	0			
*500/500	0	0	100			
300/300	0	0	100	11	11	3
200/200	0	0	0	5	4	5
100/100	0	0	0	4	5	7
TOTAL	0	0	33	7	7	5

*One replication only.

procedure.

B.C. South and B.C. North

Detection rates for regional groups from B.C. South and B.C. North are presented in Table 22. In accordance with the standards, item 20 was found to exhibit DIF for the LRU procedure. The MHC and LRN procedures did not detect DIF for this population. DIF detection rates by the LRU index were low with equal rates of DIF being identified for samples of 300 and 200. The LRU index did not detect DIF for the sample of 100. FP rates were low for all three indices with the LRU index having the highest rate and the MHC index having the lowest.

Table 22

DIF detection rates and FP for the Inferential Comprehension Subtest for groups from the B.C. South and B.C. North.

Sample Size	Mean Percent of DIF	Mean Percent of FP		
	Item 20			
	LRU	MHC	LRU	LRN
*1000/1000	100			
*750/750	0			
*500/500	100			
300/300	20	3	3	6
200/200	20	3	4	3
100/100	0	1	5	3
TOTAL	13	2	4	4

*One replication only.

B.C. South and the Coast

Detection rates for regional groups from B.C. South and the Coast are displayed in Table 23. In accordance with the standards, the MHC and LRU indices detected DIF for items 12 and 28 and the LRN index identified DIF for item 36. In general, the LRU index was a slightly better indicator of DIF than the MHC index with the LRU index identifying 4% more DIF. The detection of DIF was the same for the MHC and LRU procedures for item 12 for samples of 300 and 100, however, the LRU procedure detected DIF for item 28 for the sample of 300 examinees per group, whereas, the MHC procedure did not. The LRN procedure performed poorly as an indicator of DIF because the index only detected a low rate of DIF for the sample of 200. The results showed that the LRU procedure had a 3% higher FP rate than the MHC and LRN procedures.

Table 23

DIF detection rates and FP for the Inferential Comprehension Subtest for groups from B.C. South and the Coast.

Sample Size	Percent of DIF Items Identified					Mean Percent			Mean Percent of FP		
	Item 12		Item 28		Item 36						
	MHC	LRU	MHC	LRU	LRN	MHC	LRU	LRN	MHC	LRU	LRN
*1000/1000	100	100	100	100	100						
*750/750	100	100	100	100	100						
*500/500	100	100	100	100	0						
*300/300	100	100	0	100	0	50	100	0	14	14	0
200/200	0	0	20	20	20	10	10	20	6	6	5
100/100	20	20	0	0	0	10	10	0	0	6	4
TOTAL	18	18	9	18	18	14	18	9	4	7	4

Note. To reflect the appropriate percentages, the totals are calculated for the total number of replications at each sample size.

*One replication only.

B.C. North and the Coast

Detection rates for regional groups from B.C. North and the Coast are presented in Table 24. As shown in Table 24, the MHC and LRU procedures detected DIF for item 12. The LRN procedure did not detect DIF for this population. Detection rates for the MHC and LRU procedures were low and the LRU performed somewhat better than the MHC statistic by identifying 9% more DIF. The highest FP rate was detected by the LRU index and the lowest by the MHC index.

Table 24

DIF detection rates and FP for the Inferential Comprehension Subtest for groups from B.C. North and the Coast.

Sample Size	Mean Percent Item 12		Mean Percent of FP		
	MHC	LRU	MHC	LRU	LRN
*1000/1000	100	100			
*750/750	100	100			
*500/500	0	0			
*300/300	0	0	7	7	6
200/200	20	20	4	5	5
100/100	20	40	5	9	6
TOTAL	18	27	5	7	6

Note. To reflect the appropriate percentages, the totals are calculated for the total number of replications at each sample size.

*One replication only.

Unexpected Findings for Six Regional Groups

While differences among the comparison of urban and rural regional groups were expected, the spuriously high DIF detection rates found for six of the ten regional comparisons for the Inferential Comprehension Subtest, were unexpected. Since an excessive number of DIF items was found for the comparison of the urban and rural groups, these results were not included in this study. These six comparison groups and the number of DIF items that each identified are listed below:

for groups from Vancouver Island & the Coast, 14 DIF items were detected;

for groups from Vancouver Island & B.C. North, 15 DIF items were detected;

for groups from Vancouver Island & B.C. South, 16 DIF items were detected;

for groups from the Lower Mainland & the Coast, 13 DIF items were detected;

for groups from the Lower Mainland & B.C. North, 12 DIF items were detected;

for groups from the Lower Mainland & B.C. South, 16 DIF items were detected.

Thus, on average, DIF was detected in 14 of the 16 items (88%) for each of the six regional comparison groups.

Possible explanations for the extreme results obtained by the six regional comparison groups were considered. An examination of the signs of the Z values for the DIF items indicated that the responses provided by the students of these six groups were distributed equally for both groups, therefore, the items on the subtest did not appear to give an unfair advantage to one group over the other. In addition, a review of the mean test scores, displayed in Table 25, for the regional groups suggested that regional groups were similar in mean ability. Only a 3% difference was found between the mean scores of any of the regional groups.

Attributing the excessive detection rates to the short length of the subtest cannot explain why these results were found for 60% of the regional groups and not for the remaining 40% of the groups whose performance was not affected by the length of the test. The large number of DIF items cannot be satisfactorily explained by examinees who did not respond to questions because the number of items not answered by examinees was greater in the Literal Comprehension Subtest (No response for 1.7% of items) than the Inferential Comprehension Subtest (No response for 1.3% of the items).

The excessive number of DIF items found for the six regional comparison groups appears to be related to urban and rural differences among the groups. As was previously stated in Chapter Three, both the Lower Mainland and Vancouver Island were urban regions which were considerably different from the Coast, B.C. North and B.C. South Regions: they included numerous

Table 25

The distribution of the mean and standard deviation scores for regional groups for the Inferential Comprehension Subtest.

Regional Group	Mean Score	Standard Deviation
Lower Mainland	10.791	3.229
B.C. South	10.872	3.264
Vancouver Island	11.021	3.217
B.C. North	10.682	3.172
Coast	11.217	3.446

urban centres which were highly populated with diverse ethnic populations. The Coast, B.C. North, and B.C. South Regions were found to be similar because they covered larger geographic areas and the populations were more spread out and isolated. Communities were smaller for these regions and throughout these areas larger pockets of aboriginal groups might be found. According to staff of the Ministry of Education in British Columbia, the Coast Region included a large number of aboriginal students.

According to the Technical Report (Jeroski, 1989) which provided the results of the 1988 student assessments, 49%, 9%, 6%, and 5% of the grade four students indicated that they used different readers that were listed on the questionnaire, 17% of the grade four students indicated that they used readers that were not on the list, and 16% of the grade four students indicated that they did not use a reader at all. This may be an indication that the curriculum taught by teachers in rural areas may not be the same as the curriculum taught by teachers in the urban centres for the development of inferential comprehension skills. In addition, the teaching methodologies used by the urban and rural teachers may have had an impact on the performance of the students. It would be interesting to find out if the different readers and the teaching methodologies used in the classrooms were consistently used by urban and rural teachers.

An examination of the 36 item test showed that the five reading passages included on the test were about animals or fish and one was about the creator of Garfield. It was also noted that 15 of the 16 inferential test items were placed at the end of the series of the questions for each reading passage. This may have had an impact on the way the students responded to the questions; by the end of the passage they may have tired of the content area. In addition, the inferential comprehension items were considered the more difficult questions because they required students to use more sophisticated reading and thinking skills. This indeed was reflected in the mean performance on the test items; the student population had a mean performance of 74% on the Literal Comprehension Subtest and a mean performance of 68% on the Inferential Comprehension Subtest. Inferential items 19 and 35 were found most difficult with 25% and 34%, respectively, of the students answering the questions correctly.

Perhaps these unexpected findings can be attributed to any number of different combinations of the factors which have been previously considered. Further investigation of the urban and rural comparison groups would prove insightful and worthwhile.

Summary of Overall DIF Identification

A summary of the overall detection rates for the MH and LR procedures for the Inferential Comprehension Subtest is displayed in Table 26. The detection rates for this subtest over all conditions were very low. The LRU procedure performed slightly better than the MHC procedure in the detection of DIF and this index also showed very similar FP rates when compared to the MHC and LRN indices. The overall results indicated that sample size does effect the DIF detection rates of the MHC and LRU procedures. While DIF was not detected by the MHC and LRU indices for samples of 750, it must be kept in mind that this sample size was only used in two of the comparison groups for the detection of DIF. Therefore this finding is not surprising. While, both the MHC and LRU indices performed best in the detection of DIF for samples of 500, it should be noted that only two comparisons were made at this sample size. Both indices performed poorly in

Table 26

Summary of the DIF detection rates and FP for the Inferential Comprehension Subtest.

Sample Size	Number of Replications	Mean Percent of DIF			Mean Percent of FP		
		MHC	LRU	LRN	MHC	LRU	LRN
750/750	10	0	0	10	10	11	9
500/500	10	40	40	20	5	6	7
300/300	33	17	26	5	6	8	4
200/200	50	16	18	10	4	5	5
100/100	50	8	11	0	3	7	7
Total	153	16	19	9	6	7	6

the detection of DIF for samples of 200 and 100. Overall, the LRN procedure performed poorly in the detection of DIF across all samples.

FP rates were similar for all indices with the LRU procedure having a slightly higher rate. FP rates for all three indices were slightly higher than expected; at the .05 level of significance one would expect a 5% possibility of error.

Consistency of Direction of MHZ Values for DIF Items

An examination of the DIF items exhibiting positive and negative MHZ values was necessary to determine if DIF was consistently found in the same direction for the items. The results of the assessment of the MHZ values showed that there was complete agreement in direction for each replication of each DIF item.

Agreement of MHC and LRU Indices

A summary of the agreement and disagreement of the MHC and LRU indices is shown in Table 27. It can be seen that the MHC and LRU indices agreed 5% and 93% of the time when classifying items as DIF or NO DIF, respectively. The mean percent disagreement was very low. This means the same decision would result over all replications 93% of the time. It is recognized

Table 27

Agreement and disagreement of the MHC and LRU indices for the Inferential Comprehension Subtest.

COMPARISON GROUPS	Mean Percent Agreement Regarding Classification of Items		Mean Percent Disagreement
	DIF	NO DIF	
Gender Groups for Total Population	6	91	3
Gender Groups for Lower Mainland	6	92	2
Gender Groups for B.C. South	3	95	2
Gender Groups for Vancouver Island	8	89	3
Gender Groups for B.C. North	3	94	3
Gender Groups for the Coast	5	92	3
Lower Mainland & Vancouver Island	6	94	0
B.C. South & B.C. North	2	95	3
B.C. South & the Coast	3	94	3
B.C. North & the Coast	5	90	5
MEAN PERCENT OF TOTAL	5	93	3

that the DIF or NO DIF agreement does not mean that the decision was correct. It simply indicates that agreement between the two indices occurred.

Discussion of Overall Findings

Research questions posed in Chapter One of this study were concerned with the degree of DIF exhibited by test items, how well the MH and LR procedures performed in the detection of DIF, the extent of the agreement between the MHC and LRU procedures in the identification of DIF, and the stability of these procedures across different sample sizes and over replications. The overall findings are discussed in relation to these questions.

Detection Rates for the MH and LR Procedures

The mean percent of DIF detected by the MHC, LRU, and LRN procedures followed similar patterns in both the Literal Comprehension and the Inferential Comprehension Subtests, however, detection rates for all three indices were higher for the Literal Comprehension Subtest. This difference may be a result of the total number of comparisons made for each subtest: the results for 16 comparison groups were included for the Literal Comprehension Subtest while the results for only 10 comparison groups were included for the Inferential Comprehension Subtest. The MHC, LRU, and LRN procedures detected 10%, 9%, and 7% more of the DIF items, respectively, for the Literal Comprehension Subtest than the Inferential Comprehension Subtest. The LRU procedure performed slightly better than the other two indices in the detection of DIF for both subtests.

The detection rates for all procedures across all conditions were lower than expected. These findings may be explained by the procedures used to define the DIF detection standards in this study. Larger sample sizes were used to set the standards for the identification of DIF items: for 14 of the 16 comparison groups, samples of 1,000 and 750 were used as standards to detect DIF and for 10 of these groups, samples of 500 were also used as standards to detect DIF. Thus the majority of the DIF detection results were based on samples of 300, 200, and 100. While the patterns of DIF detection for smaller sample sizes might be similar to those found by larger sample sizes, there is little doubt that larger sample sizes would have provided higher detection rates than those provided by the smaller sample sizes.

It was interesting to note that DIF was not detected in six of the regional comparison groups for the Literal Comprehension Subtest: these groups included the Lower Mainland and Vancouver Island, the Lower Mainland and B.C. North, B.C. South and Vancouver Island, B.C. South and B.C. North, Vancouver Island and the Coast, and B.C. North and the Coast.

The FP rates for the MHC and LRN procedures were the same for the two subtests, while the FP rates for the LRU procedure were slightly higher for both subtests. FP rates for MHC were

higher than expected while the LRU and LRN were similar to those found in a previous study.

Agreement Between the MHC and LRU Procedures

In general, agreement between the MHC and LRU procedures in the classification of items for both the Literal Comprehension and Inferential Comprehension Subtests was almost identical; agreement for the classification of items consistently flagged for DIF was low while agreement for the classification of items consistently flagged for not exhibiting DIF was very high. Disagreement for the classification of items by both procedures was very low in both subtests. Keep in mind that the items flagged for exhibiting DIF and not exhibiting DIF included all items showing agreement, in this particular instance, the use of standards to define DIF were not taken into consideration.

Stability of the MH and LR Procedures Across Sample Sizes and Over Replications

The DIF detection rates of the MHC, LRU, and LRN procedures varied considerably across sample sizes for the two subtests; in general, the MHC and LRU procedures showed that a decreasing amount of DIF was detected as the sample size decreased. The MHC and LRU indices showed higher DIF detection rates for samples of 750 and 500 for the Literal Comprehension Subtest and higher DIF detection rates for samples of 500 and 300 for the Inferential Comprehension Subtest. Detection rates for both the MHC and LRU indices were very low for samples of 200 and 100 for both subtests; samples of this size should be used with caution. The results for the LRN index showed that detection rates varied for both subtests without following a particular pattern; detection rates were best for the samples of 100 for the Literal Comprehension Subtest and for the samples of 500 for the Inferential Comprehension Subtest. Lowest DIF detection rates for the LRN index were found for the samples of 200 for the Literal Comprehension Subtest and the samples of 100 for the Inferential Comprehension Subtest. It should be kept in mind that the findings for the LRN index are based on the detection of three DIF items for the Literal Comprehension Subtest and four items for the Inferential Comprehension Subtest.

Overall, the LRU procedure performed slightly better than or as well as the MHC procedure in the detection of DIF across all sample sizes.

Based on the findings, it appears that samples of 750 or more should be used to obtain stable DIF estimates which identify 50% or more of the DIF items.

In the final chapter, a summary of the findings, limitations of this study and suggestions for future research are discussed.

CHAPTER V

SUMMARY AND CONCLUSIONS

The MHC, LRU, and LRN procedures were compared in this study to answer three research questions: (1) How well do these procedures perform in the detection of items which exhibited DIF and items which did not exhibit DIF? (2) Do the MHC and LRU methods show agreement in the identification of DIF items? and, (3) How effective are each of these indices over sample size and replications? The answers to these questions are discussed in a summary of the findings followed by a discussion of the limitations of this study and suggestions for future research and implications for practitioners.

Summary of Findings

Test data from a grade four British Columbia provincial reading test, consisting of a 20 item Literal Comprehension Subtest and a 16 item Inferential Comprehension Subtest, were analyzed to determine the detection rates of the indices, the agreement and stability for the three indices over sample sizes and replications. Subtest items flagged by the associated significance tests of the MHC, LRU, and LRN methods were reviewed and classified as DIF items according to the pre-determined standards. Based on the selection criteria for DIF items, both subtests appeared to contain DIF items.

The DIF detection rates, over all conditions, were low for the MHC, LRU, and LRN indices: 26%, 28%, and 16% of the DIF items, respectively, were detected for the Literal Comprehension Subtest, while 16%, 19%, and 9% of the DIF items, respectively, were detected for the Inferential Comprehension Subtest. The LRU procedure performed slightly better than the MHC and the LRN procedures in the detection of DIF items for both subtests. The MHC and LRN indices had FP rates of approximately 6% for both subtests while the LRU index showed slightly higher rates of 7% and 8% for the subtests. The FP rates for all three indices were slightly higher than expected.

Agreement between the MHC and LRU indices in the classification of items consistently flagged for exhibiting DIF or not exhibiting DIF was almost identical for both subtests: agreement in the classification of items flagged for DIF was 5% for both procedures and agreement in the classification of items flagged for not exhibiting DIF was 92% and 93% for the Literal and Inferential Comprehension Subtests, respectively. Disagreement in the classification of items was very low (3%) for both subtests.

In assessing the stability of the indices in the detection of DIF across sample sizes, a great deal of variability was found. Thus, all the statistical indices displayed considerable instability in the detection of DIF items across sample sizes. This supports the findings of previous research (Hambleton & Rogers, 1988). Overall, there was a tendency for the indices to detect a decreasing amount of DIF as sample sizes decreased. This was expected and supports the findings of previous research (Mazor et al., 1991; Baghi & Ferrara, 1990). Detection rates were highest for both the LRU and MHC indices, 59% and 54%, respectively, for samples of 750 for the Literal Comprehension Subtest. The use of samples of 200 and 100 when using the MH and LR procedures is not recommended. Based on the findings of this study, the use of samples of 750 are recommended in order to obtain fairly stable DIF estimates in DIF detection studies.

Spuriously high DIF detection rates were found in six regional comparison groups for the Inferential Comprehension Subtest. While it appeared that the problem might be related to urban and rural differences in regional groups, further investigation of this finding would be desirable.

In conclusion, the results of this study confirm the findings of previous studies and support the use of the MH and LR procedures for the detection of DIF. While both methods appeared to produce similar results in the detection of uniform DIF, the LR procedure has the added advantage of having an index to detect nonuniform DIF.

Limitations of the Study

One of the major limitations of this study is the use of real data. By using real data, the amount of DIF for each item on the subtests is unknown. Therefore, it is assumed that the items flagged by the indices and classified by the established DIF criteria, do exhibit DIF.

This leads one to ask the following questions which also create limitations for a study of this nature: What standards were used to determine when an item exhibited DIF or NO DIF? Were these standards adequate, too lax, or too stringent? Certainly, this is of prime importance since the standards established will have a major impact on the type of results obtained. This was indeed the case for the findings of this study. By using the larger sample sizes to set the DIF standards for each comparison group, the DIF detection rates for the indices appeared to be lower than they might have been because the majority of the findings were generated for samples of 300 or less.

Using two subtests which were relatively short in length may have had an impact on the results for the smaller sample sizes of 300 or less.

Identification of the ethnic origins of the students taking these subtests would have been beneficial since it would have provided more precise information regarding the composition of the students in the regional groups. Thus, a clearer delineation of the regional comparison groups might have made it possible to pinpoint and explain differences found in these groups.

The findings found in this study apply to the specific tests used in this study and generalization may be limited; however, it should be kept in mind that the effect of sample size found in this study has been found elsewhere.

Suggestions for Future Research and Implications for Practitioners

There is a need for future studies to investigate the reliability of standards used to define DIF items. Determining what standards provide optimal results for the statistical indices would certainly give researchers more confidence in the results of the study.

A replication of the present study using data from different standardized tests, longer test lengths, and larger sample sizes for the detection of DIF would be beneficial in finding support for the findings of this study. This would provide evidence for the use of these indices over all sample sizes.

Based on the results of this study, the following suggestions are provided for practitioners who are involved with DIF detection investigations: 1) The use of large samples (at least 750) is recommended in order to obtain fairly stable DIF estimates; 2) The use of longer tests (at least 20 items) will provide a better measure of the results; 3) How the standards are set up to determine which items exhibit DIF will have a major impact on the results; therefore, this area should be given serious deliberation during the research design; 4) While this study supports the use of both the MH and LR DIF detection indices in certain circumstances, the use of the LR procedure is recommended because this procedure measures both uniform and nonuniform DIF. For those who use the MH procedure, measures should be taken to detect nonuniform DIF.

There are a number of important generic questions which need to be addressed by educational practitioners in relation to DIF detection practices. If standardized tests are used by provincial educational ministries in Canada, do assessment units of these provinces conduct DIF detection investigations for these tests? Should this be a standard operational procedure? Will this make a meaningful difference in the results obtained in the testing process? What role should university measurement specialists play in this process? Finally, would it be possible to convince provinces who administer standardized tests to include ethnic group designators on these tests?

The findings from this study would indicate that the use of both the MH and LR procedures in the detection of DIF in future studies would definitely be worthwhile.

REFERENCES

Baghi, H., & Ferrara, S. (March, 1989). A comparison of IRT, delta plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations in the Maryland Test of Citizenship Skills. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Baghi, H., & Ferrara, S. (February, 1990). Detecting differential item functioning using IRT and Mantel-Haenszel techniques: Implementing procedures and comparing results. Paper presented at the annual conference of the Eastern Educational Research Association, Clearwater, FL.

Berk, R.A. (Ed.). (1982). Handbook of Methods for Detecting Test Bias. Baltimore, Maryland: The Johns Hopkins University Press.

Camilli, G., & Smith, J.K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a Jackknife test for detecting biased items. Journal of Educational Statistics, 15(1), 53-67.

Clauser, B., Mazor, K.M., & Hambleton, R.K. (March, 1991). Examination of various influences on the Mantel-Haenszel statistic. Paper presented at the meeting of the American Educational Research Association, Chicago.

Clauser, B., Mazor, K.M., & Hambleton, R.K. (March, 1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. Paper presented at the meeting of the American Educational Research Association.

Crocker, L., & Algina, J. (1986). Introduction to Classical & Modern Test Theory. Orlando, FL: Holt, Rhinehart and Winston, Inc.

DoMauro, G.E. (April, 1990). Effects of representation of gender groups in the examinee population on the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Ellis, B.B. (1990). Assessing intelligence cross-nationally: A case for differential item functioning detection. Intelligence, 14(1), 61-78.

Diamond, E.E. (April, 1981). Item bias issues: Background, problems, and where are we today? Paper presented at the annual meeting of the American Educational Research Association, CA.

Donoghue, J.R., & Allen, N.L. (April, 1991). "Thin" versus "thick" matching in the Mantel-Haenszel procedure for detecting DIF. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Fisk, Y.H. (January, 1991). A brief overview of three classes of methods for detecting item bias. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, CA.

Gafni, N. (April, 1991). Differential item functioning: Performance by sex on reading comprehension tests. Paper presented at the annual meeting of the Academic Committee for Research on Language Testing, Jerusalem, Israel.

Gutierrez, J. (1989). Characteristics of the distribution of the Mantel-Haenszel delta under different conditions of the null hypothesis: A Monte Carlo study. Unpublished doctoral dissertation, University of Ottawa.

Hambleton, R.K., & Cook,L.L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14(2), 75-96.

Hambleton, R.K., & Rogers, H.J. (May, 1988). Detecting biased test items: Comparison of the IRT area and Mantel-Haenszel methods. Paper presented at the annual meeting of AERA, New Orleans, LA.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), Test Validity (pp. 129-144). New Jersey: Lawrence Erlbaum Associates, Inc.

Holland, P.W., & Thayer, D.T. (August, 1986). Differential item functioning and the Mantel-Haenszel procedure. Program Statistics Research (Technical Report No. 86-69), Princeton, NJ, Educational Testing Service.

Ironson, G.H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16(4), 209-225.

Ironson, G.H., & Craig, R. (1982). Item bias techniques when amount of bias is varied and score differences between groups are present. Final Report, National Institute of Education, Grant NIE-G-81-0045, Washington, DC.

Jeroski, S. (1989). The 1988 British Columbia assessment of reading and written expression: Technical report. Victoria, BC, Student Assessment Branch, Ministry of Education.

Mazor, K.M., Clauser, B.E., & Hambleton, R.K. (April, 1991). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Mellenberg, G.J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Merz, W.R., & Rudner, L.M. (March, 1978). Bias in testing: A presentation of selected methods. Paper presented at the annual meeting of the American Educational Research Association,Toronto, Ontario.

Rogers, H.J., & Swaminathan, H. (April, 1990). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Paper presented at the annual meeting of AERA, Boston, MA.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1979). The problem of item bias: A comparison of techniques. Final Report, National Institute of Education, Grant NIE-G-78-0084, Washington,DC.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). Biased item detection techniques. Journal of Educational Statistics. 5(3), 213-233.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17(1), 1-10.

Ryan, K.E. (April, 1990). The performance of the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Ryan, K.E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. Journal of Educational Measurement, 28(1), 325-337.

Scheuneman, J.D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement, 24(2), 97-118.

Scheuneman, J.D. (April, 1990). Assessing the utility of item response theory models: Differential item functioning. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6(4), 317-375.

Shepard, L., Camilli, G., & Williams, D.M. (April, 1983). Accounting for statistical artifacts in item bias research. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Shepard, L.A., Camilli, G., & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22(2), 77-105.

Shermis, M.D., & St. George, R. (April, 1990). Item bias in mathematics achievement: The progressive achievement tests for mathematics. Paper presented at the annual meeting of the National Council for Educational Measurement, Boston, MA.

Simon, M.G. (1987). Statistical and subjective bias analyses of translated educational achievement. Unpublished doctoral dissertation, University of Toronto.

Subkoviak, M.J., Mack, J.S., Ironson, G.H., & Craig, R.D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21(1), 49-58.

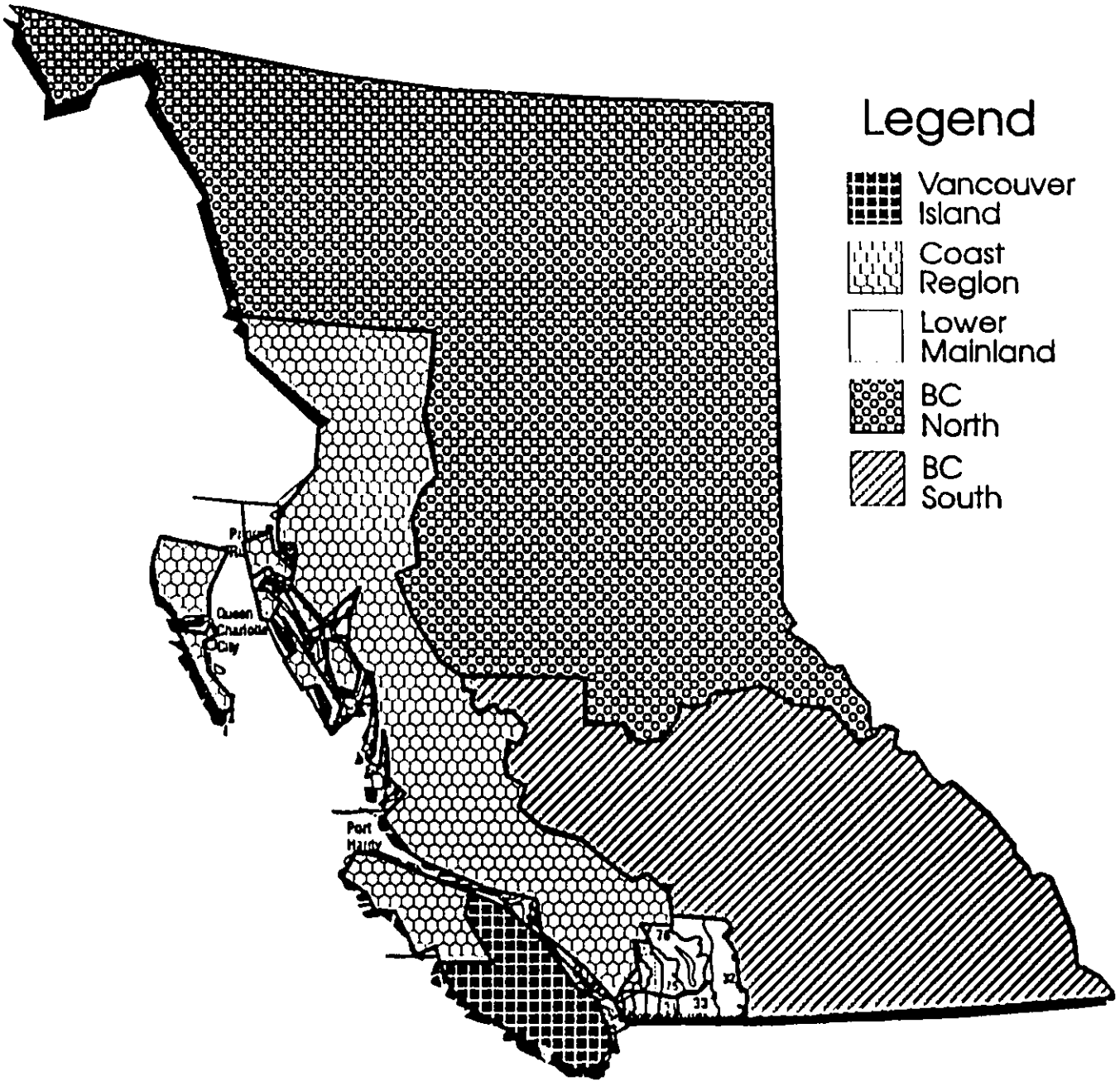
Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

Wright, D.J. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History Assessment. Journal of Educational Measurement, 26(1), 55-66.

APPENDIX A

BRITISH COLUMBIA SCHOOL REGIONS



APPENDIX B

BRITISH COLUMBIA SCHOOL DISTRICTS