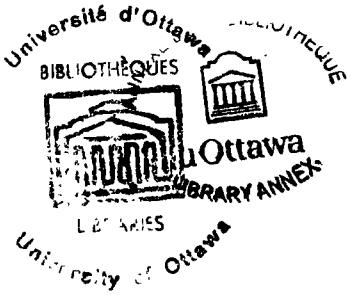
AN EMPIRICAL INVESTIGATION OF THE HUI-TRIANDIS MODEL TO ASSESS THE MEASUREMENT EQUIVALENCE OF DIFFERENT CANADIAN SUBPOPULATION (CULTURAL) GROUPS

By: RALPH KELLETT



Thesis submitted to
the School of Graduate Studies and Research
in partial fulfillment of the requirements for the
PhD degree in Educational Mesurement and Evaluation

University of Ottawa

ABSTRACT

A prevalent and serious criticism of standardized tests has been that they are biased in content, procedure, and usage against members of culturally different minority groups. Over the last 15 years progress has been made in defining the critical features of cultural bias and in devising operational techniques for assessing its dimensions. However, the application of varied strategies and techniques in the study of item or test bias has often led to conflicting findings. To provide a framework within which a multi-strategy approach to the study of test bias could occur, Hui and Triandis (1985) proposed a model for the cross-cultural comparison of test performance. The model related, in an hierarchical manner, a continuum of equivalence assumptions to various measurement strategies used to challenge the assumptions. The purpose of this study was to empirically investigate the usefulness of certain aspects of the Hui and Triandis model for demonstrating the equivalence in test performance of culture group members relative to members of a norm group.

Subjects for the study were selected to represent both culturally different and norm members of the population. Culture group subjects were drawn from among Canadian Chinese immigrants and native peoples. Two groups of native peoples were included, those living on and off reserves. These two groups were considered to represent people at two different levels of what was referred to as a cultural exposure continuum (Laveault, 1983). Norm group subjects were selected from among people not known to belong to an identifiable culture group. Chinese and norm subjects were, subsequently, divided at random into two groups each to provide a means for making control group comparisons. That is, differences in test performance of

culture group subjects with norm subjects might be expected, but subjects divided at random into two groups from the same population should not differ. All subjects were 17 to 24 years of age, with 11 to 14 years of schooling, enrolled in similar high school or college programs, and were drawn from the same geographical areas.

To obtain test performance data, a total of 3,312 subjects were administered four sub-tests of the General Aptitude Test Battery (GATB). Two of the sub-tests given were not language-related (Computation and Three Dimensional Space), and two were language-related (Vocabulary and Arithmetic Reasoning). Language-related and non-language-related sub-tests were included in order to account for the effect on test performance of fluency in the language of the test (in this case English). Test-retest reliabilities for these sub-tests have been reported in the range of .83 to .85.

Equivalence or non-equivalence of the test performance of culture group subjects relative to norm subjects was determined by the use of six measurement techniques. The six techniques selected for use in the study were Comparative Factor Analysis, the Modified Caution Index, (C*), the Standardized Appropriateness Index, (L_z), the Standardized Extended Caution Index (ECIZ4), Distractor Analysis, and Correct/Incorrect Response Analysis, (ΔMH). These six techniques were taken to represent four measurement strategies of the Hui and Triandis model.

To determine if the model was useful, three study hypotheses, derived from the tenets of the Hui and Triandis model, were subjected to examination. The first hypothesis, which stated that non-equivalence in test performance of culture group subjects relative to norm subjects should not be found for the control group comparisons using all six measurement

techniques, was supported. The second hypothesis stated that non-equivalence should be found for other than control group comparisons using at least some of the measurement techniques. This hypothesis was generally supported, but the descriptive nature of the study made it difficult to draw firm conclusions. The third hypothesis stated there would be consistency in judgements of equivalence or non-equivalence when different measurement techniques representing the same measurement strategy were applied. The third hypothesis further stated that the hierarchical nature of the model would be found to hold true. This third hypothesis was only partially supported. Consistency in judgements was obtained using one pair of techniques representing the same strategy; however, consistency was not obtained when a second pair of techniques representing a different strategy was used. It was concluded that the second set of techniques probably best represented separate strategies. The hierarchical nature of the model was seen to hold true. In the end, a modified version of the Hui-Triandis model was proposed for further study. Overall, the model was found to be useful, but strong conclusions could not be reached due to the descriptive nature of the study.

Using real-life data to study the usefulness of the Hui-Triandis model perhaps had the advantage of permitting the findings to be of practical value. However, there was no way to be certain that subjects sampled as part of a culture group were, in the final analysis, different in a cultural sense from those belonging to the norm group. It was also difficult in the study to isolate the effects of cultural differences on test performance from the influence of other factors such as ability level differences. It was, therefore, suggested that future researchers might examine the usefulness of the model using <u>both</u> simulated and real-life data. It was

also suggested that further research is needed to determine the statistical distributions of some of the measurement technique indices used, specifically C*, L_z and ECIZ4 indices. For other measurement techniques (ΔMH, Distractor Analysis, Comparative Factor Analysis) further research would be helpful in establishing criteria to judge when test performance is equivalent or not. Finally, the generalizability of findings from this study was said to be limited. Further investigation of the model using culture group subjects other than Chinese and natives and using tests other than the GATB, would be worthwhile.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Marc Gessaroli for his assistance throughout this study. His observations, recommendations and direct help, particularly in data analyses, were invaluable. His encouragement truly enhanced the learning experience.

The co-operation and support of Employment and Immigration Canada, which enabled the gathering of test data, were also greatly appreciated.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

LIST OF TABLES

Page

LIST OF TABLES (cont'd)

Page

LIST OF TABLES (cont'd)

Page

LIST OF TABLES (cont'd)

Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

A prevalent and serious criticism raised by test critics against standardized testing is that achievement and aptitude tests may be biased in content, procedure, and usage toward disadvantaged minority groups in the population. It is claimed that some of the traditional psychometric tests hold questionable validity for assessing the scholastic or vocational aptitudes of members of culturally different minority groups (Block & Dworkin, 1976; Cole & Bruner, 1971; Feuerstein, 1979; Ginsburg, 1972; Riessman, 1974; Sattler, 1982; Williams, 1971).

Over the last 15 years progress has been made in defining the critical features of cultural bias and in devising operational techniques for assessing the various dimensions of bias (e.g., Berk, 1982; Jensen, 1980, 1984; Thorndike, 1982). While the researchers have used a wide variety of tests, methodologies, bias criteria, and samples (largely drawn from American society) five strategies for detecting or providing evidence of cultural variation, or bias, have gained preponderance. One strategy, focused at the test level, has been to look for cultural group differences in construct validity (Cronbach & Meehl, 1955; Jensen, 1980, 1984; Reynolds, 1983). A second approach, which examines the question at the item level, has been the comparison of proportions of correct answers across cultural (subpopulation) groups (Angoff & Ford, 1973; Cleary & Hilton, 1968; Ironson, 1982). A third approach, also directed at the item level, has been a cognitive assessment of bias in item content and format performed by culture group representatives (Berk, 1982; Peterson & Novick, 1976). The fourth focuses on the application of the test and has examined cultural group

differences in predictive validity (Hunter, Schmidt & Rauschenberger, 1977; Jensen, 1980; Linn, 1973). A fifth and more recent approach has been the use of Item Response Theory (IRT) methods as a substitute for proportion correct statistics in studying item and test bias (Hambleton, 1983; Hambleton & Swaminathan, 1985; Lord, 1977, 1980; Trabin & Weiss, 1983).

One difficulty in applying these varied approaches to the study of item or test cultural bias is they often provide conflicting results about the presence of bias toward a particular cultural or subpopulation group (Drasgow, 1984; Linn, 1982). Another problem area is that they focus on the equivalence (absence of bias) of items or tests across cultural groups at different levels. Drasgow (1987) points out the importance of studying in a heterogeneous population the extent to which two distinct properties of a test are satisfied. The first, measurement equivalence, can be said to hold if individuals with equal standing on the trait measured by the test, but drawn from different cultural or subpopulation groups, have equal expected test scores. This property is frequently tested using the first, second, third and fifth strategies mentioned above. A test is biased at this level if the test does not provide measurement equivalence. Relational equivalence, a second property, refers to the extent that a bivariate measure of association of test scores with criterion scores is identical across relevant cultural or subpopulation groups. This property is usually tested using the fourth strategy mentioned above.

In employment and educational testing relational equivalence is usually studied using job or classroom performance as the second criterion variable. The bivariate relation studied in this case is the regression of performance on test scores, including intercepts, slopes, and standard errors of estimate (Drasgow, 1987). Using regression analysis to demonstrate

relational equivalence is often called "differential prediction", a subject which is thoroughly reviewed in the literature (Schmidt & Hunter, 1981; Petersen & Novick, 1976).

Justification for the "fair use" of achievement and aptitude tests with the members of social minority or culturally different groups in the population has largely, in the past, been based on the demonstration of relational equivalence (Linn, 1982, 1973). For example, some researchers (Hunter, Schmidt & Rauschenberger, 1977; Schmidt & Hunter, 1981) have argued that cognitive ability tests are equally valid for minority and majority applicants for employment or for college admission in that they do not (based on differential prediction studies) underestimate the expected job or classroom performance of minority groups. However, dependence on relational equivalence evidence alone to justify the use of tests with minority or culturally different individuals has not been considered acceptable by many. Linn (1984) pointed out that there is often no evidence of an absence of bias in the criterion measures used in the differential prediction studies. In two recent court cases in the United States (Drasgow, 1987) evidence of relational equivalence was not accepted as justification for the use of the tests involved, and the parties who administered the tests were directed to provide evidence of what Drasgow (1987) calls measurement equivalence. The need to demonstrate measurement equivalence as well as relational equivalence before tests are used with culturally different individuals was also recently cited by Fairweather (1986).

While the meaning of relational equivalence seems clear and the methods used in its study are well established, a common understanding of measurement equivalence, as defined earlier, and a well-developed set of

techniques for its study do not exist. The problem seems to be that the current notions of item and test bias are too limited. What is needed, then, is a framework within which the utility of various strategies, used alone or in combinations, for ascertaining measurement equivalence can be gauged. Malpass (1977) suggests that the starting point in the development of a framework for the study of cross-cultural equivalence in testing is to first address the problem of equivalence at the abstract level. Then the researcher can move, in examining test equivalence, from abstract theoretical categories to concrete local operationalizations by the development and application of relevant procedures. The general approach offered by Malpass is incorporated in a model for cross-cultural comparison recently published by Hui and Triandis (1985).

The purpose of this study is to investigate the usefulness of the Hui and Triandis model in determining the measurement equivalence of tests across cultural, or sub-population groups. In fact, the Hui and Triandis model provides a framework for the study of cross-cultural equivalence in tests which includes both measurement and relational equivalence. Since the focus of this study is on measurement equivalence alone, a modification to the Hui and Triandis model, based on the concept of measurement equivalence as defined by Drasgow (1987), is presented in the Review of Literature. In the end, only those aspects of the Hui and Triandis model related to measurement equivalence are investigated in this study. Relational equivalence is not given further consideration. Finally, It should be noted that the usefulness of the Hui and Triandis model is assessed in this study using real-life, rather than simulated data (as recommended by Kok, Mellenbergh, and Van der Flier [1985]).

This study should contribute to a better understanding of what constitutes equivalency, or lack thereof, in test performance across various groups. It should help to clarify the operational definition of equivalency and provide a well-defined set of techniques for demonstrating measurement equivalence. Such a set of techniques would complement the existing, well-established methods and procedures used to demonstrate relational equivalence. With a well defined set of techniques it might be possible to assist psychologists, educators, school and vocational counsellors, and teachers to avoid conflict in the establishment of a legal foundation for the decision to use or not use tests with, for example, particular immigrant or native peoples groups. The latter outcome is particularly important in view of the fact that most research in this area is based on American data and American racial groups. There is, therefore, a need for this type of investigation to be carried out within the Canadian context.

In the following chapter, (Review of Literature), the conceptual framework, or model, proposed by Hui and Triandis (1985) for the study of cross-cultural equivalence in tests is described. The tenets of the model are then presented. In order to focus on measurement equivalence, a modified model is proposed. It is the modified model which is, in fact, investigated in this study. The Review of the Literature concludes with a formal statement of the purpose of the study and a statement of the research hypotheses. In Chapter Three, the methodology for investigating the model is presented, and in Chapter Four, the results of the investigation are given and discussed. Chapter Five includes conclusions and recommendations.

CHAPTER II

REVIEW OF THE LITERATURE

In this chapter, a model for demonstrating cross-cultural equivalence in tests is presented. This model, offered by Hui and Trindis (1985), is subsequently modified to focus on measurement equivalence. The tenets of the modified model are described, upon which basis hypotheses are proposed to enable investigation of the validity of the model.

The Hui-Triandis Cross-Cultural Equivalence Model (Hui & Triandis, 1985)

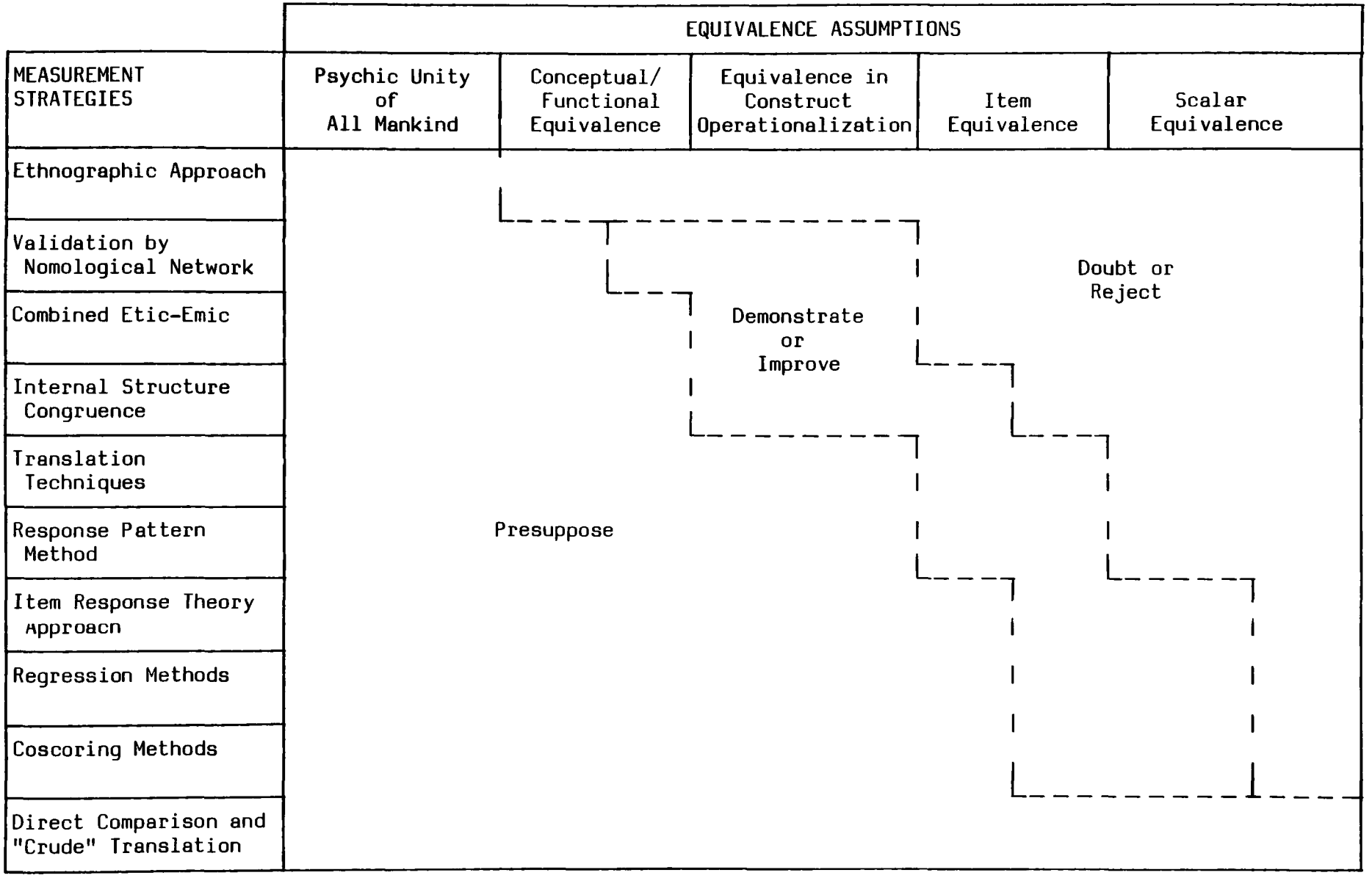
A model for the cross-cultural comparison of test performance has been proposed by Hui and Triandis (1985). The model relates various cross-cultural equivalence assumptions or conditions (for tests) to a specified domain of cross-cultural measurement strategies. It accounts for both measurement and relational equivalence (as discussed by Drasgow, 1987). A representation of the model is given in Figure 1. Some discussion of the equivalence assumptions, or conditions, and cross-cultural strategies is needed, as well as an explanation of the correspondence between strategies and assumptions.

The Equivalence Assumptions

Hui and Triandis (1985) developed the equivalence assumptions continuum of Figure 1 on the basis of the psychological conception of two continua:

FIGURE 1

The Hui-Triandis (1985) Model of Cross-Cultural Equivalence in Tests



an abstraction-concreteness continuum and a universality-cultural specificity continuum. The progression of equivalence assumptions in the model (from left to right in Figure 1) can be considered parallel to both an abstraction-concreteness and a universality-cultural specificity continuum. At the extreme left, the "Psychic Unity of All Mankind" is the weakest assumption of equivalence, existing at a very abstract and culturally universal level. Scalar equivalence is the strongest assumption, existing at a concrete and culturally specific level. For the universality-cultural specificity continuum, one can view a concept as being stable over time and space, so that it might have meaning for all people in all locations. That is, it might be universally understood. For example, "happiness" is probably understood by most people in most cultures in some general way. At the other end, a concept might only be understood by people in one location at one point in time; that is, it might be culture specific. For example, "upward mobility" would have meaning in North America but would not be meaningful in remote aboriginal cultures. For the abstraction-concreteness continuum, a concept might be so abstract that any operationalization of it would be impossible, or it might be concrete, including considerable detail of an operational definition. Triandis (1978) suggests that the universality-specificity and abstraction-concreteness continua are closely interrelated. For example, if ability is defined as life sustaining skills, then people in most cultures have it, and probably to a similar degree (the concept is defined in abstract terms and is culturally universal). If ability is defined as life-sustaining skills in an electronic-technological age, specifically as the knowledge and skill needed to operate a micro-computer, only some people in some cultures would possess it and to

different degrees (the concept is defined in concrete terms and is culture specific). Thus, moving from the abstract to the concrete end of one continuum normally implies movement from universality in cultural application to specificity.

In addition to its link with the universality-specificity continuum, the abstraction-concreteness continuum is also related to the precision with which cross-cultural differences on a given construct can be detected. For instance, if "happiness" were compared across two cultural groups in a population, the results would probably be quite imprecise and judgemental. However, comparing a more concrete construct like psychopathology, which is perhaps related to happiness, across the same cultures could be achieved through examination of the number of hospitalized mental patients per thousand people in each cultural group. To make meaningful comparisons it is necessary to move from abstraction to concreteness and at the same time universality to cultural specificity.

In general, researchers proceed according to Hui and Triandis (1985) from the abstract end of the one continuum and assume or demonstrate that a concept, attitude, attribute, value, or property is common and universal. Then, a research procedure can be considered equivalent across cultural groups studied up to the point at which assumptions of a more concrete or specific nature are required. After that point, the research procedure can no longer be used across the cultural groups with the expectation that equivalent results may be attained.

Researchers have written about the concept of "cross-cultural equivalence" as a prerequisite for comparisons across cultural and ethnic boundaries. However, different labels and descriptors have been used

(Triandis, 1978). In describing their model of cross-cultural equivalence, Hui and Triandis (1985) attempted to summarize the ideas related to several major types of equivalence, resulting in the "equivalence assumptions" categories of Figure 1. Each of those categories needs some explanation. Note that the assumptions are presented in the order they appear on the "equivalence assumptions continuum", which is from the left (abstract and universal) to the right (concrete and culturally specific). All of the following descriptions are taken from Hui and Triandis (1985).

Psychic unity of all mankind

The first and most abstract kind of equivalence is the "psychic unity of all mankind". Simply put, this assumption implies that all people in all cultures can be characterized by the same general properties or characteristics, such as emotions, attitudes, and intellect. The characteristics are known and understood in the abstract form universally across cultures.

Conceptual/functional equivalence

At the next level, "conceptual/functional equivalence" of a construct is said to exist if the construct can be <u>meaningfully discussed</u> in the cultures concerned. For example, the construct of weight lacks equivalence in a comparison of oranges and love, not because love cannot be measured, but because weight is irrelevant as an attribute of love. Hui and Triandis (1985) define conceptual equivalence as an equivalent understanding across

cultures of a notion, at the cognitive level. They refer to functional equivalence as being the setting of equivalence goals at the behavioral level. Conceptual and functional equivalence, they argue, are closely linked because they both pertain to the similarity or difference between the goals of two behaviours. For instance, "weight" is related to juiciness, size and quantity in the case of oranges, but lacks similar relationships in "love".

Equivalence in construct operationalization

"Equivalence in construct operationalization" refers, according to Hui and Triandis (1985), to the use of the same procedures (to measure the construct), with the same meaning assigned to the procedures across the different cultures. For example, equivalence across mute and general populations would not exist if the construct of aggression were operationalized through measures of verbal insults.

Item equivalence

At a more concrete level, equivalence can be viewed in terms of "item equivalence". Once it is assumed or demonstrated that a construct has a similar meaning in two cultures (conceptual/functional equivalence), and is operationalized in similar ways (equivalence in construct operationalization), the next task is to show that the construct can be measured by the same instrument. Only in doing this can cultures be numerically compared. Item equivalence implies that <u>each test item</u> means the same thing to people

from different cultures. A test that lacks item equivalence for subjects from two cultures is in reality two separate tests, one for each culture. In this case a comparison of test scores is misleading and illegitimate. According to Hui and Triandis (1985), each item of a test could mean the same thing to subjects from culture A as it does to those from culture B (achieving item equivalence) without the subjects' scores from the different cultures being measured on the same scale.

Scalar equivalence

A test has scalar equivalence for two cultures if the other types of equivalence are attained and if it can be shown that the construct is measured on the same scale. For scalar equivalence to exist a numerical value on the test's scale refers to the same degree, intensity, or magnitude of the construct regardless of the population of which the examinee is a member. This definition implies that one examinee's test performance is equivalent to another's only when the test scores are placed on a common (ability) scale. In practice, scores are placed on a common scale when the item parameter estimates used to derive examinee ability scores are obtained through sample-free (IRT) methods. This type of equivalence is ideal for cross-cultural comparison, but is usually difficult to achieve (Triandis, 1978).

Hui and Triandis (1985) point out that a clear demarcation of the last four equivalence assumptions does not exist. In particular, the boundary between conceptual/functional equivalence and equivalence in construct operationalization, on the one hand, and between item and scalar equivalence

on the other hand, is often not clear. What is evident, according to Hui and Triandis, is that equivalence in construct operationalization cannot occur in the absence of conceptual/functional equivalence. Neither can scalar equivalence exist without having item equivalence first established.

Measurement Strategies

For their model of Figure 1, Hui and Triandis (1985) explain that the measurement strategies for cross-cultural comparison can be examined from the top down, that is, from a general to a specific level, and are hierarchical. However, the strategies array is not intended to represent a continuum. Each of the 10 strategies are discussed below, as presented by Hui and Triandis (1985) in their model description. Each strategy is described under a section title indicating the equivalence assumption or assumptions the strategy can be used to challenge.

Assumption challenged - psychic unity of all mankind

Ethnographic Approach

The ethnographic approach is totally descriptive, only indicating the cultural groups that are represented in the comparisons, and characterizing the groups. With this strategy there is little danger of inaccurate comparison across cultures since it provides little precise information about cross-cultural differences.

Assumption challenged - conceptual/functional equivalence and equivalence in construct operationalization

Validation by Nomological Network

Validation of a test by a nomological network is a strategy recommended by Cronbach and Meehl (1955) in which multiple tentative assumptions of cross-cultural applicability of operationalized constructs are made and then empirically tested. The argument presented in this strategy is that if a construct has the same meaning across cultures, it should enter into the same empirical relationships in the cultures. Thus, if the networks resulting from the testing appear similar, it is concluded that instruments used in the process are cross-culturally applicable and equivalent. If they were not similar, similar relations in the nomological networks for different cultures would presumably not have been discovered. For example, assume two cultures (A and B) are being studied. A test to predict "job success" is administered to subjects from both cultures. It might be assumed that high scores on the test would be related to a good safety record on the job (few accidents), low absenteeism, and high productivity (number of units produced per shift). Further, it might be assumed that a good safety record was related to high productivity, as was low absenteeism. If the inter-correlation of test scores with the three criterion variables, and the inter-correlations of criterion variables with each other are similar for the two cultures, it would be concluded the test is cross-culturally applicable.

Clearly, the nomological network approach is insensitive to cross-cultural non-equivalence at the test item level, nor is the question of scalar equivalence addressed. Moreover, the multiple tentative assumptions empirically tested with this strategy require that criteria external to the test (as shown in the example) have to be used. If the networks turn out to be similar for two cultures, it is probably justified to state that the constructs in question have cross-cultural conceptual/functional equivalence and equivalence in construct operationalization. When they do not resemble each other, it cannot be determined whether it is the construct being examined by the test or the external criteria that are non-equivalent culturally. Furthermore, the correlations between constructs and criteria are as often due to experimenter bias and common method variance as to a similarity of the nomological networks.

Assumption challenged - equivalence in construct operationalization

Etic-Emic Approach

Another generalized strategy cited by Hui and Triandis (1985) is that of the combined etic-emic approach. Following this method the researcher identifies an "etic" (culture general) construct appearing to have universal status. Then, "emic" (culture specific) ways of measuring the construct are developed and validated. At the end, the "emically defined etic" construct is used in cross-cultural comparison. It should be noted that the culture general (etic) construct is defined and measured in culture specific (emic) ways. This approach is tantamount to developing equivalent or parallel

theories of social behaviour across cultures, which presumably have similar structures. Applying the strategy is, therefore, limited to only highly abstract constructs which are operationalized somewhat differently in the different cultures. Although Hui and Triandis (1985) assert that this strategy can be used to demonstrate equivalence in construct operationalization, it is difficult to see how it can do so when the construct of concern is not necessarily defined operationally and measured the same way in the cultures being compared. Thus, this strategy will not be applied in the current study, as explained in the later section, A Modified Model Proposed for Investigation.

Assumption challenged - equivalence in construct operationalization

Internal Structure Congruence

Researchers frequently address the question of cross-cultural equivalence of test scores by investigating and comparing across groups the internal structure of the construct. It is reasoned that if a construct is the same across cultures, it should have the same components and exhibit the same relations among components across cultures. Statistical procedures are usually employed to understand the anatomy of the construct. The most popular method of comparing internal structures across groups is that of exploratory and confirmatory factor analysis, and more specifically, a comparison of correlation matrices.

A step-wise procedure is usually employed where covariance and then correlation matrices are compared across groups. These matrices inter-

relate either test part (sub-test) scores or test item scores. Factor structures and factor weights for the groups extracted from analysis of the inter-correlations may then be compared to assess similarity of construct structure. Hui and Triandis (1983) also propose that multi-dimensional scaling may be used as a technique within this strategy. In this case, a small sample from each culture judges the similarity in meaning of pairs of items to be used in an instrument. Knowing that, for example, judges from Culture A do not consider a certain dimension as important as do judges from Culture B, equivalence in cross-cultural construct operationalization can be doubted.

Hui and Triandis (1985) indicate that comparing the internal structure of a construct across cultural groups, as suggested above, should not be done in the absence of a proper theory. For example, an hypothesized factor structure for a test, based on data from at least one culture, should be possible. Too often, sporadic and unplanned comparisons of internal structures are made as part of cross-cultural studies, usually having inconclusive results. However, it should be noted that Hui and Triandis (1985) indicate in their model (Figure 1) that the Internal Structure Congruence strategy can be employed to demonstrate, in some way, item equivalence. Even when factor analytic techniques are used, equivalence of test items across groups, as defined earlier (<u>each item</u> means the same thing to people from different cultures), cannot be demonstrated. Similarity of factor structures across groups, even when factor weights are similar, does not imply an absence of item bias (non-equivalence of any one item across groups). It implies simply that the <u>inter-relationship</u> among items for the different groups is comparable.

Assumption challenged - item equivalence

Translation Techniques

In the model of Figure 1, Hui and Triandis define another strategy for demonstrating the cross-cultural equivalence of a test, one based on translation validation and linguistic adaptation of tests. In this case the assumption is made that cross-cultural application of the test involves translation into a different language. The authors suggest that a number of techniques exist, such as back-translation, bilingual and committee approach, decentering, and pretests, which aim to put the test in different languages while preserving the same ideas across the linguistic boundaries.

Hui and Triandis (1985) contend that the end product of a translation approach could be a test reflecting the same underlying idea and meaning the same thing to test-takers from different cultural and linguistic groups. A translated test may, in fact, have the same underlying ideas with the same constructs defined the same way across linguistic groups; that is, the tests may be found to have similar factor structures. However, contrary to the view expressed by Hui and Triandis (1985) item equivalence would not necessarily exist as the content, in most cases, of the items would not always be the same. Exceptions to this premise would probably be tests not involving direct language usage, such as tests of perception and spatial ability (Werner & Campbell, 1970). This strategy is clearly appropriate only when a translation of test content is required.

Assumption challenged - item equivalence

Response Pattern Method

For this strategy, Hui and Triandis (1985) point out that it is based on a comparison of item response patterns, taken as a whole over tests, of individuals in relation to a group or of one group in relation to another. In most techniques subsumed under this strategy an "expected pattern" of test item responses is constructed based on the matrix of individual total test scores and item parameter estimates (usually only item difficulty values). Then an individual's pattern of responses (correct or incorrect) can be compared to the expected pattern. Similarly, the pattern of responses for one group can be compared to that of another. Hui and Triandis (1985) note that according to Van der Flier (1982), the rationale for this approach is that high scorers should not miss easy items, nor low scorers get hard items correct, unless the items do not mean the same thing to the individuals, or cheating, copying, or some other irregularity has occurred.

Hui and Triandis (1985) also refer to other techniques, such as one advocated by Angoff (1972), which compare group response patterns across all the items in a test with one another. This is done by correlating the order rankings of item difficulty for each pair of cultures concerned. In this case, the difference in the order of difficulty may reflect the fact that certain items are more or less difficult for one culture, but not for the other culture.

The response pattern method, as a measurement strategy for determining the cross cultural equivalence of tests can be used to challenge the assumption of item equivalence. However, scalar equivalence as defined for the Hui and Triandis (1985) model (Figure 1) cannot be implied using any of the techniques normally applied as part of this strategy. Another strategy is needed to demonstrate equality of scales on a test across cultural groups.

Assumption challenged - item and scalar equivalence

Item Response Theory Approach

Hui and Triandis (1985) suggest that item response theory (IRT) methods can be used to demonstrate both item and scalar equivalence across cultural groups. They note that IRT techniques have the advantage of allowing comparisons to be made across groups using criteria which are independent of the groups being compared. IRT is based on a mathematical model which relates the probability of a correct response to an item to examinee ability and to item parameters. The item parameters are derived based on a large reference group of subjects. Calibration of items from a large reference group using the item response function (a mathematical model) is intended to render the item parameter estimates group independent. This is unlike the group determined item difficulty mentioned in the previous section on the item response pattern approach. The IRT approach to item bias study (item non-equivalence) is to compare the differences between item characteristic curves (ICC) for examinees classified into various groupings (Lord, 1977).

The ICC, obtained using the reference group, represents the probabilities of correct response to an item by examinees possessing different levels of the ability (latent trait) being measured. By statistically comparing differences between ICCs one can establish if differences exist, and thus, when equivalence on an item can be claimed or denied. Because IRT methods place all examinees on the same ability scale (Hambleton, 1983) equivalent ICCs for two groups implies scalar equivalence across the groups on that item.

As disadvantages of the IRT approach, Hui and Triandis (1985) mention that the assumption of uni-dimensionality of test items has to be met, and that a large number of items and subjects are needed to obtain stable parameter estimates. They also suggest that any conclusion of equivalence is contingent on the demonstration of other types of equivalence at a more abstract level, for IRT is not concerned with the conceptualization or operationalization of constructs.

In proposing the use of IRT methods as a measurement strategy in their model, Hui and Triandis (1985) have confounded measurement strategies with techniques. All of the other approaches for challenging equivalence assumptions, which are included in the model of Figure 1 and which have been discussed up to this point, seem to be global strategies. They include: forming nomological networks; using emically-defined etic constructs for different culture groups; examination of underlying constructs; comparison of translated test versions for equivalence; study of the patterns of examinee or group responses across all items of a test. In fact, IRT methods, as discussed by Hui and Triandis (1985) are used to challenge the test item and scalar equivalence assumptions of the model through a

comparison of group responses to individual test items. The measurement strategy applied in this case should probably be termed "within-item analysis", and is certainly not limited to techniques employing IRT. Indeed, differences on test items can be detected by a variety of non-IRT methods such as the chi-square approaches (Ironson, 1982), log linear models (Kok, Mellenbergh, & Van der Flier, 1985; Van der Flier, Mellenbergh, Adler, & Wijn, 1984), standardization approach (Kulick, 1986), and the Mantel-Haenszel procedure (Holland & Thayer, 1986). The problem in applying within-item analysis to the issue of equivalence of a whole test is that of finding criteria to judge when a test is equivalent or not across groups, based on examination of individual test item equivalence (or absence of bias). The problem exists whether IRT or non-IRT methods of item bias study are used.

Another reason for suggesting here that the inclusion of IRT methods as an overall measurement strategy detracts from the usefulness of the model is that IRT methods can also be applied in Response Pattern Analysis, something not mentioned by Hui and Triandis (1985). Tatsuoka and Linn (1983) show how an individual's pattern of responses across items on a test can be compared to that of a group, or to an expected pattern of responses, using either IRT or observed item responses and standard summary statistics, such as the number or percentage of people in a group answering an item correctly. The use of IRT in Response Pattern Analysis is not new, with Wright (1977) and Levine and Rubin (1979) having done early work in what has become known as "appropriateness measurement". Thus, it might be more fruitful if IRT were to be included in the model of Figure 1 as a possible technique which could be used with either the Response Pattern Strategy or a Within-Item Analysis

Strategy, rather than as a separate measurement strategy. If IRT methods (or techniques) are applied in the Response Pattern or Within-Item Analysis Strategies, then the strategies can be used to challenge the scalar as well as item equivalence assumption. This is so because IRT appropriateness measurement (unlike non-IRT approaches) involves the use of a common ability scale for all examinees. The latter points will be addressed again in the development of a modified cross-cultural equivalence model to be applied in this study.

Assumption challenged - scalar equivalence

Regression Methods

Hui and Triandis (1985) state that this strategy for determining equivalence of a test across cultural (or other) groups is dependent on the use of an external criterion (or criteria). The use of the strategy enables conclusions to be drawn about what was referred to in the Introduction as "relational equivalence". Using this strategy involves a comparison across groups of some bivariate measure of association of test scores with criterion scores. If the measures of association (test scores with criterion scores) do not differ significantly for the groups, the test, in terms of its application, is said to be equivalent for the groups. Typically, the bivariate relation studied is the regression of criterion scores (such as job performance ratings, teacher grades, grade point average) on test scores, with statistical tests applied to the intercepts and slopes to determine if there are significant group differences in these parameters.

Hui and Triandis (1985) point out that Poortinga (1975) suggested that regression methods could be used to demonstrate scalar equivalence of a test. The case presented was that the regression parameters of the criteria to which the test scores are generalized can only be the same for the populations studied if the sets of test scores (for the populations compared) were obtained from numerically equivalent scales.

While the Poortinga (1975) approach is attractive because of its simplicity and economy, it does not take into account the effect of unreliability in the criterion measure, and differences in examinees' variability across the groups. In other words, regression parameters may not be the same for two groups of subjects simply because of unreliability in the criterion measured. As well, a criterion which is biased in terms of some extraneous factor would provide rival explanations for differences in regression parameters across groups (other than real differences in test score meaning). Finally, concluding that test scores would have to be derived from numerically equivalent scales in order for regression parameters to be equal is not tenable. The test scores for two groups, for example, need only be in the same general order of magnitude, but not necessarily on numerically equivalent scales, to produce regression slopes which are not significantly different. It is suggested that scalar equivalence of test scores across groups can only directly be imputed if examinees' scores are placed on the same scale (or in the same metric).

Assumption challenged - scalar equivalence

Coscoring Methods

The general strategy, or approach, referred to as "coscoring methods" by Hui and Triandis (1985) is based on the derivation for a set of tests, rather than a single test, of a "transcultural factor". This factor is obtained by factoring the test scores of a sample of people. Hui and Triandis (1985) refer to methods originated by Cattell (1957) through which one representative from each culture is obtained. The cultural representatives form the sample of people whose scores on a set of tests are factored. According to Cattell's (1957) transcultural method there are three ways to select a representative for each culture. One way is to simply select a person at random from each culture. In this case factor analysis will reflect both individual differences and cultural differences on the tests. A second way is to pick one person at random from each culture, then factor analyze standard scores (with each person's standard score obtained from the mean and variance of his own group). In this case, between group variance is taken out of the factoring, leaving only within-group variance. The third way is to factor analyze the mean raw score for each group on the tests, eliminating within group variance and considering only inter-group differences.

Hui and Triandis (1985) mention that two of the three means mentioned above for obtaining and factor analyzing a representative score for the cultures are problematic. The first way would result in factors on which subjects from different cultures could be scored, but when the pattern of

within group covariance is very different from that of inter-group covariance the factors derived may not be interpretable. For the second way of selecting scores representative of the cultures, "transcultural dimensions of individual deviations" (Cattell, 1957) are derived, but these are one step removed from the actual constructs underlying the tests, and in which the researcher has an interest. Only the third method of selecting representative scores for analysis seems very promising. However, the third method is nothing more than common factor analysis which will produce results the same as those obtained using the Internal Structure Congruence strategy.

Hui and Triandis (1985) indicate in their model (Figure 1) that this strategy can be used to demonstrate scalar equivalence. They suggest that application of Cattell's (1957) transcultural method would result in the derivation of transcultural factors, on which examinees from different cultures could be scored, and then their scores compared to determine (using "a common metric" which are the factor scores) equivalence of cultures across sets of tests. It is easy to see that factor scores constitute a common metric, but equivalence of cultures across factor scores is not the same thing as equivalence of numerical values across the original scores of the tests.

In addition to the difficulties cited above in interpreting results from use of the coscoring method, the strategy involves a comparison across many cultures simultaneously of examinees' performance <u>on a set of tests</u>. If the model proposed in Figure 1 is intended to explain how equivalence across cultures on a single test can be demonstrated, this strategy is of little value. Furthermore, the assertion by Hui and Triandis (1985) that

the method can be used to challenge the scalar equivalence assumption is not considered valid. The latter being true, the "coscoring method" as a strategy has nothing different to offer from the Internal Structure Congruence strategy.

Assumption challenged - all assumptions

Direct Comparison and Crude Translation

Perhaps the most commonly applied strategy for determining the equivalence of a test across cultures has been to compare culture group mean scores, often after a first (crude) translation of the test has been made from one language to another. Hui and Triandis (1985) have noted that typical procedures used with this strategy are simple t-tests or multivariate analysis of variance (MANOVA). In recent years, in an attempt to control for real ability differences across groups of subjects, a two-factor analysis of variance, with repeated measures on the second factor (test item) has been used (McCauley and Colberg, 1983). With this procedure it has been argued that a significant group by item interaction meant that the groups found the items to be of different relative difficulty. This was taken as an indication of non-equivalence of the test for the groups. A simple comparison of means would have been confounded by ability differences, but when a decision of equivalence or non-equivalence was based on the presence of a group by item interaction, ability differences were purported to be controlled.

Regardless of which specific technique is used to directly compare the test scores of groups and conclude that a test is equivalent or not, it is implicit, according to Hui and Triandis (1985) that in applying this strategy the construct being measured by the test exists in both cultures and is equivalently operationalized. Also note that in direct comparisons total test scores of examinees are not placed on the same scale, as is the case when IRT ability estimates of examinees are used. Thus, with this strategy scalar equivalence is assumed to hold, but little justification for such an assumption seems to exist. Indeed, this strategy cannot be used to directly challenge any specific equivalence assumption or assumptions.

Model Tenets

In the previous section the several assumptions necessary for cross-cultural equivalence of a test, according to Hui and Triandis (1985), have been described. They included (in Figure 1): the abstract notion that in general all people are similar to each other; the assumption of conceptual or functional equivalence of a construct; equivalence in how the construct is operationalized; item equivalence; and the strongest assumption of all, scalar equivalence. Recalling the concept of an abstraction - concreteness continuum outlined earlier, the progression of equivalence assumptions in the model (from left to right in Figure 1) can be considered parallel to the abstraction-concreteness continuum. "Psychic Unity of Mankind" is the weakest assumption made for equivalence, existing at a very abstract level, while scalar equivalence is the strongest assumption, existing at the most concrete level. Accepting or testing a certain assumption is only

meaningful when all previous assumptions are demonstrated to be true, or are simply accepted as true. For example, it is worthwhile demonstrating the equivalence of test items across cultures using IRT methods only if the constructs measured are conceptually and functionally equivalent across the cultures.

The model depicted in Figure 1 reflects a correspondence between equivalence assumptions and measurement strategies. The measurement strategies presented in the model do not operate independently of the set of assumptions. Some of the strategies can be used to demonstrate only the weakest (and most general) assumptions. For example, the ethnographic approach can not be used to demonstrate cross-cultural equivalence beyond the level of "psychic unity of all mankind". Other strategies can be applied to demonstrate the tenability of stronger assumptions; for example, the Internal Structure Congruence Strategy should effectively show if conceptual/functional equivalence and equivalence in construct operationalization exist. However, the latter strategy cannot be applied to directly demonstrate the psychic unity of all mankind. Thus, when the strategy is applied it must be accepted without direct proof that equivalence at the lower level of the "assumptions continuum" holds.

According to Hui and Triandis (1985), the different measurement strategies of the model (Figure 1) do not serve the same function and are not mutually replaceable. Each strategy can be used to directly challenge one or at most two of the equivalence assumptions. The strategies at the top (and left) of the model are applied to demonstrate the tenability of the weaker assumptions, and those at the bottom to demonstrate the stronger assumptions. Also, cross-cultural equivalence is presupposed (or accepted

without proof) as one moves from the abstract toward the concrete end of the assumptions continuum, up to the point where an assumption is too strong for the data set. At that point, the appropriate strategy from the model is selected to demonstrate or deny equivalence. It should be understood from viewing Figure 1 that a finding of non-equivalence using strategies higher in the array should imply non-equivalence for the same test and groups using a strategy lower in the array. If the Response Pattern Method, for example, was applied and the test was found to be equivalent for two groups, then one might expect the Internal Structure Congruence strategy (higher in the array) to also show the test to be equivalent across the groups. Should the latter not be found, the hierarchy of measurement strategies (top to bottom) in the model would not be correct.

The purpose of the "Doubt or Reject" region in Figure 1 is to make it clear that the assumptions lying in this region are not tested by a given strategy. Indeed, the assumptions lying in this region are stronger than the ones tested by a given strategy, in that the strategy being used does not produce evidence strong enough to support a conclusion of equivalence at a more concrete level. Regrettably, many people begin to apply a test across cultures in the belief that the test is equivalent in all ways simply because the test's constructs or items were validated by one particular method. If the same test is to be used for cross-cultural comparison, all the equivalence assumptions up to and including scalar equivalence should hold.

A Modified Model Proposed for Investigation

The Hui and Triandis (1985) model provides a broad framework for relating measurement strategies to assumptions or conditions which should be met in order to establish equivalence in test performance across cultures. However, it seems to unnecessarily confound the issues of measurement and relational equivalence. Indeed, as presented by Hui and Triandis, there is no distinction made in the model of Figure 1 between measurement and relational equivalence. It should be recalled that measurement equivalence refers to the similarity of a single test across cultures at the whole test or item level. Relational equivalence, on the other hand, refers to the accuracy of predictions made from a test to a criterion.

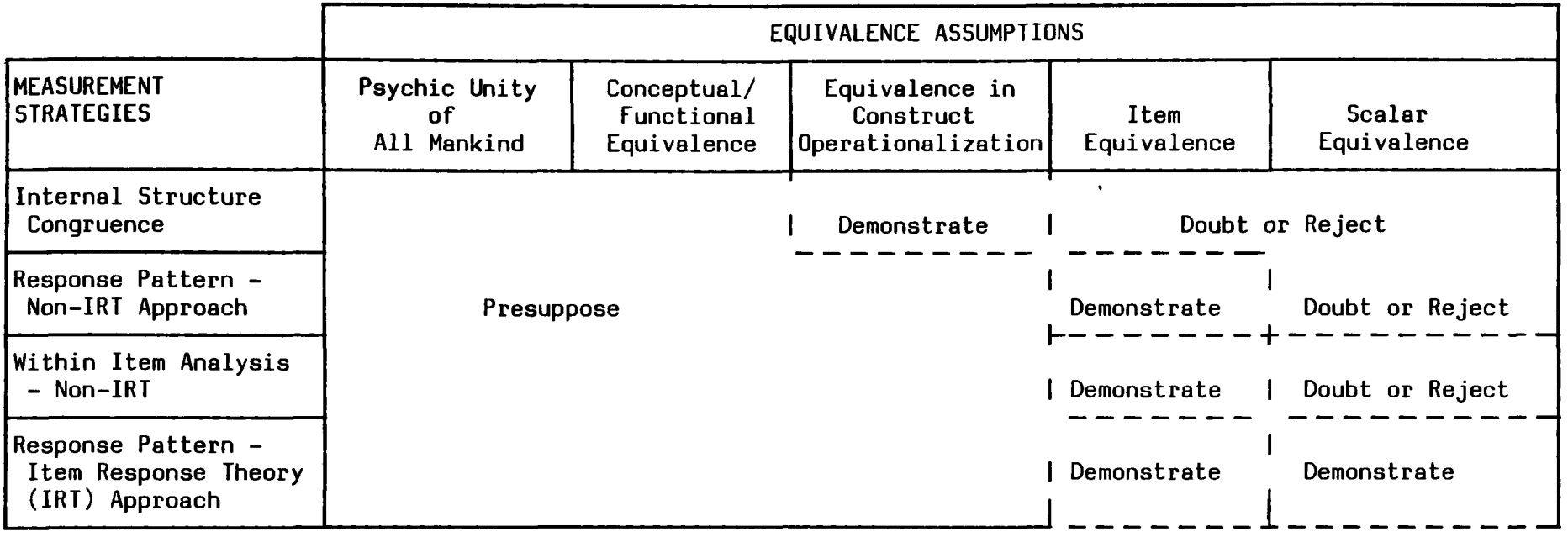
As discussed in the introduction to this study, the most fundamental issue in the "fair use" of tests across cultures is that the test items, and the test as a whole, mean the same thing to examinees regardless of the cultural or social group, or geographical area, they represent (Darlington, 1971; Drasgow, 1987; Linn, 1984; Peterson & Novick, 1976). Such similarity in a test across cultures is subsumed under the heading of measurement equivalence, as defined by Drasgow (1987). Only after a test is shown to possess measurement equivalence can the issue of relational equivalence (or equivalence in the application of the test) be properly addressed. In fact, studies focusing on relational equivalence assume equivalence in predictor test meaning (Schmidt & Hunter, 1981). Additionally, relational equivalence can really only be demonstrated when equivalence across cultures of the criterion measure is shown to exist. Thus, it is suggested that some additional assumptions or conditions (beyond those included in the model of

Figure 1) related to the equivalence of criteria measures across cultures are needed in any cross-cultural equivalence model that incorporates the notion of relational equivalence. As well, some different measurement strategies than those of Figure 1 would probably be required to challenge the additional assumptions. Clearly, two models would be appropriate for the study of cross-cultural equivalence: one focusing on equivalence of a single test across cultures (measurement equivalence), and one providing a framework for consideration of equivalence in application of the test (relational equivalence). In this study, the model of Figure 1 (Hui and Triandis, 1985) is modified in order to provide a model focused on measurement (single test) equivalence (see Figure 2).

Because the model of Figure 2 is meant to provide a framework within which the equivalence across cultures of a single test can be considered (measurement equivalence), the Hui and Triandis (1985) measurement strategy of Regression Methods is not included in the modified model. As explained earlier, regression methods relate predictor test to criterion performance and are used to demonstrate relational equivalence. The strategy of Validation by Nomological Network is also left out of the modified model because that strategy relates peoples' responses on a test to hypothesized outcomes which are operationalized through the use of criteria external to the test itself. As is the case when regression methods are applied, a conclusion of non-equivalence of a test across cultures, based on a dissimilarity of nomological networks for the cultures, could occur simply because the external criteria used in building the networks are not equivalent across the cultures.

FIGURE 2

A Modified Model of Cross-Cultural Equivalence in Tests



Four other measurement strategies included in the Hui and Triandis (1985) model of Figure 1 are not found in the modified model (Figure 2). First, the Ethnographic Approach is left out because it is totally descriptive providing no opportunity for testing hypotheses of equivalence of a test across cultures. Second, in the Combined Etic-Emic approach the constructs underlying a test can be operationally defined and measured in culture-specific ways, meaning that the construct definition may be different for different cultures. While this approach may permit a cross-cultural researcher to obtain observed measures derived for constructs relevant to each culture, empirically testing the equivalence of different constructs across cultures is illogical. Thus, the Combined Etic-Emic Approach is excluded from the modified model. Third, Translation Techniques are not considered to be an appropriate part of a model intended to provide a framework for determination of the cross-cultural equivalence of a single test, where the studied test is given in a language (for example, English) common to people of the different cultural groups. In fact, the issue of equivalence of different language versions of a test, and how equivalence in the translation and adaptation of a test is achieved, is really beyond the scope of both the Hui and Triandis and the modified (Figure 2) model. Coscoring Methods are not included in the modified model because those methods depend upon the simultaneous analysis, using factor analytic techniques, of a <u>set of tests</u>. The purpose of this study is to describe and empirically investigate a model for cross-cultural equivalence of a <u>single</u> test, as mentioned earlier. Finally, the Direct Comparison and Crude Translation strategy is not included in the modified model for two reasons.

First, it cannot be used to directly challenge any one specific assumption. Second, the techniques used with this strategy have little to offer beyond those included in the modified model under the Within-Item Analysis strategy.

In the modified model it should also be observed that Item Response Theory (IRT) is not considered to be a separate "measurement strategy". It is argued, earlier in this section, that one can determine equivalence of a test across groups by applying techniques that compare response patterns across whole tests of examinees from different groups. Such an approach is properly termed a strategy, and was presented earlier under the heading, Response Pattern Methods. As presented by Hui and Triandis (1985) Response Pattern Methods make use only of non-IRT techniques. However, with IRT "appropriateness measurement", IRT procedures are followed for estimating item and person parameters of each test item. These parameters are then used to compare examinees' patterns of responses to an expected pattern. The IRT procedure is similar to the non-IRT one described by Hui and Triandis (1985), except that IRT is used to estimate parameters. When IRT procedures are used, examinee ability is placed on a common scale, so a finding of equivalence for a test using this approach demonstrates item and scalar equivalence. When non-IRT procedures are used in response pattern analysis, only item equivalence can be shown. For the above reason, the Response Pattern Strategy is displayed in two parts of the modified model (Figure 2).

As a technique, IRT is also commonly employed, as discussed by Hui and Triandis (1985), in item bias studies, where groups are compared within-

items. The item characteristic curves (ICC) for different groups are compared to determine if a test item functions the same across the groups. In this case, IRT can be taken as a technique, for cross cultural comparison at the test item level, and clearly serves to demonstrate both item and scalar equivalence. "Within-item analysis" can also be conducted using non-IRT approaches which are far more economical. However, when non-IRT techniques are used, it is not possible to place examinee item responses on a common scale, and thus enable demonstration of scalar equivalence. Examples of non-IRT within-item techniques are the iterative logit method (Van der Flier, Mellenbergh, Adler, Wijn, 1984), the standardization method (Durans & Kulick, 1986), and the Mantel-Haenszel procedure (Holland & Thayer, 1986). In this study, within-item non-IRT analysis is applied as a measurement strategy in order to challenge the item, but not the scalar equivalence assumption. A non-IRT approach is applied because of its simplicity and economy.

A final difference can be seen between the models of Figure 1 and Figure 2. In the Hui and Triandis model (Figure 1), the correspondence of measurement strategies with the equivalence assumptions is made within three bands or areas, labeled "Presuppose", "Demonstrate or Improve", and "Doubt or Reject". Since all of the strategies are simply approaches which can be applied to either show or not show equivalence of a test across cultural groups, the use of the term "improve" in the model is not considered to be acceptable. Moreover, in the model of Figure 1 some strategies are shown as only partially challenging an assumption. A strategy can either be used to challenge a particular assumption (condition) or it can not. Therefore, in

the modified model the position of the vertical lines are adjusted to better reflect what is really possible.

It is, then, certain aspects of the Hui and Triandis (1985) model applied with more specificity (Figure 2) that are investigated in this study. Selection of only certain aspects of the Hui and Triandis (1985) model for investigation does not alter the basic tenets, which were enumerated earlier. It is these tenets that are to be put to the test. Nor does the selection of particular aspects of the original model (Figure 1) for investigation change the complementarity of the measurement strategies. Complementarity is an important aspect of the model if a multi-strategy approach to the study of cross-cultural equivalence of tests is to be productive. Indeed, in the modified model of Figure 2 the function of a strategy in challenging assumptions, as distinguished from various methods or techniques, is clarified. In a later section (methodology) various techniques or methods which can be applied as part of a strategy are presented. At that point, techniques are selected to represent each strategy of the modified model.

The Hui-Triandis (1985) model, or any other version of it, has not been previously subjected to empirical investigation. However, several studies have been conducted in which a multi-strategy approach was applied to a determination of cross-cultural equivalence. For example, Miller, Slomczynski, and Schoenberg (1981) examined the validity of indicators of authoritarian-conservatism across several countries using the combined etic-emic and internal structure congruence approaches. Hui and Triandis (1983) also used multiple strategies, namely multidimensional scaling,

factor analysis, and nomological validation in demonstrating that one aspect of locus of control was generalized across mainstream American and Hispanic sub-populations.

Further investigation of the issue of a multi-strategy approach to cross-cultural equivalence study, particularly in the area of test equivalence, is clearly desirable. The usefulness of a multi-strategy approach to the determination of the equivalence of a test across cultures can perhaps best be examined if such an approach is first explained by a model, and then tenets postulated for the model are empirically tested. The purpose of this study is to investigate the usefulness of the Hui and Triandis model, as modified in Figure 2, by empirically testing the tenets of the model. Note that the model tenets are the same for the original Hui and Triandis and the modified model.

Purpose of the Study and Hypotheses

In the previous section a modified model for the cross-cultural comparison of test performance is proposed. The modified model (Figure 2) is, in fact, a version of the Hui and Triandis model (Figure 1). It focuses on the issue of measurement equivalence in a single test. The modified model, like the Hui and Triandis model, provides a framework within which the value of a multi-strategy approach for demonstrating the equivalence of test performance across cultural groups can be gauged. To be useful, each of the strategies included in the model must have associated with it techniques or methods which can be used to represent the strategy. Further,

it must be possible to apply the techniques or methods representing each strategy in such a way as to render a judgement of equivalence or non-equivalence of a single test across culturally different groups.

The judgements made about the equivalence of the test performance of specified culture groups with a "norm" group must also be consistent over the different strategies, bearing in mind the hierarchical nature of the strategies. That is, non-equivalence at the more general strategy level must imply non-equivalence at the more specific levels. The converse does not have to hold. Finally, using the multiple strategies it should be possible to determine equivalence or non-equivalence with respect to cultural factors alone by controlling through sampling and statistical procedures the impact on equivalence of other confounding variables.

There should be little doubt that this is an exploratory, rather than a confirmatory, study. Neither all the strategies proposed, nor the techniques or methods representing them, have been previously applied together to the determination of equivalence in test performance across different cultural groups. However, most of the techniques or methods have been used to identify aberrance in test scores due to cheating, errors, prompting, or inappropriate instructional coverage. Nonetheless, their use as cultural group identifiers is new. Consequently, it is necessary to first independently determine the usefulness of each strategy (represented by the various techniques). Then, the hypothesis of hierarchical consistency in judgement can be tested.

In this study judgements are to be made as to whether or not a test measures the same thing in the same way for subjects classified as belonging to a norm group in relation to those categorized as members of another

specified cultural group. However, that is not the intended outcome of the study. The focus is on methodological issues. The intention is to determine if the particular combination of strategies included in the modified model can be applied in a manner which will enable consistent judgements to be made concerning the equivalence of test performance across different cultural groups in Canada.

Therefore, in this study the following hypotheses are tested.

Hypothesis 1

For paper-and-pencil group tests, it is expected that each measurement strategy under investigation in the modified model will yield a judgement of equivalence in test performance for two groups of subjects drawn from the same population.

Hypothesis 2

It is expected that a judgement of non-equivalence in performance on tests will result using at least some of the measurement strategies in the model for each culture group in relation to a group of "norm" subjects.

Hypothesis 3

It is expected that judgements of equivalence or non-equivalence in performance for culture group subjects in relation to norm group subjects will be consistent when two measurement techniques representing the same measurement strategy are applied. Further, it is expected that the

hierarchical nature of the model (described above) will hold. The hierarchical nature will be judged valid when the following condition is met:

> If the performance on a test of a culture group is judged to be equivalent with that of the norm group in terms of a measurement strategy at the top of the strategies array (Figure 2), equivalence can (but may not necessarily) be shown by application of subsequent strategies. If non-equivalence exists using a strategy higher in the array, then non-equivalence should be found for that group by application of all the strategies lower in the array.

CHAPTER III

METHODOLOGY

The methodology used to investigate the usefulness of specific aspects of the Hui and Triandis model for cross-cultural equivalence is presented in this chapter. First, the instrument used to obtain measures of test performance is described. This is followed by a description of the sample of subjects obtained for study. The data collection procedures are then detailed, followed by a discussion of the data analysis methods used. It is in this latter section of the chapter that specific measurement techniques are selected to represent the measurement strategies of the model displayed in Figure 2.

The Measuring Instrument

The General Aptitude Test Battery (GATB) was used in this study to obtain test performance data. This test battery was developed on the basis of statistical analyses of 59 different kinds of tests being used in the United States to predict job performance in a variety of occupations. Nine GATB aptitudes were derived which seemed to provide adequate measures of all the major abilities measured by the original 59 tests. Since 1947, the GATB has been repeatedly subjected to validity analysis, with over 500 studies having been completed which demonstrate the extent to which the test battery predicts future job performance.

The GATB measures nine aptitudes on eight paper-and-pencil and four apparatus tests. However, only four sub-tests formed part of this study, three cognitive and one perceptual paper-and-pencil test, requiring a total of forty minutes including administration time. The tests used were Vocabulary, Arithmetic Reasoning, Computation, and Three Dimensional Space. A brief description of these sub-tests including reference to the aptitudes they measure, is given in Table 1. The four paper-and-pencil tests used are contained in a single test booklet with answers recorded on a separate, multiple-choice answer sheet. The number of choices is different for the various sub-tests (or parts).

TABLE 1

Description of the Four Cognitive and Perceptual Sub-tests Used in the Study

Part 2 - Computation - 6 Minutes
This test consists of a number of arithmetic exercises requiring the addition, subtraction, multiplication, or division of whole numbers. Measures Numerical Aptitude. Total number of items is 50.

Part 3 - Three-Dimensional Space - 6 Minutes
This test consists of a series of four drawings of three-dimensional objects. The stimulus figure is presented as a flat piece of metal which is to be either bent, or rolled, or both. Lines indicate where the stimulus figure is to be bent. The examinee indicates which one of the four drawings of three-dimensional objects can be made from the stimulus figure. Measures General Learning Ability and Spatial Aptitude. Total number of items is 40.

Part 4 - Vocabulary - 6 Minutes
This test consists of sets of four words. The examinee indicates which two words have either the same or opposite meanings. Measures General Learning Ability and Verbal Aptitude. Total number of items is 60.

Part 6 - Arithmetic Reasoning - 7 Minutes
This test consists of a number of arithmetic problems expressed verbally. Measures General Learning Ability and Numerical Aptitude. Total number of items is 25.

The specific four sub-tests described in Table 1 were selected for use in this study in order to account for the possible confounding of language with culture. The effect of fluency in the language of a test, which in the case of the GATB is English, has not been subjected to controlled study. However, many of the differences between culture groups in mean test scores attributed to culture (design for living) were suspected to be due to deficiencies in language fluency, where the language of the test was not the first language of the examinee. In applying the Hui and Triandis model as a framework within which the equivalency of tests across culture groups could be determined, it was desirable to establish the role of language. Thus, two sub-tests were used in the study which have a known language component (Vocabulary and Arithmetic Reasoning), and two were used which do not (Computation and Three Dimensional Space). Differences in performance indicated by any of the measurement techniques for culture groups relative to a norm group on the two sub-tests which were language affected, but not on the two which were non-language in nature, might be explained by an effect due to language fluency.

The aptitudes of the GATB are measured reliably in the types of situations in which the battery is commonly used. Studies using samples from a variety of high school, college, and adult populations with intervals between initial testing and retesting ranging from one day to three years, show test-retest reliabilities in the range of .80 to .90. The reliability coefficients (test-retest) obtained by the United States Department of Labor, Manpower Administration (1970) for the four sub-tests used in this study are shown in Table 2. Reliability coefficients (KR-20) obtained for the total sample used in this study (N=3312) are also given in Table 2. Because of

the large number of items in the sub-tests relative to the testing time allowed (for example, six minutes allowed to complete the Vocabulary Sub-test consisting of 60 items), there were a number of items at the end of each sub-test which were not reached by any of the subjects. Since items not reached by any subject provide no information at all, they were not included in analysis. In essence, the number of items actually used in analysis for all parts of the study was reduced for each GATB sub-test in order to eliminate the items not reached by anyone in the sample.

TABLE 2

GATB Sub-test Reliability Coefficients

Sub-Test	Test-Retest Reliability (N=23428)		KR-20 Reliability (N=3312)	
	Coefficient	No. Items Used	Coefficient	No. Items Used
Computation	0.84	50	0.85	41
Three-Dimensional Space	0.81	40	0.87	40
Vocabulary	0.85	60	0.88	46
Arithmetic Reasoning	0.83	25	0.73	20

Note: KR-20 reliabilities are reported for each GATB sub-test broken down by sample group in Table A9 at the Appendix.

The Sample

In selecting subjects for this study, it was decided at the outset that the usefulness of the Hui and Triandis model in a determination of cross-cultural equivalence in test performance could be investigated using as few as two culture groups in addition to a base (norm) group. Among the many cultural groups to be found in Canada, the Chinese are relatively easy to identify and, as an immigrant group, are known to retain a strong cultural identity after arriving in Canada. Thus, Chinese were selected in this

study as one culture group. This group was sampled in a sufficiently large number to permit randomly dividing it into two halves for the purpose of providing control group comparisons (i.e., to test Hypothesis One). A second cultural group whose members are accessible is Canada's native people. Two native peoples groups were included in the study, one group whose members were not resident on reservations (off-res) and one group with members living on reservations (on-res). Further clarification of each of the above two groupings is provided below. The norm group consisted of adults not known to belong to any of the major cultural groups in Canada. This group was also divided to provide control group comparisons (i.e., to test Hypothesis One). A reference group, consisting of "norm" subjects was used in the study. The purpose for having a reference group is explained below. Table 3 provides a description of the entire sample.

TABLE 3

Descriptive Statistics of the Entire Sample

GROUP								
	NO. IN GROUP	MALE		FEMALE		MEAN AGE (YEARS)	STAN. DEV. AGE	MEAN EDUCA-TION YEARS	STAN. DEV. EDUCA-TION
COMPARISON GROUPS:		#	%	#	%				
Chinese 1	333	190	57	143	43	19.4	2.70	12.4	1.03
Chinese 2	347	189	55	158	45	19.6	1.45	12.5	0.78
Native (On-Reserve)	386	207	54	179	46	18.6	3.59	11.6	0.91
Native (Off-Reserve)	371	191	52	180	48	19.7	4.59	11.7	2.89
Norm 1	366	194	53	172	47	18.8	2.53	12.6	1.92
Norm 2	374	192	54	182	46	18.7	2.46	12.8	2.30
SUB-TOTAL	2,177	1,173	54	1,002	46	19.1	2.89	12.3	1.64
Reference Group	1,135	618	54	517	46	19.0	2.68	12.6	0.97
GRAND TOTAL	3,312	1,791	54	1,521	46	19.1	3.03	12.4	1.64

As mentioned above both Chinese subjects and norm subjects were randomly divided into two group pairs, with each pair being approximately equal in size. The purpose of this division was to provide for control group comparisons. For the first hypothesis of this study the expectation was stated that each measurement technique under investigation would yield a judgement of equivalence in test performance for two groups of subjects drawn from the same population. In essence, the control group comparisons were used to validate the correct functioning of each measurement technique in determining the equivalence or non-equivalence of test performance. While the use of a measurement technique may show a culture group to differ from the norm group in test performance, groups of subjects drawn from the same population should not be found to differ. This approach of applying a within-group versus between-group comparison was successfully used by Ironson and Subkoviak (1978) in an item-bias study. Note finally, that while two norm groups (1 and 2) were used for making control group comparisons, all comparisons of culture groups to a norm group were made to Norm 1 (for purposes of consistency).

A brief explanation of how the norm group subjects were selected, as well as the purpose of the "reference group" is required. Norm groups (1 and 2) were obtained by randomly selecting subjects from among a total group of "norm" subjects (N=1,875). Norm subjects were identified as those not belonging to any recognized culture group, who declared on the biographical section of test answer sheets that their "heritage" was English, and first language spoken (primary language) was English. The answer sheets of subjects included in data collection but not meeting these criteria were simply not used in the study. Comparisons with the two norm groups were

used, as mentioned, to ascertain the effectiveness of each measurement technique. However, the IRT techniques applied in the study (which are discussed in a later section of this chapter) required sample sizes greater than 1,000 to obtain item parameter estimates. Reference group subjects (N=1,135) were used for this purpose. In addition, by using a separate group of norm subjects for the reference group in order to estimate item parameters (which were used in the computations for all groups), both norm groups (1 and 2) were treated like the culture groups. In the end the two norm groups and reference group were comparable in terms of all key variables.

To ensure the culture and norm groups were as similar as possible, the 3,312 subjects were obtained from the same three geographical areas of the country (Ontario, Manitoba, and British Columbia), from comparable educational backgrounds, fairly balanced according to sex, and were within a limited range of age (see Tables 3 and 4). Subjects were 17 to 24 years of age, and had all completed 11 to 14 years of schooling. All were enrolled, at the time of testing, in a high school, community college, or technical school program. Indeed, all subjects for all groups were taken from the last two or three years of high school (grades 11, 12, or 13), enrolled in a general (neither enriched nor remedial) program, or were taken from the first or second year of community college or technical school. Post-secondary subjects were all enrolled in programs containing academic curricular elements (excluding purely vocational programs such as hairstyling). This restriction in age, educational level, and education program was applied to reduce the potential spread in ability measured by the four sub-tests of the aptitude battery used in the study. Note, as well, that

TABLE 4

Geographical Distribution of the Sample (% by Region by Group)

REGION	G R O U P						
	CHIN 1	CHIN 2	NATIVES (ON-RES)	NATIVES (OFF-RES)	NORM 1	NORM 2	REF
B.C.	57.7	59.0	39.4	42.8	40.1	41.9	42.2
MAN.	18.2	16.6	22.5	25.1	12.0	11.7	13.4
ONT.	24.1	24.4	38.1	32.1	47.9	46.4	44.4
TOTAL %	100.0	100.0	100.0	100.0	100.0	100.0	100.0

while most subjects were under 22 years of age, some older subjects (up to 24 years) were included in the study, in approximately equal numbers for the different groups, to ensure a sufficient sample size for Natives off-reserve.

Two distinct groups of native people were used in this study. The reason for including the two different native people groups was to account for the effect of degree of cultural exposure. This confounding variable is discussed next.

In explaining potential sources of difference in mean aptitude scores of minority and majority groups, the United States Department of Labor, Manpower Administration (1970) noted the possibility that some or all of the difference could be explained by differences in prior experiences. They explained that not all minority group members have had the "cultural exposure" assumed for the majority group. In a paper on inter-cultural evaluation, Laveault (1983) proposed a model of "cultural change" which could be used (in part) to group the subjects of this study according to amount of cultural exposure. Laveault explained that in the interaction of two cultures four results can occur at the cognitive level: (1) integration, referring to the acquisition by the receiving culture of the elements

of the host culture in which the receiving culture's own elements, even the positive ones, are replaced; (2) acculturation, which refers to the acquisition by the members of the receiving culture of all the elements of the host culture without the loss by the receiving culture of its own important cultural elements; (3) imitation, in which the elements of the host culture are reproduced by the receiving culture, but without them being internalized; and (4) isolation, referring to the maintenance of separation of cultural elements. Laveault (1983) suggested that differences in educational and psychological test performance may often be more a reflection of the actual result of cultural interaction (as described above) than of differences in the attribute measured. In this study, an attempt was made to classify subjects who were considered to belong to a particular cultural group (native peoples) into one of the first two categories of the Laveault model, based primarily on duration of exposure to (or immersion in) the "norm" culture. For native peoples place of residence (reserve or not) was the factor used to classify subjects according to duration of exposure to the "norm" (white) culture. In fact, for the native subjects classified as "on-reserve", all were attending reserve schools. Those native subjects were relatively homogeneous in background, generally having spent all of their school years in the same reserve school. Native subjects labeled as "off-reserve", on the other hand, had a history of attendance at various urban schools. They were fairly heterogeneous in terms of school background.

It should be noted that the second and third hypotheses of this study were investigated by comparing the test performance of culture group subjects to that of Norm 1 group subjects. Thus, the test performance of

Native (on-reservation) subjects and Native (off-reservation) subjects was compared independently to the performance of Norm 1 subjects. The effect of the cultural exposure variable was observed, in part, as a difference in the frequency of a non-equivalence finding for the Native (on-reservation) to Norm 1 comparison relative to the Native (off-reservation) to Norm 1 comparison. It was expected that the group further away from the norm group in terms of cultural exposure (Native on-reservation) would be found non-equivalent in test performance more often than the group (Native off-reservation) closer to the norm group. The effect of cultural exposure was also examined by comparing the test performance of subjects in the two native groups to each other. If subjects in the groups were distinct, perhaps as a result of being at different points on the cultural exposure continuum, differences in test performance between the groups would be expected to occur.

Some clarification is needed of criteria applied in selection of subjects for the immigrant (Chinese) culture group. In identifying someone as part of a cultural or subpopulation group where cultural origin is from outside Canada, the chief criterion for inclusion in the group would have to be period of residency in Canada. In other words, it is assumed that those who immigrate to Canada move, in terms of cultural identity, from isolation to imitation to acculturation to integration over a period of time (Gordon, 1964; Greeley & McCready, 1975; Kluckholn, 1967; Laveault, 1983). Thus, a cut-off point in time could be set (somewhat arbitrarily) that would allow classification of an individual as a member of a cultural (immigrant) group up to the cut-off point. For this study, a five-year residency was used as the cut-off point. Thus, all Chinese included in the sample were immigrants to Canada, with arrival in the country within the last five years.

To ascribe significant differences in test performance to cultural group membership alone requires that the effect on test performance of other variables be determined and subsequently controlled. The variables or factors which may be expected to confound the interpretation of study results should be listed. Two such factors have already been discussed, that of amount of cultural exposure and language fluency. Other possibilities are age, sex and educational level.

The characteristics of subjects in the seven groups (six comparison groups and the reference group) with respect to the three variables (sex, age, and educational level) are displayed in Table 3. The proportion of males and females across the groups was approximately equal, with a computed chi-square statistic for the contingency table of sex by group being 2.65. At the .05 level the observed chi-square value was not significant. On the other hand, the groups were significantly different in age (F probability of .0000), and in educational level (F probability of .0000). Although age was different across groups, there is no known effect of age on GATB performance until sometime after age 35 (United States Department of Labor, Manpower Administration, 1970). All subjects in the sample were 17 to 24 years of age. The mean educational level was statistically different across groups. However, the largest difference in mean educational level was 1.2 years. According to findings from studies conducted by the United States Department of Labor, Manpower Administration (1970), education appears to exert an influence on GATB performance only when differences in group means exceeds two full years of schooling. In the end, it can be stated that there probably were no meaningful differences across the groups with respect to age, sex or educational level.

Data Collection Procedures

Test data for the sample were gathered through the administration of the four cognitive and perceptual sub-tests of the GATB to the subjects, concurrently collecting the biographic information needed to conduct the analyses related to the confounding variables. The GATB answer responses were recorded on a separate, machine-readable answer sheet which contains blocks for inclusion of the biographic data. The test administrator had to input the culture group code on all answer sheets following an administration session. Standardized procedures for test administration were followed, so that separate administration of the test battery to each culture group did not introduce a bias (in terms of administrator effect).

The data were collected from February to April, 1988 in three provinces as mentioned earlier: Ontario, Manitoba, and British Columbia. Except for subjects resident on native reserves, all other subjects took the tests in a major urban centre (Toronto, Ottawa, Cornwall, Winnipeg, Brandon, Vancouver). All testing sessions were held in school or college classrooms, with the maximum number tested at one time being 40 subjects. The average class size for testing was 25 subjects. Testing was conducted in public secondary, private secondary and native secondary reserve schools, as well as in community colleges and post-secondary technical schools. Answer sheets were optically read in batches, and the data were transferred by disquette to the University of Ottawa mainframe computer.

Individual subjects were not financially compensated for their participation in the study. School principals, school administrators, teachers and coordinators were asked to encourage the adult students

(subjects) to take the 40 minutes of testing. In many cases donations were made to school funds in order to compensate the students and staff for their time and effort. The purpose of the study and its possible contribution in education were explained to participating schools, and in this way cooperation was obtained.

Data Analysis

In this section, a number of techniques which could be applied under each of the measurement strategies in Figure 2 are reviewed. Justification is given for selecting a certain one or two techniques to represent a strategy. Subsequently, the procedures followed in applying the techniques, in order to support or fail to support the hypotheses stated in Chapter Two, are outlined.

Selection of Techniques

On the following pages, a number of techniques are critically reviewed to determine their potential for use in this study as a means to represent each of the measurement strategies of the modified model for cross-cultural equivalence in test performance (Figure 2). The techniques are discussed under headings which cite the measurement strategy to be represented and the assumption or condition to be challenged by the strategy.

Assumption: Equivalence in construct operationalization

Internal Structure Congruence

According to the model of Figure 2 (Chapter Two), one assumption underlying any statement of cross-cultural equivalence of test scores is that of equivalence in construct operationalization. Whether or not the assumption holds true can be shown by comparing the factor structure obtained for a set of test scores belonging to one group of subjects to that obtained from scores belonging to another group. While the factor structures for different groups could not be expected to be identical, a high degree of congruence would establish at least the equivalence of construct operationalization across cultural groups (Hui & Triandis, 1983; Irvine & Carroll, 1980).

A number of approaches have been used in the past to establish the similarity or congruence of factor patterns across groups. For example, Sung and Dawis (1981) calculated "coefficients of factor congruence" (Harman, 1964) in order to demonstrate the similarity of factor patterns across race and sex groups on sixteen Ball Aptitude Battery (BAB) tests. Applying the Harman coefficient, DeFries, Vandenberg, McClearn, Kuse, Wibon, Ashton and Johnson (1974) reported similar findings on the same tests for Hawaiians of Japanese and of European ancestry. Hennessy and Merrifield (1976) also obtained the same results for high school seniors of Black, Hispanic, Jewish, and Caucasion-gentile stock. Insua (1983) compared, using the Harman coefficient, the factor structures obtained for American and Spanish-speaking (South American) subjects on the WAIS-R, having translated

the WAIS-R into Spanish, and found few differences in factor structure. Somewhat different results were obtained by MacArthur (1975), who discovered different factor patterns for different ethnic samples (Canadian Central Inuit, Greenland Inuit, Nsenga Africans, Alberta Whites) on 29 cognitive ability tests.

The fact that MacArthur (1975) obtained different results to those of Sung and Dawis (1981) is not surprising when both the nature of the tests used in the different studies and the statistical procedures followed are examined. Sung and Dawis sought similarity in factor patterns across three race and two sex groups on a single, multiple aptitude test battery, which included cognitive, perceptual and psychomotor sub-tests. They used exploratory factor analysis to derive 10 factors from the test items, then compared the factor structures for the groups using a procedure recommended by Harman (1964). MacArthur studied the factor patterns for four groups on 29 different cognitive tests, using principal component analysis. MacArthur did not explain how he determined similarity or difference in "factor structure" across the groups.

A comment on the statistical procedures used by Sung and Dawis (1981), DeFries, et. al. (1974), and Hennessy and Merrifield (1976) is needed at this point. All three groups of researchers conducted exploratory factor analysis followed by computation of a "coefficient of congruence" (Harman, 1964). As pointed out recently by McDonald (1985), the use of confirmatory rather than exploratory factor analysis permits the researcher to decide the number and location of factors on rational, substantive grounds. Thus, the results of decisions will commonly be more parsimonius than is the case with traditional results. Moreover, using the common factor analytic methods

suggested by McDonald, an estimate for the goodness of fit of a data set to the hypothesized factor structure is possible. When normality is assumed a chi-square test is also available. McDonald has, in fact, offered a model for use in comparative factor analysis such as that to be undertaken in this study. The model permits the testing of hypotheses concerning the equality in correlation and covariance matrices and factor patterns. It is the Comparative Factor Analysis approach suggested by McDonald (1985), for the reasons he has cited and which are given above, that was used in this study to represent the Internal Structure Congruence strategy.

Assumption: Item equivalence

Response Pattern - Non-IRT Approach

Various researchers have formulated psychometric criteria and derived indices for the comparison of an individual's pattern of test responses (over all items on a test) to that of a group (Tatsuoka & Tatsuoka, 1982a). Some of the same indices have been applied to groups (rather than to individuals alone) in an attempt to identify groups with atypical instruction or experiental histories that alter the relative difficulty ordering of the items (Harnisch & Linn, 1981a). Van der Flier (1982) has suggested that response pattern methods can be used both to identify persons whose test scores are not comparable with the other scores in a certain group, and to classify an individual in such a way that his or her test score is compared with a relevant group (for example, to groups in a developing country westernized to varying degrees). In fact, a variety of

response pattern indices have been developed and applied in studies of test performance differences among examinees (Harnisch, 1983).

In this study the Modified Caution Index (Harnisch, 1983) is used to represent the measurement strategy of Response Pattern - Non-IRT Approach. The Modified Caution Index, like Sato's Caution Index (Sato & Kurata, 1977) from which it was derived, is obtained from the matrix of an examinee's item responses (as 1 or 0) by the group-determined item difficulties (p-values) arranged in descending order. Looking at the item response matrix, it is expected that the individual would get the easiest items correct, the more difficult ones incorrect. The computed Caution Index shows the extent to which the individual's score pattern is Guttman scalable. The larger the index, the greater the departure from a Guttman scale and the less likely the pattern reflects "normal" performance across the items.

The Modified Caution Index was developed in order to avoid having the extreme scores often associated with Sato's Caution Index (Harnisch & Linn, 1981a). The extreme Caution Index values usually occurred when a high scoring examinee missed a single very easy item. The Modified Caution Index yields a lower bound of zero and upper bound of one. The Modified Caution Index can be computed from the following formula.

The Modified Caution Index (C_i^*) for the i^{th} examinee is defined as:

$$C_i^* = \frac{\sum_{j=1}^{N_{i.}} (1-U_{ij})N_{.j} - \sum_{j=N_{i.}+1}^{J} U_{ij} N_{.j}}{\sum_{j=1}^{N_{i.}} N_{.j} - \sum_{j=J+1-N_{i.}}^{J} N_{.j}}$$

where i = 1, 2, ... I, indexes the examinee
j = 1, 2, ... J, indexes the item
U_{ij} = (1 if examinee i answers item j correctly,
0 if examinee i answers item j incorrectly)
$N_{i.}$ = total correct for the i^{th} examinee
$N_{.j}$ = total number of correct responses to the j^{th} item.

The computer program Student Problem Package (SPP) (Harnisch, Kuo & Torres, 1983) was used to compute a Modified Caution Index (C*) for each subject in a group. Differences in mean C* values of groups could then be subsequently determined using analysis of variance (ANOVA).

There are other indices which might be used to represent the response pattern strategy. For example, the personal biserial coefficient (Donlon & Fischer, 1968) has been used to detect atypical test performance for many years. This index is based on the analysis of number-right score patterns. While offering simplicity, the personal biserial index is more confounded with total score than would be desirable for an index designed to identify individuals or groups with unusual response patterns.

In addition to the personal biserial coefficient, Harnisch and Linn (1981b) reviewed the development and implications of seven other response pattern indices. They subsequently examined (empirically) the inter-relationships among the indices, and the relationships of the indices to total score, using data from two tests which form part of a state wide assessment program. The eight indices investigated were the Sato Caution Index (Sato & Kurata, 1977), the Modified Caution Index (Harnisch & Linn, 1981a), the Personal Point-Biserial Index (attributed to Brennan, 1980 and reported in Harnisch and Linn, 1981a), Personal Biserial Correlation (Donlon & Fischer, 1968), Agreement and Disagreement Indices (Brennan, 1980), the Dependability Indices (Kane & Brennan, 1980), the Van der Flier Index (Van der Flier, 1977), and the Norm-Conformity Index of Tatsuoka and Tatsuoka

(1982a). Harnisch and Linn's purpose was to obtain an index, without recourse to item response theory, which could effectively reveal school and regional differences in test score patterns, presumably reflecting differences in curriculum coverage. They concluded that the Modified Caution Index was most appropriate for identifying differences in curriculum coverage across schools and regions based on a comparison of test item response patterns. Their decision was founded chiefly on the fact that the Modified Caution Index showed the least relationship to total test score.

In a later paper, Harnisch (1983) reviews a number of item response pattern approaches (including those employing Item Response or Latent-Trait Theory) from the viewpoint of their application in educational practice. As a means of analyzing ethnic or cultural group differences in test scores, or school differences in curriculum coverage, he argues that comparisons of number-right item response patterns using Sato's Caution Index or the Modified Caution Index appear to be a superior technique (among those not employing Latent-Trait Theory). Those indices are, he argues, sufficiently well developed for practical use and have had broad application. In a more recent study, Miller (1986) obtained results which show how the Caution Index can be used to identify school classes with an atypical pattern of time allocation across content areas.

Harnisch (1983) applied the Modified Caution Index in a study of students participating in the National Assessment of Educational Progress (NAEP) for mathematics. He was able to detect large ethnic group differences on the index. Specifically, he found significantly higher index values for blacks across each of the quintiles of the test score distribution than for non-blacks. Of course, it could be argued that these differences in response patterns occurred because of real differences in

ability, a factor not controlled by this method of analyzing test performance equivalence. However, among all the techniques currently available (using non-IRT test theory) for comparing an individual's test item response pattern to that of a group, and for comparing groups in terms of consistency of response patterns, Sato's Caution Index and the Modified Caution Index seem least affected by total test scores (Harnisch, 1983; Harnisch & Linn, 1981b). They are, perhaps, also least affected by real differences in ability. For those reasons, and because the Modified Caution Index does not have extreme scores associated with it (as does Sato's index), it was selected for use in this study. Once a Modified Caution Index (C*) value was computed for each subject in each group, mean differences in C* among groups were compared using analysis of variance (ANOVA).

Assumption: Item equivalence

Within-Item Analysis - Non-IRT

Ironson and Subkoviak (1979), as well as Hui and Triandis (1985) refer to the use of IRT methods in detecting item bias. However, within-item analysis can also be conducted using various techniques based on non-IRT test theory (Jensen, 1984). Further, Hui and Triandis (1985) indicate that scalar as well as item equivalence can be demonstrated using IRT, but perhaps not with non-IRT, methods. They note that when IRT item-characteristic curves (ICCs) for a test item are compared across groups, examinee responses are automatically placed on a common scale (through the estimation of item and person parameters) (Lord, 1977). This is not the case with non-IRT item bias methods (Dorans & Kulick, 1986).

In this study, two non-IRT techniques are used to represent the Within-Item Analysis strategy. One is based on an analysis of correct/incorrect item responses. It includes a means of controlling for differences in ability level of examinees. The other is based on an analysis of only incorrect responses (distractor choice). It does not include a means for controlling ability differences.

Distractor Analysis Technique

Veale and Foreman (1983) suggest that examinees' responses to incorrect options (distractors or foils) provide more and better information concerning cultural bias than their responses to correct options. They argue that for a particular cultural group, a foil may actually be the correct response, and even if it is incorrect for all groups, it (the foil) may attract or repel members of some groups more than others because it contains culture-specific stimuli. They propose that the presence or absence of cultural variation can be tested by computing a "coefficient of attractiveness or pull" for the foils and comparing the value of the coefficient across identifiable cultural groups.

The procedure developed by Veale and Foreman (1983) provides three levels of analysis to establish the presence or absence of cultural variation in a test. First, cultural variation is defined in terms of the proportions of examinees in each of the various cultural groups selecting each foil, among those answering the item incorrectly. The proportions are called "pull indices", and the hypothesis tested is that there is no cultural variation if the pull indices are equal across cultural groups. Equality of the pull indices (for each test item) is established by

computing a chi-square statistic from the cultural groups by foils contingency table. Second, the degree of cultural variation in an item is determined by computing a statistic (λ), after the method recommended by Goodman and Kruskal (1954) and by computing Cramer's V statistic. The statistic measures the proportional decrease in the expected number of errors in predicting a selected examinee's foil response, obtained by incorporating knowledge of the examinee's cultural group. The λ statistic is needed when the expected frequencies in 20% or more of the contingency table cells are less than five. Clearly, λ does not discriminate between foil response configurations in which groups are most attracted to the same foil but in different degrees. However, the chi-square detects this kind of cultural variation. Kohr (1977) recommends that when using the Veale and Foreman procedure the Cramer's V statistic also be computed for each item. This statistic provides a measure of association or homogeneity of response distributions across groups. The λ index and Cramer's V statistic are computed as shown below:

λ is:

$$\lambda_f = \sum_g \frac{P_{gf} - P_{.f}}{1 - P_{.f}}$$

where g = cultural groups 1, 2, ... g
P_{fg} = largest proportion in the g^{th} row of the contingency table
$P_{.f}$ = largest marginal proportion among the foils (f)

Cramer's V is:

$$V = \left[\frac{\chi^2}{[N \min(G-1, F-1)]} \right]^{\frac{1}{2}}$$

where χ^2=computed chi-square value
N=total number of foil responses
G=Number of cultural groups
F=Number of foils

A third level of analysis has been added by Kohr (1977). It is a total test chi-square and generalized V which can be obtained for an entire set of items, so that equivalence or non-equivalence of the test can be judged. The total test chi-square is simply the sum across items of the computed chi-square values, with degrees of freedom being the sum of the degrees of freedom across all the items. The generalized V is the average V taken across all the items, but weighted by applying the respective culture groups sizes (n).

It has been countered by some that most differences in test performance, revealed by distractor choice analysis, between cultural groups are attributable to real differences in achievement (or ability) levels. For example, Levine and Drasgow (1983) found an orderly relation between ability decile and option choice for many items in an analysis of 9,900 examinee responses to the Graduate Record Examination-Verbal section (GRE-V). They concluded that the pattern of incorrect option choice in a test is probably related to ability. However, conflicting findings were obtained in a study reported by Veale (1988) in which achievement test data of 500 eighth grade students were analyzed. In the latter case, the Veale and Foreman (1983) procedure was shown to detect the presence of cultural variation even when item p-values were homogeneous. Had the item p-values differed greatly across the groups, detection of cultural variation using the Veale and Foreman (1983) procedure might easily have been said to be a reflection of achievement or ability differences across groups.

In using the Veale and Foreman procedure operationally to determine if members of cultural groups differ from those in a population "norm" group in terms of their responses to test items, some test information is lost since only those members of the groups who responded incorrectly to the items are

compared. Both items marked correctly and items omitted are not included in the analysis. For speeded tests, one aspect of the latent ability or skill being measured presumably has to do with speed of response as well as the interaction of speed and accuracy. In just analyzing distractor responses, important differences between cultures in test performance may not be detected when the fact is ignored that some items are omitted because examinees cannot arrive at any conclusion whatever, while others are omitted because examinees do not reach the items within the test time limit. Veale and Foreman (1983) make no mention of this potential problem; however, they do emphasize quite correctly that analysis of distractor responses may provide clues to some of the reasons for differences in test performance across cultural groups. Thus, the Veale and Foreman (1983) procedure described above for analysis of distractor choices is used in the study as one technique representing the Within-Item Analysis strategy of the cross-cultural equivalence model. The technique is applied with some reservation because of the potential confounding of ability with cultural differences, but should provide a point of comparison for another technique applied under this strategy, one which controls for ability level.

Correct/Incorrect Response Analysis Technique

Van der Flier, Mellenbergh, Adler, and Wijn (1984) distinguish between item bias methods which do not take ability differences into account (unconditional) and those which do (conditional). When unconditional methods are applied, bias is defined in terms of the method used. For methods using the item difficulty (p) value, bias is said to exist when analysis of variance applied to an n-item by g-group factorial design with

dependent measure being the p-values or arc sine transformed p-values reveals a significant item by group interaction. For methods which replace p-values by the corresponding standard normal deviate (z-values), bias is again defined as a significant group by item interaction. However, in conditional methods, item bias is defined conditioned on the ability level. In this case an item is not biased if the probability of a correct response is the same for subjects of two or more groups at a given ability level. In item response theory (IRT), item bias is determined by comparing the item characteristic curves of different groups, with bias judged to exist if the curves are not identical, or if a group's item data do not fit the model (Lord, 1977).

In conditional methods not applying IRT, the observed score is typically but not necessarily used as the ability index. Ironson (1982) describes two non-IRT, conditional methods for detecting item bias, the chi-square approaches and loglinear models. Two other approaches are also possible, the standardization approach and the Mantel-Haenszel procedure. (Dorans & Kulick, 1986; Holland & Thayer, 1986; Kok, Mellenbergh, Van der Flier, 1985; Van der Flier, Mellenbergh, Adler & Wijn, 1984). It is a Mantel-Haenszel (MH) statistic that is used in this study for the Correct/ Incorrect Response Analysis Technique. The derivation of the MH statistic used, and justification for its selection, are provided on the following pages.

The Mantel-Haenszel (MH) procedure for detecting differential item performance (DIP) across two groups of examinees can be considered a natural outgrowth of the various chi-square procedures (Marascuilo & Slaughter, 1981; Mellenbergh, 1982; Scheuneman, 1979). The procedure was developed by Mantel and Haenszel (1959), and has been widely used in the biomedical field. Its use in educational and behavioral science research is a relatively recent occurrence.

The MH procedure is based on 2X2 contingency tables for each ability category. In practice, ability is measured by the total test score, with the score divided into K intervals or levels. Unlike the usual chi-square methods, the MH procedure permits testing from the contingency table of the null hypothesis (Ho) of no DIP across two groups against a specific alternative hypothesis (H_1). Thus, the MH chi-square statistic (MH-CHISQR) is distributed as a chi-square with just one degree of freedom. MH-CHISQR provides, therefore, the uniformly most powerful unbiased test of Ho versus H_1 (Cox, 1970). It corresponds but is not identical to the single degree of freedom chi-square test used with the iterative logit method (Mellenbergh, 1982) for testing of "no bias" against the alternative of "uniform bias". The MH procedure is not, however, iterative.

For K 2X2 tables of the form shown below, the MH null and alternative hypotheses are as follows:

Score on Item

GROUP		1	0	TOTAL
	R	PR_j	QR_j	1
	F	PF_j	QF_j	1

K 2X2 tables

where R is the reference group and
F is the focal group

Null Hypothesis (Ho) is

$$\frac{PR_j}{QR_j} = \frac{PF_j}{QF_j} \qquad J = 1, 2, \ldots K$$

Alternative Hypothesis (H_1) is

$$\frac{PR_j}{QR_j} = \alpha \frac{PF_j}{QF_j} \quad J = 1, 2, \ldots K$$

for $\alpha \neq 1$

Clearly, $\alpha = 1$ in H_1 corresponds to Ho.

The MH-CHISQR is then given by,

$$\text{MH-CHISQR} = \frac{(|\sum_j A_j - \sum_j E(A_j)| - \frac{1}{2})^2}{\sum_j \text{Var}(A_j)}$$

where $\text{Var}(A_j)$ is defined by

$$\text{Var}(A_j) = \frac{NR_j \, NF_j \, Ml_j \, MO_j}{T_j^2(T_j - 1)}$$

and T_j is the total number of reference and focal group members in the J^{th} matched set, NR_j is the number of those who are in R (the reference group) and of these A_j answered the item correctly. NF_j is the number in F (the focal group), while Ml_j is the total number of correct responses to the item for the J^{th} matched set and MO_j is the number of incorrect responses (Holland & Thayer, 1986).

In the MH procedure an estimate of α can also be obtained. In fact, α is the common odds-ratio across the 2X2 tables, that is, this estimator compares the odds of success for the two groups on the item under study, when the groups are matched on K categories of ability (taken from their total test scores). The estimator is given by

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

where T_j and A_j are as defined above,
C_j represents the number of examinees in F answering the item correctly
B_j and D_j are the number of examinees who responded incorrectly in the R and F groups respectively.

In essence, the reference group is used to establish a standard against which the performance of the focal group is compared. The scale of the estimator (αMH) is from zero to infinity, with α = 1 being the null value. According to Holland and Thayer (1986): "The value of αMH is the average factor by which the odds that a member of R is correct on the studied item exceeds the corresponding odds for a comparable member of F. Values of αMH that exceed one correspond to items on which the reference group performed better on average than did comparable members of the focal group". Values of αMH less than one signal items on which the focal group did better than comparable members of the reference group.

Holland and Thayer also suggest that a log transformation of αMH can be made to place the estimator on the ETS delta scale. The transformation is

$$\Delta MH = -2.35 \ln(\alpha MH).$$

ΔMH is interpreted as the average amount more difficult that a member of R found the studied item than did comparable members of F. Negative values of ΔMH indicate that R group members found the item easier than F group members.

In relation to the chi-square methods for detecting DIP such as those offered by Scheuneman (1979), Marascuilo and Slaughter (1981), and Shepard, Camilli, and Williams (1985), the MH procedure has the advantage of providing a single degree of freedom chi-square test of the null hypothesis (of no DIP) against a specific alternative hypothesis. The usual chi-square tests are multi-degree of freedom, so that statistical power is often weak.

Power is typically increased in such cases by reducing the number of score categories, and thus increasing the sample size per category. An undesirable consequence of score category reduction is diminished quality in ability matching. A second advantage of the MH procedure over typical chi-square approaches is that with MH the degree of bias, or DIP, in an item can be measured by use of the α MH or Δ MH parameter. However, the MH procedure, when only the chi-square value is used (MH-CHISQR), is affected by the well-known problem with chi-square tests of falsely rejecting the null hypothesis due to large sample sizes. With the MH procedure it also must be assumed that non-uniform bias does not exist.

Comparing the MH procedure to the standardization method (Dorans & Kulick, 1986) shows that the two approaches measure very similar phenomena. Whereas standardization compares the probabilities of success for two groups, matched on K ability levels, the MH procedure compares the odds of success for the two groups. In a comparative study of the two methods for detecting DIP, using real life data, Wright (1986) shows that the MH estimator (Δ MH) and standardization statistic (Dstd) are highly related. Correlation between the two estimators of DIP ranged from .94 to 1.00 across samples of differing size and using different numbers of score intervals. Wright concludes that further research is needed to determine what practical differences in the two methods might be important. He does note that the MH procedure is particularly suitable for situations with sparse data, meaning relatively small 2X2 tables (low frequencies).

In an examination of the consistency across several statistical procedures used to detect DIP between Irish and American students, Welch, Akerman, Doolittle, and Hurley (1987) discovered that there is a great deal of overlap among the procedures (which included Angoff's delta-plot, a

modified Angoff, two IRT methods, ordinary least squares regression, principal component analysis, and the MH procedure). They reported correlations amongst the DIP statistics ranging from .51 to .97, with the average correlation being .91. Most methods produced practically the same results (that is, they identified most of the same items as having or not having DIP). In contrast to both iterative and IRT methods, the MH procedure does not require calibration of items to obtain parameter estimates and thus costs less to use.

In summary, several advantages for using the MH procedure to detect DIP over chi-square, iterative logit, or standardization methods can be cited. In relation to other chi-square approaches, the MH procedure uses a single degree of freedom chi-square test, providing greater statistical power. With the MH procedure, degree of DIP can also be measured using either the α MH or Δ MH estimator. Compared with the iterative logit approach, calibration of items to obtain parameter estimates is not needed with the MH procedure, so MH is cheaper to use. The MH procedure can also be used with small sample sizes (N=250), whereas both the iterative logit and standardization methods require larger samples (usually over 500 subjects) (Carlson & Sarrazin, 1984). Because of these advantages, and its current popularity of use in educational measurement, the MH procedure is used in the current study. Specifically, the Δ MH estimator of DIP is applied, in preference to the α MH index. The Δ MH is preferred because it places DIP measures on a symmetric scale with zero as the null value. In the next section (Procedures Using Techniques), criteria are given for determining when the ΔMH value of an item is indicative of DIP. Further, a means of establishing equivalence or non-equivalence of a whole test (an entire item set) for two groups using ΔMH is provided in the next section.

Assumption: Item equivalence and scalar equivalence

Response Pattern - Item Response Theory (IRT) Approach

Item response theory (IRT) is used by researchers to solve a variety of measurement problems that are very difficult to address with methods drawn from classical (non-IRT) theory. Appropriateness measurement is one area where IRT has recently been applied. In using IRT, rather than non-IRT approaches, it is necessary to first specifically define what is meant by an "appropriateness index". It is defined as a measure of goodness of fit of a very general psychometric model to the individual examinee's item-by-item pattern of responses. Levine and Rubin (1979) suggest that IRT appropriateness indices are most useful as broad, low-power screening devices for identifying a proportion of examinees whose test responses require further, specific analysis.

Appropriateness measurement involves a two-step procedure in which item parameter estimation occurs first followed by person measurement where an appropriateness index is computed. This procedure is similar to the practice in IRT of estimating parameters (difficulty, discrimination, guessing) characterizing items or the test as a whole, and then of estimating person characteristics such as ability (Levine & Drasgow, 1983). In the first step, item parameters and the probability of a correct response to an item by an examinee with a given ability ($P_i(\theta)$ according to Levine & Drasgow, 1983) are estimated using any of the one, two, or three-parameter logistic formulae (Birnbaum, 1968). Based on the values $P_i(\theta)$ the probability of a specified (expected) pattern of item responses is obtained. In computing the probability of an expected pattern which is conditioned on

ability, either a "constant ability" model or "variable ability" (Gaussian) model may be applied. In the constant ability model, a person's ability is considered to be constant across all items of the test. For the Gaussian model ability is allowed to vary from item to item. In the second step of the process, the objective is to determine if any of the formulae (for the expected pattern of responses) generated in step one fit the person's data. The degree of fit is measured by an appropriateness index.

A number of IRT appropriateness indices have been developed and studied using both simulated and real-life data. For the purposes of this study, two of these indices (based on a constant ability model) are selected for use. They are the L_z and ECIZ4 indices (Drasgow, Levine & Williams, 1987; Tatsouka & Tatsouka, 1982b). The derivation of these indices is explained below.

The L_z index was derived from another appropriateness index, Lo. Levine and Drasgow (1983) have explained the deriviation of Lo as follows:

"Suppose that for all values of θ, a pattern U* appears improbable in the sense that Prob(U*$|\theta$) is very small. Then, U* is badly fit by all sub-models for individual data developed in the test norming stage" (step one).

Aberrance of an individual's pattern of item responses is indicated by a relatively low maximum of the function of θ, Prob(U*$|\theta$). An atypical examinee pattern is expected to be relatively low because it is not likely that high-ability people miss easy items or low-ability people pass hard items. The index (Lo) is calculated from:

$$Lo(U^*) = \max_{\theta} \text{Prob}\ (U^*|\theta)$$

The Lo index, as first proposed by Levine and Rubin (1979), is considered by Levine and Drasgow (1983) to be one of the simplest, yet

effective, aberrance indices. Nonetheless, the effect on the index of omitted responses seems to be a problem. Indeed, examinees omitting different items are in some way taking different tests. A low value of Lo, for example, obtained by an examinee who omitted a substantial number of items, may be less indicative of aberrance than a higher Lo value achieved by an examinee who omitted fewer, and different, items. This problem of omitted items has been alleviated by a standardization process. Drasgow, Levine, and Williams (1987) have shown that the "standardization process" is useful in accounting for item omitting as well as in controlling the confounding of ability and appropriateness.

Drasgow, Levine, and Williams (1987) described a new index (L_z) based on the transformation of the Lo index to a standard normal distribution. The detection rates of the new index for both actual and simulated data were compared, as well as the performance of the index under conditions of unrestricted and restricted omitting. They showed that L_z has an empirical distribution reasonably close to the standard normal distribution, but that the distribution was not completely independent of estimated ability. However, the effects of this lack of complete independence from estimated ability were fairly small. Similarly, little difference in detection capability arose under the two conditions of omitting.

The L_z index is computed as follows (Drasgow & Guertler, 1987):

$$L_z = \frac{Lo - E(Lo)}{[Var(Lo)]^{1/2}}$$

where $E(Lo) = \sum_{i=1}^{n} (P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + [1-P_i(\hat{\theta})] \ln [1-P_i(\hat{\theta})])$

and $Var(Lo) = \sum_{i=1}^{n} P_i(\hat{\theta}) [1-P_i(\hat{\theta})](\ln[P_i(\hat{\theta})/(1-P_i(\hat{\theta})])^2$

Another standardized IRT appropriateness index, which is analogous to the modified caution index yet provides for a control on ability (through the use of IRT procedures) is the "standardized extended caution index" or ECIZ4 (Tatsuoka & Tatsuoka, 1982b). The ECIZ4 index provides a measure of the degree to which the pattern of item difficulties in the general examinee population covaries with the item difficulties experienced by a particular examinee as indicated by his or her pattern of right and wrong answers. The standardized extended caution index is calculated from (Drasgow & Guertler, 1987):

$$ECIZ4 = \frac{\sum [(P_i(\hat{\theta})-U_i)(P_i(\hat{\theta})-\bar{P})]}{[\sum P_i(\hat{\theta})(1-P_i(\hat{\theta}))(P_i(\hat{\theta})-\bar{P})^2]^{1/2}}$$

where $\bar{P}$ is the mean probability of correct responses at ability θ over the test items and $P_i(\hat{\theta})$ is the probability of answering the i^{th} item correctly at estimated ability level θ.

Because ECIZ4 is a standardized index which provides for a control on ability, yet is analogous to the modified caution index, it is used in this study to represent the Response Pattern - IRT Approach Strategy, in addition to the L_z index of Drasgow, Levine, and Williams (1987).

Using simulated data for the Scholastic Aptitude Test (SAT), Levine and Rubin (1979) evaluated the Lo index (from which L_z is derived). They posed two questions which need to be answered before any appropriateness index can be used in practice. The questions were:

(1) What is the effect on aberrance indices of using estimated item parameters?

(2) What is the effect of estimating item parameters from samples containing aberrant examinees (which is certainly the usual case)?

In a series of four empirical studies, using both simulated and actual test data, Levine and Drasgow (1983) studied the questions posed above. In

response to the first question about the effect of using estimated item parameters on aberrance detection, they found close agreement between appropriateness indices computed with simulated parameters and with estimated parameters. Using estimated parameters, they concluded, would not significantly alter detection rates. In examining solutions to the second question, they found that including unidentified aberrants in the screening sample did not seriously degrade item parameter estimates. Further, they showed that the three-parameter logistic model provided high detection at low false-alarm rates (rate of identifying examinees as "aberrant" when their response patterns are, in fact, "normal"). Unfortunately, no explanation was given as to what constituted "high" detection. The model was successfully used to identify spuriously low examinees.

In using the IRT strategy to investigate cross-cultural equivalence, three additional issues arise. The first is the choice of an IRT model for multiple-choice tests. The second is the relative effectiveness of various IRT appropriateness indices. The third is the issue of sample size needed for parameter estimation as applied to computation of appropriateness indices. Using Graduate Record Examination-Verbal (GRE-V) item test results, Drasgow (1982) examined all three issues. He found that appropriateness measurement does not seem greatly affected by the differences between the Rasch and three-parameter logistic model. However, the three-parameter model produces generally better detection of spuriously low examinees. The use of moderately small sample sizes (N of approximately 300) also did not degrade appropriateness measurement. Appropriateness measurement appeared feasible, then, in much smaller samples than previously believed. Also, as shown by Levine and Drasgow (1983), good aberrance

detection with actual and simulated data could be obtained with any of the appropriateness indices despite model misspecification and parameter-estimation error.

In conclusion, it is pointed out that because of the merits of a "standardized" appropriateness index over any of the non-standardized ones, the L_z index is used, as mentioned above, in the present study. Another standardized index, ECIZ4, which controls for ability differences, is also used because of its analogy to the Modified Caution Index.

Although both the L_z and ECIZ4 indices are normally used to indicate the aberrance of an individual examinee's response pattern from an expected pattern, they can be used in group comparisons. This is done by computing a mean L_z and ECIZ4 value for the groups compared (from individual examinee index values), then by comparing means using analysis of variance (ANOVA). Further details of this procedure are provided in the next section.

Procedures Using Techniques

In order to test the hypotheses of this study, which were stated in Chapter Two, a two-stage procedure was needed. In the first stage, the first hypothesis was challenged. It should be recalled that the first two hypotheses relate to the usefulness of each of the measurement strategies included in the modified model (Figure 2), as represented by specific measurement techniques, in determining the equivalence or non-equivalence of a single test for a culture group in relation to a norm group. Differences in test performance (non-equivalence) between members of groups could be found as a result of a technique or techniques not functioning correctly rather than as a consequence of true cultural differences. Thus, the

proposition (hypothesis one) that two samples from the same culture or norm group should not be different in test performance was first tested. If the two parts (control group pairs) were found to be equivalent in test performance using a specific technique the measurement technique was assumed to function correctly. In this case, the test performance of two groups of Chinese (Chinese 1 and Chinese 2), and two norm groups (Norm 1 and Norm 2) were compared for each of the four GATB sub-tests listed in Table 1. Each of the techniques, representing the measurement strategies of the modified model, and shown in Table 5 below, were used in the comparisons. Note that control group pair comparisons were made, as were all other comparisons for each of the four GATB sub-tests separately, treating each sub-test as a single test.

TABLE 5

Measurement Techniques Applied in the Study and the Strategies Represented

MEASUREMENT TECHNIQUES USED	MEASUREMENT STRATEGY REPRESENTED
Comparative Factor Analysis	- Internal Structure Congruence
Modified Caution Index (C*)	- Response Pattern - Non-IRT Approach
Distractor Analysis Correct/Incorrect Response Analysis (Δ MH)	Within-Item Analysis - Non-IRT
Standardized Appropriateness Index (L_z) Standardized Extended Caution Index (ECIZ4)	Response Pattern - IRT Approach

In the second stage of procedures, the second and third hypotheses were challenged. It will be recalled that under the second hypothesis an expectation of some differences for different culture groups in test performance shown by use of some of the techniques (representing the various measurement

strategies of the modified model) was stated. This expectation was investigated by applying each of the techniques in Table 5, separately for each GATB sub-test, in an analysis of item or test responses for the four culture groups (Chinese 1, Chinese 2, Native on-reserve, Native off-reserve) in relation to a norm group (Norm 1). Also, the performance of the two native groups was compared in each case to help determine the effect of the confounding variable, degree of cultural exposure.

For the third hypothesis, the measurement strategies of the model, as represented by the specific techniques in Table 5, were related to the continuum of equivalence assumptions, or conditions. Under the hypothesis, it is postulated that non-equivalence in test performance of a culture group with the norm, demonstrated using a technique representing a measurement strategy at the top of the strategies array (see Figure 2 of Chapter Two), should imply non-equivalence for the same groups on the same sub-test of the GATB being demonstrated using all other techniques representing measurement strategies lower in the array. As well, techniques representing the same strategy should for the same group comparisons, lead to the same judgement of equivalence or non-equivalence in test performance. The third hypothesis was tested by aggregating the results (in terms of equivalence or non-equivalence decisions) from stage one procedures. Then the patterns of decisions reached were examined. To examine the patterns of decisions, a group by decision by measurement technique table (like that shown in Table 6) was constructed.

Somewhat different analytical methods were followed in applying each of the measurement techniques discussed in the previous section. The different methods are outlined below.

TABLE 6

Example Decision Table
Resulting from Application of the Modified Model

Test: (GATB Sub-Test Used)

0 = non-equivalence
1 = equivalence

MEAUREMENT TECHNIQUE	GROUP COMPARISON MADE						
	CHIN 1 VS CHIN 2	NORM 1 VS NORM 2	CHIN 1 VS NORM 1	CHIN 2 VS NORM 1	NATIVE (ON-RES) VS NORM 1	NATIVE (OFF-RES) VS NORM 1	NATIVE (ON-RES) VS NATIVE (OFF-RES)
Comparative Factor Analysis	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1
Modified Caution Caution Index (C*)	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1
Distractor Analysis	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1
Correct/Incorrect Response Analysis (Δ MH)	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1
Standardized Appropriateness Index (L_z)	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1
Standardized Extended Caution Index (ECIZ4)	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1

Comparative factor analysis

In the previous section of this chapter, a description of techniques which may be used to represent the Internal Structure Congruence strategy of the modified model was given. It was concluded at that juncture that the general approach to determining congruence of internal (factor) structures recommended by McDonald (1985) would be used in this study. McDonald has suggested that hypotheses concerning the equality of correlation and covariance matrices first be tested, and if inequality is found, then the equality of factor structures may be tested.

Thus, a two-step procedure was followed in this study to enable a judgement of equality or inequality of internal (factor) structure for the matrix of correct-incorrect test responses of examinees belonging to two different groups (for example, Chinese 1 and Norm 1). In the first step, the intercorrelation matrix of item responses (coded as 1=correct, 0=incorrect) in a single test was computed for the two groups being compared. The hypothesis of equality of the correlation matrices was then tested. If the matrices were found not to be equal, a second step consisting of a determination of a factor structure for the item responses of the two groups, then a simultaneous comparison of the factor structures (not the factor loadings) of the two groups, was performed.

It should be noted at this point that in the first step described above, the computation of correlation matrices, the use of ordinary product moment correlations is not recommended. Because the observed (dependent) variables are categorical (0=incorrect, 1=correct for each test item response), it is preferable to compute estimates of tetrachoric correlations, and then analyze the matrix of such correlations by the

Unweighted Least Squares (ULS) method, rather than by Maximum Likelihood (ML) methods (Jöreskog & Sörbom, 1985). When ULS methods are used a chi-square test of goodness-of-fit (for one group's correlation matrix in relation to another) is not possible. Instead, other measures of "fit" must be used to judge equality or inequality of the correlation matrices of the groups compared. Jöreskog and Sörbom (1985) provide several means of assessing "fit" in their program, Lisrel VI. Three of the measures available in the program (described below) were used in this part of the current study.

The three measures to assess "fit" used in this study are a Goodness-of-Fit index (GFI), the root mean square residual (RMSR), and the magnitude of non-standardized residuals.

The GFI can be considered a measure of the relative amount of variances and covariances jointly accounted for by the model. Unlike chi-square statistics, GFI is independent of the sample size and relatively robust against departures from normality. However, the distribution of the statistic is unknown, so there is no stated standard to judge what constitutes good or poor fit (Jöreskog & Sörbom, 1985). Based on general statements made by others examining data similar to that used in this study, a value of GFI greater than or equal to .90 is taken as a criterion for "good fit".

The RMSR can be considered a measure of the <u>average</u> of the residual variances and covariances. The value of RMSR is interpreted as implying "good fit" or "poor fit" in relation to the sizes of the observed variances and covariances. By looking at data sets similar to those used in this study, a value for RMSR of less than or equal to 0.120 is taken as an indication of "good fit". For the actual matrix of residuals

(non-standardized, obtained after "fitting") a criterion is set (somewhat arbitrarily) of 10% of residuals greater than or equal to .200 as constituting a poor fitting model.

In summary, the first procedural step in comparative factor analysis, as applied in this study, was the computation and comparison of the tetrachoric correlation matrices for the item responses (0 or 1) of two groups. If all of the following conditions were met, the matrices were judged equal, otherwise they were considered not equal, and a second step was required.

1) GFI $\geq$ .90
2) RMSR $\leq$ .120
3) (% Residuals $\geq$ .200) $\leq$ 10%

Should the hypothesis of equal correlation matrices be rejected, a simultaneous comparison of factor structures was made, as described earlier. This additional step is only necessary when and if equality of correlation matrices is not found. The criterion for judging "good fit" of the factor structures is the same as the criteria (above) used for comparison of the correlation matrices.

It can be seen that comparative factor analysis was conducted separately for each GATB sub-test and for each group pair given in Table 6. Upon completion of the analyses for any one GATB sub-test, values of equivalence (=1) or non-equivalence (=0) of the sub-test for each group pair compared were entered into a decisions by measurement techniques table (like Table 6).

The modified caution index

The Modified Caution Index (C*) was computed separately for each GATB sub-test and for each subject. The value of C* is based on a comparison of

subjects' number right response patterns (across an entire sub-test) to a pattern derived for subjects in the six culture and norm groups on the sub-test (regardless of individual culture group membership), as explained in the previous section. Then the C* values for subjects could be averaged across each group (culture or norm), and the mean C* values for the six groups compared. The Student Problem Package (SPP) computer program (Harnisch, Kuo, & Torres, 1983) was used to compute the Modified Caution Indices (C*). The program uses the item response matrix (0 or 1) for all subjects in the six groups (N=2177) to establish item difficulty (p) values, arrange items in ascending order of difficulty and subjects' total scores (sum of items correct) in descending order and compute an index of departure from the "expected" pattern of responses. The Modified Caution Index thus calculated has a lower bound of zero and upper bound of one.

To determine if groups differ from one another in terms of average C*, analysis of variance (ANOVA) was used (subject to meeting all the assumptions required for ANOVA). If the omnibus test indicated there was a significant difference among groups (at the .05 level), post-hoc procedures (Scheffe's) were conducted. Pairs of groups which were significantly different in C* value were then identified. This process allowed for simultaneous comparisons of Chinese groups 1 and 2, Norm groups 1 and 2, Native (on-res) and Native (off-res), as well as each culture group (Chinese 1 and 2, Native on and off reserve) with the Norm 1 group. A significant difference (at the .05 level) in C* values between groups was taken to imply non-equivalence of the sub-test across the groups. Once analysis was completed for a GATB sub-test, values of equivalence (=1) or non-equivalence (=0) for each group pair compared were entered into a decisions by measurement techniques table (see Table 6).

Standardized appropriateness index (L_z) and standardized extended caution index (ECIZ4)

These two measurement techniques are applied to the subjects' vectors of correct (1) and incorrect (0) responses to each of the four GATB sub-tests (separately) in a three-step procedure.

Step One

For each sub-test, item difficulty, discrimination, and guessing parameters were estimated for the three-parameter logistic model of IRT using item responses for a reference group of "norm" subjects (N=1135). This reference group was derived from the same pool of norm subjects as Norm groups 1 and 2, as described earlier in the section titled The Sample. Parameter estimation was achieved using the LOGIST computer program (Wingersky, Barton, & Lord, 1982).

Step Two

Two standardized IRT appropriateness indices (L_z and ECIZ4) were computed for all subjects in all groups (N=2177) following the computational procedures described in the previous section. The computer program used to compute L_z and ECIZ4 was one written by Drasgow (1987).

Step Three

Analysis of variance (ANOVA) was applied (subject to meeting all needed assumptions) to the L_z and ECIZ4 scores for all subjects across the six

groups (Chinese 1 and 2, Native on and off reserve, Norm 1 and 2). If the omnibus test indicated the presence of a significant "group effect" (at the .05 level), post-hoc procedures (Scheffe's) were followed to identify which pairs of groups had significantly different L_Z or ECIZ4 values. A significant difference in the L_Z or ECIZ4 values was taken to imply non-equivalence for the sub-test across the relevant groups for that measurement technique (L_Z or ECIZ4). Values of equivalence (=1) or non-equivalance (=0) were subsequently entered into a decisions by measurement techniques table (see Table 6).

Distractor analysis

The procedures cited in the previous section were followed to compare the pattern of distractor choices of each culture group with the Norm 1 group. To determine if the measurement technique was functioning correctly, the distractor choice patterns of the two Chinese groups were compared, as were those of the two norm groups. For each GATB sub-test the computer program CULVAL (Kohr, 1977) was used to compute a total test chi-square (χ^2) statistic and a generalized Cramer's V statistic for each pair of groups compared. Computation of these two statistics for the total test from item by item comparisons was explained in the previous section. In all, the seven comparisons shown below were made.

(1) Chinese 1 with Chinese 2

(2) Norm 1 with Norm 2

(3) Chinese 1 with Norm 1

(4) Chinese 2 with Norm 1

(5) Native (on-res) with Norm 1

(6) Native (off-res) with Norm 1

(7) Native (on-res) with Native (off-res)

The decision rules applied to establish equivalence or non-equivalence for each group pair on each sub-test using the total test χ^2 and generalized Cramer's V statistics were those recommended by Kohr (1977) and Veale and Foreman (1983). The criteria for rejection of the hypothesis of "no cultural variation", that is, rejection of the equivalence of the sub-test across the two groups compared, are:

$$\chi^2 \leq .05 \text{ and } V \geq 0.15$$

The results from the above analyses in terms of equivalence (=1) or non-equivalence (=0) of test performance was entered into a decision by measurement techniques table (see Table 6).

Correct/incorrect response analysis (Δ MH)

As discussed in the previous section, the Mantel-Haenszel (MH) procedure for detecting differential item performance was used in this study as one technique representing the Within-Item Analysis strategy for determining cross-cultural equivalence. Specifically, the Δ MH statistic was selected to test the hypothesis of differential item performance (DIP). As defined by Holland and Thayer (1986), ΔMH is interpreted as the average amount more or less difficult that a member of a reference (R) group found the studied item than did comparable members of a focal (F) group. Negative values of Δ MH indicate that R group members find the item easier than F group members. Positive values indicate that R group members find the item more difficult than F group members. The estimator ΔMH is a log transformation of α MH, with the transformation putting the statistic on the ETS delta scale. The Δ MH has an expected value of zero if no DIP is present.

To determine if a whole test was equivalent (no DIP present) or non-equivalent (DIP is present) for each pair of groups compared (Table 6), a criterion was needed to first establish for a computed value of ΔMH if DIP was present in the test item. While there is no single value for Δ MH which everyone agrees signifies DIP, Educational Testing Services has apparently been using Δ MH critical values of -0.65 to +1.53 (Hambleton & Rogers, 1988). To provide an interval that gives equal weight to DIP which negatively as well as positively affects the focal group, Arrasmith and Dizinno (1986) suggest an acceptance range (for no DIP present) of Δ MH between -1.50 and +1.50. The latter values were used in this study.

Thus, for each pair of groups compared using Δ MH on each GATB sub-test, test items were declared equivalent (no DIP) if:

$$-1.50 \leq \Delta \text{ MH} \leq +1.50$$

Based on the above criterion, items in the same sub-test may be biased for (positive values of Δ MH) or against (negative Δ MH) focal group members. Thus, to determine if a <u>whole sub-test</u> contains DIP which is in favour, or not in favour, of focal group members, some criterion is needed. The criterion must take into account the fact that some items may be found to have positive Δ MH values outside the +1.50 range, and others negative Δ MH values outside the -1.50 range. For the purposes of this study, a sub-test was labeled as containing DIP (non-equivalent for the groups compared) if the <u>absolute difference</u> in the percentage of items having positive Δ MH values greater than +1.50, as opposed to those with negative Δ MH values less than -1.50, equalled or exceeded 10%. For example, if a 20 item test is found to have two items with Δ MH values greater than +1.50 (10% positive DIP) and five items with Δ MH values less than -1.50 (25% negative DIP), the absolute difference is 15%. Such a difference was taken

as constituting significant DIP which was not in favour of (biased against) the focal group. This critical value (10% difference) is, admittedly, somewhat arbitrary, but Martois, Pickard, and Stiles (1988) report such values as frequently occurring on state-wide achievement tests suspected of being biased against minorities.

Once analyses were completed for all comparison groups, as outlined above, rejection of the no DIP hypothesis was coded as non-equivalence (=0) in test performance for the groups compared, while failure to reject the hypothesis was considered as equivalence (=1). The values for equivalence or non-equivalence were subsequently entered into a decision by measurement techniques table (see Table 6).

CHAPTER IV

PRESENTATION AND DISCUSSION OF THE RESULTS

The results of this study are presented in two sections. In the first section, the results are given from analyses of the sample data using each measurement technique applied in this study. Follow-up analyses, sometimes necessitated by the results of initial analyses for a particular technique, are also described in the first section. Note that the results obtained using each measurement technique are not presented (in Section One) in the order the measurement techniques appear in the model of Figure 2 (i.e., from top to bottom). Results obtained using the Standardized Appropriateness Index (L_Z) and the Standardized Extended Caution Index (ECIZ4) are given immediately following the display of results for the Modified Caution Index (C*). This is done to facilitate comparison of the techniques, all of which are based on a Response Pattern Analysis approach. In the second section, results are displayed which were obtained by combining data from analyses reported in the first section.

Following the presentation of results, the findings of this study are discussed in relation to each of the three hypotheses stated in Chapter Two. The discussion of results is divided, then, into three sections, each given a heading to reflect the hypothesis under consideration. Overall findings are then finally presented.

Section One Results

After initial analyses of the sample data for the four GATB sub-tests used in this study, it became evident that the test lengths (number of items

in each sub-test) shown in Table 1 had to be adjusted to reflect the reality of the sample's test performance. Because the sub-tests were each timed, there were some items in some sub-tests not reached by anyone. Inclusion of those items in analyses would only distort results, so the actual test lengths (number of items) used in analyses were truncated to the test lengths shown in Table 7. It should be noted that test reliability was maintained in spite of the reduced test lengths. The internal consistency reliabilities (KR-20) of the truncated sub-tests were given earlier in Table 2 (Chapter Three).

TABLE 7

General Aptitude Test Battery (GATB) Revised Description of Sub-Tests

TEST	ORIGINAL NO. OF ITEMS	NUMBER OF ITEMS USED	TIME (MINUTES)	LANGUAGE EFFECT
Computation	50	41	6	NO
Three Dimensional Space	40	40	6	NO
Vocabulary	60	46	6	YES
Arithmetic Reasoning	25	20	7	YES

In Table 8, the means and standard deviations of the raw scores for each group included in the study, and for each sub-test, are displayed. Subsequently, the results are given of the analyses of the data using each measurement technique (separately).

TABLE 8

Means and Standard Deviations of Sub-Test Raw Scores - By Group

Sub-Test Computation

GROUP	MEAN	STD DEV	CASES
Chinese 1	26.06	4.38	333
Chinese 2	25.71	4.39	347
Native (On-Res)	20.13	4.66	386
Native (Off-Res)	17.33	5.58	371
Norm 1	20.95	4.72	366
Norm 2	21.03	4.65	374
Reference	21.36	4.64	1135
TOTAL	20.76	6.37	3312

TABLE 8 (cont'd)

Sub-Test Three Dimensional Space

GROUP	MEAN	STD DEV	CASES
Chinese 1	24.03	6.25	333
Chinese 2	23.61	6.49	347
Native (On-Res)	19.59	5.31	386
Native (Off-Res)	17.83	6.32	371
Norm 1	20.14	6.16	366
Norm 2	20.27	6.15	374
Reference	20.64	6.16	1135
TOTAL	20.76	6.37	3312

Sub-Test Vocabulary

GROUP	MEAN	STD DEV	CASES
Chinese 1	11.07	5.18	333
Chinese 2	11.22	5.38	347
Native (On-Res)	15.96	5.34	386
Native (Off-Res)	14.11	6.25	371
Norm 1	20.06	6.12	366
Norm 2	20.54	6.04	374
Reference	20.66	6.56	1135
TOTAL	17.35	7.17	3312

Sub-Test Arithmetic Reasoning

GROUP	MEAN	STD DEV	CASES
Chinese 1	10.86	2.75	333
Chinese 2	10.65	2.52	347
Native (On-Res)	9.00	2.86	386
Native (Off-Res)	7.76	3.03	371
Norm 1	10.31	2.76	366
Norm 2	10.26	2.80	374
Reference	10.42	2.89	1135
TOTAL	10.00	2.98	3312

Comparative Factor Analysis Technique

As described in Chapter Three, a two-step procedure was followed in applying the comparative factor analysis approach recommended by McDonald (1985). In the first step, the inter-correlation matrix (tetrachoric

correlations) of subjects' responses (coded as correct =1 or incorrect =0) to the items of a single GATB sub-test was computed for each of the groups using the LISREL VI computer program (Jöreskog & Sörbom, 1985). The hypothesis of equality of correlation matrices was then tested for each of the following group pairs: Norm 1 and 2; Chinese 1 and 2; Chinese 1 and Norm 1; Chinese 2 and Norm 1; Native On-Reserve and Norm 1; Native Off-Reserve and Norm 1; Native On-Reserve and Native Off-Reserve. The equality hypothesis was not rejected if the following three conditions were met for both the two groups compared:

1) Goodness-of-Fit Index (GFI) $\geq$ .90;
2) Root Mean Square Residual (RMSR) $\leq$ .120; and
3) (Residuals $\geq$ .200) $\leq$ 10%

In attempting to compute a tetrachoric correlation matrix for each group on each GATB sub-test, it was discovered that for a number of items near the end of each sub-test the variance of item responses was extremely low. Computation of the tetrachoric inter-correlations proved to be impossible if items having a percentage of correct responses less than ten per cent were included in the analysis. Consequently, the number of items used for each GATB sub-test in the comparative factor analysis portion of the study was somewhat less than the number cited in Table 7. In fact, the number of items used in this portion of the study, for each sub-test, was reduced to those shown in Table 9, below.

TABLE 9

Item Set Used in Comparative Factor Analysis

SUB-TEST	NUMBER OF ITEMS USED WITH OTHER MEASUREMENT TECHNIQUES	NUMBER OF ITEMS RETAINED IN ANALYSIS
Computation	41	28
Three Dimensional Space	40	30
Vocabulary	46	33
Arithmetic Reasoning	20	15

Tables 10 to 13 contain the results from the first-step of this procedure for each of the four GATB sub-tests and for the seven comparisons of group matrices outlined above.

TABLE 10

Tests of Equality of Correlation Matrices: GATB Sub-Test Computation

GROUP PAIR	N	GFI	RMSR	% RESIDUALS $\geq$ .200	EQUALITY CONDITIONS NOT MET (*)
Norm 1	366	.931	.107	7.9	
Norm 2	374	.937	.109	8.7	
Chinese 1	333	.951	.095	6.1	
Chinese 2	347	.946	.091	5.6	
Chinese 1	333	.923	.119	9.0	
Norm 1	366	.938	.108	7.4	
Chinese 2	347	.932	.113	7.9	
Norm 1	366	.906	.119	9.3	
Native On-Res.	386	.958	.089	4.2	
Norm 1	366	.935	.094	4.2	
Native Off-Res.	371	.937	.108	9.3	
Norm 1	366	.952	.107	8.2	
Native On-Res.	386	.950	.098	5.1	
Native Off-Res.	371	.957	.102	4.9	

TABLE 11

Test of Equality of Correlation Matrices:
GATB Sub-Test Three Dimensional Space

GROUP PAIR	N	GFI	RMSR	% RESIDUALS $\geq$ .200	EQUALITY CONDITIONS NOT MET (*)
Norm 1	366	.967	.077	.1	
Norm 2	374	.963	.026	.1	
Chinese 1	333	.964	.079	.1	
Chinese 2	347	.962	.076	.1	
Chinese 1	333	.999	.011	.0	
Norm 1	366	.999	.011	.0	
Chinese 2	347	.958	.080	.1	
Norm 1	366	.968	.076	.1	
Native On-Res.	386	.918	.095	4.6	
Norm 1	366	.943	.100	6.2	
Native Off-Res.	371	.956	.082	1.8	
Norm 1	366	.962	.083	2.9	
Native On-Res.	386	.934	.085	1.9	
Native Off-Res.	371	.948	.089	2.9	

TABLE 12

Tests of Equality of Correlation Matrices:
GATB Sub-Test Vocabulary

GROUP PAIR	N	GFI	RMSR	% RESIDUALS $\geq$ .200	EQUALITY CONDITIONS NOT MET (*)
Norm 1	366	.939	.092	5.0	
Norm 2	374	.938	.090	4.6	
Chinese 1	333	.926	.083	2.1	
Chinese 2	347	.936	.080	1.8	
Chinese 1	333	.871	.110	9.1	*
Norm 1	366	.928	.100	4.4	
Chinese 2	347	.892	.104	7.0	*
Norm 1	366	.929	.099	6.2	
Native On-Res.	386	.932	.088	4.1	
Norm 1	366	.937	.093	5.0	
Native Off-Res.	371	.948	.098	4.1	
Norm 1	366	.929	.099	4.4	
Native On-Res.	386	.931	.089	3.9	
Native Off-Res.	371	.953	.092	2.8	

TABLE 13

Tests of Equality of Correlation Matrices: GATB Sub-Test Arithmetic Reasoning

GROUP PAIR	N	GFI	RMSR	% RESIDUALS $\geq$ .200	EQUALITY CONDITIONS NOT MET (*)
Norm 1	366	.952	.090	7.6	
Norm 2	374	.955	.088	4.8	
Chinese 1	333	.942	.091	5.7	
Chinese 2	347	.937	.088	3.8	
Chinese 1	333	.929	.101	4.8	
Norm 1	366	.950	.092	2.9	
Chinese 2	347	.920	.101	8.6	
Norm 1	366	.946	.095	6.7	
Native On-Res.	386	.959	.087	6.7	
Norm 1	366	.951	.091	7.6	
Native Off-Res.	371	.953	.093	7.6	
Norm 1	366	.948	.094	7.6	
Native On-Res.	386	.972	.071	5.4	
Native Off-Res.	371	.970	074	5.2	

The only sub-test for which all groups compared were not judged to have equal correlation matrices was sub-test Vocabulary. For that sub-test both Chinese groups (1 and 2) were found to have correlation matrices different from the Norm 1 group. Consequently, the second step of the comparative factor analysis approach used in this study was followed for the Vocabulary sub-test. In this second step, the factor structure of the sub-test was first determined from the item response inter-correlation matrix of the Norm 1 group. Unrestricted exploratory factor analysis was used with a two-factor structure being obtained. Varimax rotation was applied to the initial factor structure with factors having eigenvalues greater than one being retained.

It should be recalled from the previous chapter that the simulataneous comparison of factor structures using LISREL VI involves the computation of deviations of each group's factor structure pattern from the Norm 1 group pattern. The Norm 1 group factor structure was generated through exploratory factor analysis, as previously explained. The structures of the Chinese 1 and Norm 1 groups, and the Chinese 2 and Norm 1 groups were then compared. With the LISREL VI program, this comparison involves the derivation of a "pooled" factor structure, then the simultaneous comparison of each group's factor structure to the "pooled" structure. The results obtained (in terms of the same criterion values as used in correlation matrix analysis) are shown in Table 14, below.

TABLE 14

Tests of Equality of Factor Patterns: GATB Sub-Test Vocabulary

GROUP PAIR	N	GFI	RMSR	% RESIDUALS $\geq$ .200	EQUALITY CONDITIONS NOT MET (*)
Chinese 1	333	.905	.094	4.2	
Norm 1	366	.928	.099	6.1	
Chinese 2	347	.909	.095	4.2	
Norm 1	366	.929	.099	5.9	

Although the hypothesis of equality of correlation matrices was earlier rejected for the two Chinese groups in relation to the Norm 1 group for the sub-test Vocabulary, the factor structures (patterns) of these two culture groups were not found to be different from the pattern of the Norm 1 group. Thus, for each culture group with Norm 1 group comparison using the Comparative Factor Analysis Technique a conclusion of equality was reached.

Modified Caution Index (C*) Technique

The Student Problem Package (SPP) (Harnisch, Kuo, & Torres, 1983) was used to compute a Modified Caution Index (C*) for each subject in the total group (N=2177) on each GATB sub-test. Once C* was computed for each subject in the sample, the means and standard deviations of C* values for each group were computed. Tables 15(a) to 15(d) display by sub-test the means and standard deviations for the group indices.

TABLE 15

Group Means and Standard Deviations for C* by GATB Sub-Test

(a) Test: Computation

GROUP	N	C* MEAN	C* STD DEV
Chinese 1	333	.047	.069
Chinese 2	347	.073	.048
Native (On-Res)	386	.079	.065
Native (Off-Res)	371	.048	.047
Norm 1	366	.033	.051
Norm 2	374	.035	.041
TOTAL	2,177	.053	.052

(b) Test: Three Dimensional Space

GROUP	N	C* MEAN	C* STD DEV
Chinese 1	333	.117	.104
Chinese 2	347	.130	.049
Native (On-Res)	386	.090	.083
Native (Off-Res)	371	.093	.091
Norm 1	366	.084	.086
Norm 2	374	.093	.085
TOTAL	2,177	.101	.071

TABLE 15 (cont'd)

(c) Test: Vocabulary

GROUP	N	C* MEAN	STD DEV
Chinese 1	333	.174	.111
Chinese 2	347	.195	.049
Native (On-Res)	386	.125	.064
Native (Off-Res)	371	.197	.103
Norm 1	366	.085	.069
Norm 2	374	.089	.072
TOTAL	2,177	.143	.066

(d) Test: Arithmetic Reasoning

GROUP	N	C* MEAN	STD DEV
Chinese 1	333	.110	.050
Chinese 2	347	.085	.082
Native (On-Res)	386	.155	.091
Native (Off-Res)	371	.358	158
Norm 1	366	.079	.092
Norm 2	374	.081	.116
TOTAL	2,177	.146	.079

To simultaneously compare the mean C* values, for each sub-test, of all six groups in the study (Norm 1, Norm 2, Chinese 1, Chinese 2, Native On-Reserve, Native Off-Reserve), a one-way analysis of variance (ANOVA) was conducted. Tables 16 to 19 contain the ANOVA summary tables and tables of post-hoc tests (Scheffe's) for the analyses of C* by group, taken over each of the four GATB sub-tests.

It should be noted for most of the GATB sub-test comparisons with C* that the homogeneity of variance assumption of ANOVA was not met. However, it has been shown that ANOVA is robust to violations of this assumption (Lindquist, 1953) when the group sample sizes are large and approximately equal, as they were in this study. The C* index is, however, known to be approximately normally distributed in the population.

TABLE 16

ANOVA Summary Table for C* by Group - Sub-Test Computation

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	.7054	5	.1411	.9407	.4534
Error	325.6123	2171			
Total	326.3177	2176			

At the .05 significance level, there was not a significant group effect. Post-hoc procedures were, therefore, not needed.

TABLE 17 (a)

ANOVA Summary Table for C* by Group - Sub-Test Three Dimensional Space

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	.5746	5	.1149	2.56	.0259
Error	97.6340	2171	.0450		
Total	98.2086	2176			

Because the group effect was significant at the .05 level, post-hoc procedures using Scheffe's statistic were conducted. The pairs of groups which were significantly different at the .05 level are shown in Table 17(b).

TABLE 17(b)

Results of Scheffe Test for Pair Wise Comparison of Groups - Sub-Test Three Dimensional Space

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (On-Res)						
NAT (Off-Res)						
NORM 1		*				
NORM 2						

* indicates significant comparison at .05 level

TABLE 18(a)

ANOVA Summary Table for C* by Group - Sub-Test Vocabulary

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	4.7682	5	.9536	3.22	.0067
Error	643.1138	2171	.2926		
Total	647.8820	2176			

Because the group effect was significant at the .05 level, post-hoc procedures using Scheffe's statistic was conducted. The pairs of groups found to be significantly different from each other are shown in Table 18(b).

TABLE 18(b)

Results of Scheffe Test for Pair Wise Comparison of Groups - Sub-Test Vocabulary

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (On-Res)						
NAT (Off-Res)						
NORM 1						
NORM 2	(None found different)					

* indicates significant comparison at .05 level

TABLE 19(a)

ANOVA Summary Table for C* by Group - Sub-Test Arithmetic Reasoning

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	21.5766	5	4.3153	7.0000	.0000
Error	1338.9842	2171	.6168		
Total	1360.5608	2176			

Because the group effect was significant at the .05 level, post-hoc procedures were needed to identify where the differences lie. Scheffe's procedure revealed the differences shown in Table 19(b).

TABLE 19(b)

Results of Scheffe Test for Pair Wise Comparison of Groups - Sub-Test Arithmetic Reasoning

	GROUP					
	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
GROUP CHIN 1						
CHIN 2				*		
NAT (On-Res)						
NAT (Off-Res)						
NORM 1				*		
NORM 2				*		

* indicates significant comparison at .05 level.

From the results displayed in Tables 16 to 19, it can be seen that significant differences in mean C* values do not exist for either of the two control group comparisons, that is, Chinese 1 with Chinese 2 or Norm 1 with Norm 2 group. Thus, the strategy seems to be functioning correctly, according to Hypothesis One. With respect to the second hypothesis concerning the usefulness of the strategy in establishing measurement equivalence or non-equivalence of cultural groups in relation to a norm group, a group by decision table can be derived for each GATB sub-test from the contents of Tables 16 to 19. That table (Table 20) follows.

TABLE 20

Culture Group By Decision (Equivalence/ Non-Equivalence) Table for C*

GATB SUB-TEST

GROUP COMPARISON	COMPUTATION	THREE DIM SPACE	VOCAB	ARITHMETIC REASONING
Chinese 1 vs Chinese 2	1	1	1	1
Norm 1 vs Norm 2	1	1	1	1
Chinese 1 vs Norm 1	1	1	1	1
Chinese 2 vs Norm 1	1	0	1	1
Native (On-Res) vs Norm 1	1	1	1	1
Native (Off-Res) vs Norm 1	1	1	1	0
Native (On-Res) vs Native (Off-Res)	1	1	1	1

Decision Categories: Non-Equivalence = 0
Equivalence = 1

In the previous chapter, it is mentioned that indices such as the C*, which measures the departure of an individual's response pattern across items from the pattern exhibited by a group, may be more closely related than desirable to individual total test score. Total test score is thought to be a reflection to some degree of measured ability. What is sought in using an index like C* is an indication of individual aberrance, or difference, in test performance from that of a group which is not simply a reflection of true ability differences. It is also mentioned in the discussions of the previous chapter that standardized IRT-based response pattern indices may be less prone than those based on non-IRT approaches to reflect ability differences rather than differences due, for example, to cultural factors. Thus, some analysis of the relationship between individuals' total test scores and their response pattern indices (C* with the non-IRT approach; L_z and ECIZ4 with the IRT approach) should be helpful.

Indices like C*, L_z and ECIZ4 are used to flag examinees' response patterns that are atypical. Usually, such response patterns are found to exist for examinees with spuriously high or low scores. Thus, the relationship, if any, of total test score to an index such as C* might be expected to be curvilinear. To see if this was the case, the scatterplot of examinee C* values by examinee raw sub-test scores was first examined.

Based on the scatterplots for the total group of examinees (N=2177), the relationship of C* to total test score appeared to be, for all four sub-tests, curvilinear (quadratic). Consequently, for each sub-test, the quadratic relationship of C* to total score was examined on a group by group basis. To do this, each examinee's total sub-test score was squared and then correlated (Pearson Product-Moment) with the C* indices. The correlations obtained are given in Table 21 below.

TABLE 21

Correlations of Squared Total Sub-Test Scores with C* - Presented by Group

SUB-TEST	GROUP	CORR. COEFFICIENT
Computation	Total Group	.0339
	Chinese 1	.0697
	Chinese 2	-.1115
	Native (On-Res)	.0278
	Native (Off-Res)	.1225
	Norm 1	.0754
	Norm 2	.1321
Three Dimensional Space	Total Group	.1741
	Chinese 1	.0929
	Chinese 2	.1218
	Native (On-Res)	.0153
	Native (Off-Res)	.0718
	Norm 1	.3113
	Norm 2	.1932

TABLE 21 (cont'd)

Vocabulary	Total Group	.1562
	Chinese 1	-.1296
	Chinese 2	-.1397
	Native (On-Res)	.0052
	Native (Off-Res)	.0292
	Norm 1	.1548
	Norm 2	.2577
Arithmetic Reasoning	Total Group	.0486
	Chinese 1	.0728
	Chinese 2	.1178
	Native (On-Res)	.0068
	Native (Off-Res)	-.0847
	Norm 1	.1043
	Norm 2	.4058

In Table 21 it is seen that most of the correlations are relatively small. With respect to the culture groups for which a significant difference in C* value in relation to the Norm 1 group was found (see Table 20), the relationship of C* to sub-test score is no different than it is for the other groups. In fact, the strongest relationship of C* to test score is found for the Norm 2 group on sub-test Arithmetic Reasoning and Vocabulary, and for the Norm 1 group on the Three Dimensional Space sub-test.

Standardized Appropriateness Index (L_z) and Standardized Extended Caution Index (ECIZ4) Techniques

The application of these two techniques (both based on IRT procedures) to a determination of the equivalence or non-equivalence across each GATB sub-test for the control group comparisons and for the comparison of culture groups proceeded in three steps, as mentioned earlier. In the first step, item parameters (difficulty, discrimination, and guessing) were estimated for each sub-test using data (matrix of incorrect [0] and correct [1] responses)

from a reference group (N=1135). The item parameters, for each sub-test, estimated by the LOGIST computer program, are shown in Table A1 to Table A4 of the Appendix.

In the second step, a computer program written by Drasgow (1987) was used to estimate individual abilities from the LOGIST-established item parameters, and to compute the Standardized Appropriateness Index (L_z) and the Standardized Extended Caution Index (ECIZ4) for each subject (N=2177). The list of ability estimates and L_z and ECIZ4 values for subjects is too lengthy to publish, but is available from the author.

The third step in analysis consisted of a comparison using a one-way ANOVA of L_z and ECIZ4 group means. A separate analysis was conducted for each GATB sub-test and for each index (L_z and ECIZ4). An absence of significant differences across the two Chinese and the two norm groups was taken to indicate that the method was functioning correctly (Hypothesis One). Significant differences between groups other than the control group pairs reflected the usefulness of the method in determining equivalence or non-equivalence of test performance due to cultural factors (Hypothesis Two). Table 22 contains the means and standard deviations for the two indices on the four GATB sub-tests. Results of ANOVA for the L_z index are given in Tables 23 to 26. Results for the ECIZ4 index are given in Tables 28 to 31.

Similar to the case of the C* index, the homogeneity of variance assumption of ANOVA was not met for either L_z or ECIZ4 analyses. However, as stated earlier, ANOVA is robust to violations of this assumption when the sample sizes are approximately equal. Both L_z and ECIZ4 are considered to be normally distributed in the population.

TABLE 22

Group Means and Standard Deviations for L_z and ECIZ4 by GATB Sub-Test

(a) Test: Computation

GROUP	N	L_z MEAN	L_z STD DEV	ECIZ4 MEAN	ECIZ4 STD DEV
Chinese 1	333	-1.1560	3.4181	.4895	2.2983
Chinese 2	347	-1.1288	2.9700	.5367	2.1302
Native (On-Res)	386	.0986	1.6582	-.2467	1.3617
Native (Off-Res)	371	.4708	1.5197	-.5177	1.2976
Norm 1	366	.1076	1.8838	-.2588	1.4267
Norm 2	374	- .0227	1.8818	-.1836	1.3949
TOTAL	2177	- .2450	2.3780	-.0470	1.7240

(b) Test: Three-Dimensional Space

GROUP	N	L_z MEAN	L_z STD DEV	ECIZ4 MEAN	ECIZ4 STD DEV
Chinese 1	333	-1.0721	2.5351	.7184	2.0960
Chinese 2	347	- .6671	2.3939	.3542	1.9188
Native (On-Res)	386	- .0475	1.7448	-.0505	1.5111
Native (Off-Res)	371	.0484	1.8385	-.1206	1.6051
Norm 1	366	.1350	1.6551	-.2699	1.4014
Norm 2	374	- .1202	1.7907	.0058	1.5477
TOTAL	2177	- .2680	2.0480	.0930	1.7160

(c) Test: Vocabulary

GROUP	N	L_z MEAN	L_z STD DEV	ECIZ4 MEAN	ECIZ4 STD DEV
Chinese 1	333	-.9342	1.6937	.5903	1.2547
Chinese 2	347	-.9823	1.7858	.6626	1.1890
Native (On-Res)	386	.2003	1.2296	-.0793	1.1604
Native (Off-Res)	371	.4427	1.2247	-.2054	1.1487
Norm 1	366	-.0533	1.6203	.0277	1.3597
Norm 2	374	-.0935	1.5127	.0721	1.3101
TOTAL	2177	-.2140	1.6080	.1640	1.2800

TABLE 22 (cont'd)

(d) Test: Arithmetic Reasoning

GROUP	N	L_z MEAN	L_z STD DEV	ECIZ4 MEAN	ECIZ4 STD DEV
Chinese 1	333	-.3931	1.5558	.2256	1.2554
Chinese 2	347	-.4512	1.5317	.2680	1.2392
Native (On-Res)	386	.3394	1.0898	-.3314	.9909
Native (Off-Res)	371	.4159	1.1244	-.3441	.9815
Norm 1	366	-.0091	1.2554	-.1314	1.0726
Norm 2	374	-.0316	1.3920	-.1612	1.1254
TOTAL	2177	.0030	1.3680	-.0900	1.1360

TABLE 23(a)

ANOVA Summary Table for L_z by Group - Sub-Test Computation

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	849.4342	5	169.8868	32.1830	.0000
Error	11460.2062	2171	5.2788		
Total	12309.6403	2176			

Post-hoc procedures are required because of the significance (at the .05 level) of the computed F-value. Scheffe's statistic revealed the differences shown in Table 23(b) below.

TABLE 23(b)

Results of Scheffe Test for Pair Wise Comparison of Groups - Sub-Test Computation

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (On-Res)	*	*				
NAT (Off-Res)	*	*				
NORM 1	*	*				
NORM 2	*	*				

* indicates significant comparison at .05 level

TABLE 24(a)

ANOVA Summary Table for L_z by Group - Sub-Test Three Dimensional Space

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	394.1219	5	78.8244	19.5907	.000
Error	8735.1349	2171	4.0236		
Total	9129.2568				

Because the computed F was significant (at the .05 level), post-hoc procedures using Scheffe's statistic were needed. The results of post-hoc analysis are displayed in Table 24(b).

TABLE 24(b)

Results of Scheffe Test for Pair Wise Comparison of Group Means - Sub-Test Three Dimensional Space

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (RES)	*					
NAT (N-RES)	*	*				
NORM 1	*	*				
NORM 2	*					

* indicates significant comparison at .05 level

TABLE 25(a)

ANOVA Summary Table for L_z by Group - Sub-Test Vocabulary

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	618.6272	5	123.7254	53.6721	.000
Error	5004.6094	2171	2.3052		
Total	5623.2366	2176			

Since the group effect was significant (at the .05 level) for Sub-Test Vocabulary, pair wise comparisons of group means following Scheffe's procedure were carried out. The results of the comparisons are shown in Table 25(b).

TABLE 25(b)

Results of Scheffe Test for Pair Wise Comparison of Groups Sub-Test Vocabulary

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (ON-RES)	*	*				
NAT (OFF-RES)	*	*				
NORM 1	*	*				*
NORM 2	*	*				*

* indicates significant comparison at .05 level

TABLE 26(a)

ANOVA Summary Table for L_z by Group - Sub-Test Arithmetic Reasoning

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	231.1457	5	46.2291	26.1467	.000
Error	3838.4719	2171	1.7681		
Total	4069.6176	2176			

Since the group effect was significant (at the .05 level) for Sub-Test Arithmetic Reasoning, pair wise comparisons of group means following Scheffe's procedure were carried out. The results of the comparisons are given in Table 26(b).

TABLE 26(b)

Results of Scheffe Test for Pair Wise Comparison of Groups Sub-Test Arithmetic Reasoning

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (ON-RES)	*	*				
NAT (OFF-RES)	*	*				
NORM 1	*	*		*		*
NORM 2		*				*

* indicates significant comparison at .05 level.

For all four sub-tests, there were no significant differences in L_z value between the pairs of control groups. Accepting that the method was functioning correctly (Hypothesis One), the second hypothesis was studied by constructing a culture group by decision table for each GATB sub-test, in which the results contained in Tables 23 to 26 are transformed into statements of measurement equivalence or non-equivalence for each sub-test across the four culture groups in relation to the Norm group, across the two pairs of control groups, and across the two native groups. Table 27 displays these results in terms of decision categories (measurement equivalence or non-equivalence).

TABLE 27

Culture Group by Decision (Equivalence/Non-Equivalence) Table for L_z

	GATB SUB-TEST			
GROUP COMPARISON	COMPUTATION	THREE DIM SPACE	VOCAB	ARITHMETIC REASONING
Chinese 1 vs Chinese 2	1	1	1	1
Norm 1 vs Norm 2	1	1	1	1
Chinese 1 vs Norm 1	0	0	0	0
Chinese 2 vs Norm 1	0	0	0	0
Native (On-Res) vs Norm 1	1	1	1	0
Native (Off-Res) vs Norm 1	1	1	0	0
Native (On-Res) vs Native (Off-Res)	1	1	1	1

Decision Categories: Non-Equivalence = 0
Equivalence = 1

Tables 28 to 31 display the ANOVA summary tables from analysis with the ECIZ4 index, for the four GATB sub-tests.

TABLE 28(a)

ANOVA Summary Table for ECIZ4 by Group - Sub-Test Computation

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	335.0355	5	67.0071	23.7337	.0000
Error	6129.3506	2171	2.8233		
Total	6464.3862	2176			

Post-hoc procedures were needed to determine between which group pairs there were significant differences in mean ECIZ4 values. Scheffe's statistic revealed the differences shown in Table 28(b).

TABLE 28(b)

Results of Scheffe Test for Pair Wise Comparison of Group Means - Sub-Test Computation

GROUP

	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
GROUP CHIN 1						
CHIN 2						
NAT (ON-RES)	*	*				
NAT (OFF-RES)	*	*				
NORM 1	*	*				
NORM 2	*	*				

* indicates significance at the .05 level

TABLE 29(a)

ANOVA Summary Table for ECIZ4 by Group - Sub-Test Three Dimensional Space

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	229.5909	5	45.9182	16.1434	.0000
Error	6175.1960	2171	2.8444		
Total	6404.7869	2176			

Because the computed F in Table 29(a) was significant, post-hoc procedures (Scheffe's test) were needed to determine which group pairs are significantly different in mean ECIZ4 value. Table 29(b) displays the results from post-hoc analysis.

TABLE 29(b)

Results of Scheffe Test for Pair Wise Comparison of Group Means - Sub-Test Three Dimensional Space

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (ON-RES)	*					
NAT (OFF-RES)	*					
NORM 1	*		*			
NORM 2	*					

* indicates significance at the .05 level

TABLE 30(a)

ANOVA Summary Table for ECIZ4 by Group - Sub-Test Vocabulary

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	230.2001	5	46.0400	29.9839	.0000
Error	3333.5477	2171	1.5355		
Total	3563.7478	2176			

To determine between which group pairs the mean ECIZ4 value was significantly different, Scheffe's statistic was used, as shown in Table 30(b).

TABLE 30(b)

Results of Scheffe Test for Pair Wise Comparison of Group Means - Sub-Test Vocabulary

GROUP

GROUP	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
CHIN 1						
CHIN 2						
NAT (ON-RES)	*		*			
NAT (OFF-RES)	*		*			
NORM 1	*		*			
NORM 2	*		*			

* indicates significance at the .05 level

TABLE 31(a)

ANOVA Summary Table for ECIZ4 by Group - Sub-Test Arithmetic Reasoning

SOURCE OF VARIATION	SS	DF	MS	F	SIG OF F
Group Effect	126.6105	5	25.3221	20.5026	.0000
Error	2681.3311	2171	1.2351		
Total	2807.9416	2176			

Post-hoc procedures (Scheffe's test) were applied to determine between which group pairs the mean ECIZ4 was significantly different, with results as given in Table 31(b).

TABLE 31(b)

Results of Scheffe Test for Pair Wise Comparison of Group Means - Sub-Test Arithmetic Reasoning

	GROUP					
	CHIN 1	CHIN 2	NAT(ON-RES)	NAT(OFF-RES)	NORM 1	NORM 2
GROUP CHIN 1						
CHIN 2						
NAT (ON-RES)	*	*				
NAT (OFF-RES)	*	*				
NORM 1	*	*				
NORM 2	*	*				

* indicates significance at the .05 level.

For all four sub-tests, there were no significant differences in ECIZ4 value between the pairs of control groups. It was concluded that the technique was functioning correctly (Hypothesis one). In relation to the second hypothesis, a group by decision (equivalence or non-equivalence) table was derived for the four GATB sub-tests from the contents of Tables 28 to 31. The group by decision table (Table 32) is given below.

TABLE 32

Culture Group by Decision (Equivalence/ Non-Equivalence) Table for ECIZ4

	GATB SUB-TEST			
GROUP COMPARISON	COMPUTATION	THREE DIM SPACE	VOCAB	ARITHMETIC REASONING
Chinese 1 vs Chinese 2	1	1	1	1
Norm 1 vs Norm 2	1	1	1	1
Chinese 1 vs Norm 1	0	0	0	0
Chinese 2 vs Norm 1	0	0	0	0
Native (On-Res) vs Norm 1	1	1	1	1
Native (Off-Res) vs Norm 1	1	1	0	1
Native (On-Res) vs Native (Off-Res)	1	1	1	1

Decision Categories: Non-Equivalence = 0
Equivalence = 1

It was mentioned earlier that one difficulty in applying response pattern analysis to a determination of equivalence or non-equivalence in test performance is that the indices used are often closely related to examinee ability, as measured by total test score. Thus, the relationship of total test score with both the L_z and ECIZ4 indices, was examined. As well, the relationship between the two IRT-based indices (L_z and ECIZ4), both of which represent the same measurement strategy (see Table 5, Chapter Three), was looked at below.

As was discussed in relation to the C* index, the L_z and ECIZ4 indices are used to flag examinee response patterns that are thought to be atypical. Such patterns are usually associated with spuriously high or low test scores. Thus, it might be expected that any relationship of L_z or ECIZ4 with total test score would be curvilinear. To determine if this was the case, the scatterplot of L_z by test score, and of ECIZ4 by test score, for the total group of examinees (N=2177) can first be inspected. This was done for each of the four sub-tests.

Having examined the scatterplots of L_z by test score and ECIZ4 by test score, it was decided that the relationship of each index to test score was most likely curvilinear (quadratic). Consequently, the strength of the quadratic relationship was investigated for all groups taken together and separately for each of the six groups in the study. To do this, examinees' sub-test scores were squared and correlated (Pearson Product-Moment) with both the L_z and ECIZ4 indices. The correlation coefficients obtained are presented in Table 33 below.

TABLE 33

Correlations of Squared Total Sub-Test Scores with L_z and ECIZ4 - Presented by Group

SUB-TEST	GROUP	CORR. COEFFICIENT (With L_z)	CORR. COEFFICENT (With ECIZ4)
Computation	Total Group	-.2113	.2129
	Chinese 1	-.2299	.2361
	Chinese 2	-.0917	.1111
	Native (On-Res)	-.1159	.1669
	Native (Off-Res)	-.2746	.2766
	Norm 1	-.1646	.1834
	Norm 2	-.2679	.2360
Three Dimensional Space	Total Group	-.1432	.1559
	Chinese 1	-.0877	.0550
	Chinese 2	-.0781	.1205
	Native (On-Res)	.0142	.0300
	Native (Off-Res)	-.1307	.2174
	Norm 1	-.0785	.1361
	Norm 2	-.0914	.0879
Vocabulary	Total Group	-.0190	.0942
	Chinese 1	-.0154	.0911
	Chinese 2	-.0397	.1576
	Native (On-Res)	-.0336	.1382
	Native (Off-Res)	-.0471	.1342
	Norm 1	-.1106	.1010
	Norm 2	-.2280	.1870

TABLE 33 (cont'd)

Arithmetic Reasoning	Total Group	.0115	.0657
	Chinese 1	-.0195	-.0029
	Chinese 2	-.0948	.0203
	Native (On-Res)	-.0850	.1778
	Native (Off-Res)	-.0490	.3608
	Norm 1	-.0230	-.0708
	Norm 2	-.2394	-.1869

It is seen in Table 33 that all of the correlations are relatively small. Moreover, no differences are evident in the magnitudes of correlations of either L_z or ECIZ4 with total score for the culture groups found to be significantly different from the Norm 1 group on these index values (see Table 32). Note also that in most cases the correlations for L_z are negative while those for ECIZ4 are positive. This is due to the fact that low values of L_z indicate aberrance of response patterns while high values of ECIZ4 indicate aberrance.

It was also considered to be of interest to correlate the three indices used in response pattern analysis (C*, L_z and ECIZ4). The C* index was used in this study to represent one measurement strategy (Response Pattern Analysis - Non-IRT Approach). This strategy can be used to challenge the item equivalence assumption only. The L_z and ECIZ4 indices were taken to represent a different strategy (Response Pattern Analysis - IRT Approach), which is used to challenge both item and scalar equivalence. Thus, it would be anticipated that L_z and ECIZ4 would be more highly correlated with each other than with C*. The reason for examining the correlations is to determine the extent to which the two strategies, as represented by the C* on the one hand and the L_z and ECIZ4 indices on the other, are measuring the same thing. The inter-correlations of these indices (Person Product-Moment) are given in Table 34.

TABLE 34

Correlation of C*, L_z, and ECIZ4
(N=2177)

SUB-TEST	VARIABLES CORRELATED	COEFFICIENT
Computation	C* - L_z	-.9006
	C* - ECIZ4	.8424
	L_z - ECIZ4	-.9240
Three Dimensional Space	C* - L_z	-.9029
	C* - ECIZ4	.8473
	L_z - ECIZ4	-.9430
Vocabulary	C* - L_z	-.8150
	C* - ECIZ4	.6577
	L_z - ECIZ4	-.8792
Arithmetic Reasoning	C* - L_z	-.8143
	C* - ECIZ4	.7222
	L_z - ECIZ4	-.8925

The surprising finding from the correlations presented in Table 34 was that C* is more highly correlated with L_z than with ECIZ4 in all cases. This was not expected as C* and ECIZ4 are more similar indices in terms of the approach used in calculation than are C* and L_z.

Distractor Analysis Technique

As described earlier in Chapter Three, the program CULVAL was used to compare the distractor (foil) choice pattern of seven pairs of groups. (Chinese 1 and 2; Norm 1 and 2; Chinese 1 and Norm 1; Chinese 2 and Norm 1; Native (On-Res) and Norm 1; Native (Off-Res) and Norm 1; Native (On-Res) and Native (Off-Res)). The CULVAL program provided an item by item comparison of the group pairs; however, it also gave total test statistics which were used to compare the group pairs with respect to their distractor choice patterns. The total test statistics used were the chi-square value (χ^2), as well as an index of the homogeneity or association of distractor (foil) choices across groups (Cramer's V). The total test χ^2 was in fact a sum of the item by

item χ^2 values, and the Generalized Cramer's V was computed as a weighted average of the V indices, across all items on a test. The computation of these values was explained in the previous chapter. The results of this analysis for the four GATB sub-tests are given in Tables 35(a) to 35(d). For all tables, the decision rule to reject a null hypothesis of no difference between groups is $\chi^2 \leq .05$ and $V \geq 0.15$.

TABLE 35(a)

CULVAL Output for Computation Sub-Test

GROUP PAIR	N	TOTAL TEST χ^2	DF	PROB	WEIGHTED AVER CRAMER V	SIGNIFICANCE (*)
Chinese 1 Chinese 2	333 347	114.82	123	0.6911	0.113	
Norm 1 Norm 2	366 374	175.70	123	0.0010	0.115	
Chinese 1 Norm 1	333 366	280.44	123	0.0000	0.160	*
Chinese 2 Norm 1	347 366	271.90	123	0.0000	0.155	*
Native (On-Res) Norm 1	386 366	145.98	123	0.0757	0.103	
Native (Off-Res) Norm 1	371 366	389.18	123	0.0000	0.163	*
Native (On-Res) Native (Off-Res)	386 371	166.30	123	0.0049	0.113	

TABLE 35(b)

CULVAL Output for Three Dimensional Space Sub-Test

GROUP PAIR	N	TOTAL TEST χ^2	DF	PROB	WEIGHTED AVER CRAMER V	SIGNIFICANCE (*)
Chinese 1 Chinese 2	333 347	89.06	80	0.2286	0.094	
Norm 1 Norm 2	366 374	128.07	80	0.0005	0.098	
Chinese 1 Norm 1	333 366	168.92	80	0.0000	0.161	*
Chinese 2 Norm 1	347 366	162.08	80	0.0000	0.117	
Native (On-Res) Norm 1	386 366	96.83	80	0.0969	0.084	
Native (Off-Res) Norm 1	371 366	117.91	80	0.0038	0.091	
Native (On-Res) Native (Off-Res)	386 371	141.73	80	0.0000	0.158	*

TABLE 35(c)

CULVAL Output for Vocabulary Sub-Test

GROUP PAIR	N	TOTAL TEST χ^2	DF	PROB	WEIGHTED AVER CRAMER V	SIGNIFICANCE (*)
Chinese 1 Chinese 2	333 347	142.95	138	0.3725	0.088	
Norm 1 Norm 2	366 374	228.80	138	0.0000	0.127	
Chinese 1 Norm 1	333 366	586.66	138	0.0000	0.191	*
Chinese 2 Norm 1	347 366	611.48	138	0.0000	0.193	*
Native (On-Res) Norm 1	386 366	220.26	138	0.0000	0.118	
Native (Off-Res) Norm 1	371 366	408.70	138	0.0000	0.161	*
Native (On-Res) Native (Off-Res)	386 371	253.02	138	0.0000	0.120	

TABLE 35(d)

CULVAL Output for Arithmetic Reasoning Sub-Test

GROUP PAIR	N	TOTAL TEST χ^2	DF	PROB	WEIGHTED AVER CRAMER V	SIGNIFICANCE (*)
Chinese 1 Chinese 2	333 347	56.41	60	0.6079	0.099	
Norm 1 Norm 2	366 374	85.24	60	0.0178	0.113	
Chinese 1 Norm 1	333 366	168.08	60	0.0000	0.166	*
Chinese 2 Norm 1	347 366	200.67	60	0.0000	0.179	*
Native (On-Res) Norm 1	386 366	148.24	60	0.0000	0.143	
Native (Off-Res) Norm 1	371 366	289.67	60	0.0000	0.196	*
Native (On-Res) Native (Off-Res)	386 371	99.27	60	0.0011	0.107	

To determine if the measurement strategy of distractor (foil) analysis was functioning correctly (Hypothesis One) in the identification of groups where test performance differs from a norm group for what was considered to be cultural reasons, the comparisons of distractor choice patterns for control groups (the two groups of Chinese and two groups of norms) were examined. From the data displayed in Tables 35(a) to 35(d), it is evident that this strategy was functioning as expected, since the control group distractor patterns were not found to be significantly different according to the decision criteria established.

With respect to the second study hypothesis, that a technique should detect some differences among groups, the data of Tables 35(a) to 35(d) would suggest that differences among some pairs of groups did exist.

From the contents of Tables 35(a) to 35(d), a culture group by decision (equivalence or non-equivalence) table (Table 36) can be constructed which reflects, for the seven group pairs compared, the equivalence or non-equivalence of the group pairs in distractor choice patterns over the four GATB sub-tests. Table 36 is presented below.

TABLE 36

Culture Group by Decision (Equivalence/Non-Equivalence) Table for Distractor Analysis

	GATB SUB-TEST			
GROUP COMPARISON	COMPUTATION	THREE DIM SPACE	VOCAB	REASONING
Chinese 1 vs Chinese 2	1	1	1	1
Norm 1 vs Norm 2	1	1	1	1
Chinese 1 vs Norm 1	0	0	0	0
Chinese 2 vs Norm 1	0	1	0	0
Native (On-Res) vs Norm 1	1	1	1	1
Native (Off-Res) vs Norm 1	0	1	0	0
Native (On-Res) vs Native (Off-Res)	1	0	1	1

Decision Categories: Non-Equivalence = 0
Equivalence = 1

Correct/Incorrect Response Analysis (Δ MH) Technique

The Mantel-Haenszel (MH) procedure was applied to examinees' vectors of correct (=1) and incorrect (=0) responses to the items of the four GATB sub-tests. Each of the four sub-tests was analyzed separately, with values of ΔMH computed for each item and for each of the following group pairs: Norm 1 and 2, Chinese 1 and 2, Chinese 1 and Norm 1, Chinese 2 and Norm 1, Native On-Res and Norm 1, Native Off-Res and Norm 1, Native On-Res and Native Off-Res. A computer program obtained from the American College Testing Program (ACT) was used to compute the MH values (Welch, Ackerman, Doolittle, & Hurley, 1987).

The ΔMH value for each item was used to determine if the item was equivalent (no Differential Item Performance or DIP) across the group pair. Thus, for each GATB sub-test, groups were compared in pairs, as described above, with one group labeled as the reference (R) group and the other as the focal (F) group. An item was considered as non-equivalent (DIP present) if the computed ΔMH value fell outside the range: $-1.50 \leq \Delta MH \leq 1.50$. For each GATB sub-test and each pair of groups compared, the percentage of items having negative DIP ($\Delta MH \leq -1.50$) and the percentage having positive DIP ($\Delta MH \geq +1.50$) were tabulated. It should be recalled that negative DIP represents item bias in favour of the reference group, while positive DIP represents bias favouring the focal group. If the absolute difference in this percentage of items containing positive as opposed to negative DIP equalled or exceeded 10%, the sub-test was judged to be non-equivalent for the group pair.

The list of computed ΔMH values for each item of each sub-test, and for each of the seven pairs of groups compared, is too lengthy to publish in

this study. However, Tables 37(a) to 37(d) contain the mean Δ MH value over all the items of each sub-test, for each pair of groups compared. In these tables, the first group listed for each comparison is the focal (F) group and the second group named is the reference (R) group.

It should be noted that the computed mean Δ MH values are large relative to those seen with other data sets. As mentioned in Chapter Three, the expected value for Δ MH, where DIP is not present, is zero. The rather large mean values (even for group comparisons where DIP is not found) are probably a result of the speeded nature of the GATB sub-tests. Many subjects did not complete items near the end of a sub-test, having run out of time. In some instances, focal group subjects possibly responded to the time limit by randomly filling-in answers to items near the end of a sub-test. In such cases, the focal group subjects (when matched by ability level to reference group subjects) obtained a higher proportion of items correct near the end of a sub-test, resulting in large positive Δ MH mean values (e.g., sub-test computation where the Chinese groups are compared to the Norm 1 group).

TABLE 37(a)

Mean Δ MH Value for 41 Items of Sub-Test Computation

GROUP PAIR	MEAN Δ MH
Norm 1 and 2	0.299
Chinese 1 and 2	0.001
Chinese 1 and Norm 1	0.375
Chinese 2 and Norm 1	0.375
Native On-Res and Norm 1	0.208
Native Off-Res and Norm 1	0.308
Native On-Res and Native Off-Res	0.321

TABLE 37(b)

Mean Δ MH Value for 40 Items of Sub-Test Three Dimensional Space

GROUP PAIR	MEAN ΔMH
Norm 1 and 2	0.152
Chinese 1 and 2	-0.114
Chinese 1 and Norm 1	0.259
Chinese 2 and Norm 1	0.108
Native On-Res and Norm 1	0.131
Native Off-Res and Norm 1	0.493
Native On-Res and Native Off-Res	0.202

TABLE 37(c)

Mean ΔMH Value for 46 Items of Sub-Test Vocabulary

GROUP PAIR	MEAN ΔMH
Norm 1 and 2	0.088
Chinese 1 and 2	0.075
Chinese 1 and Norm 1	0.295
Chinese 2 and Norm 1	0.279
Native On-Res and Norm 1	-0.196
Native Off-Res and Norm 1	-0.395
Native On-Res and Native Off-Res	0.114

TABLE 37(d)

Mean ΔMH Value for 20 Items of Sub-Test Arithmetic Reasoning

GROUP PAIR	MEAN ΔMH
Norm 1 and 2	0.039
Chinese 1 and 2	-0.011
Chinese 1 and Norm 1	-0.219
Chinese 2 and Norm 1	-0.174
Native On-Res and Norm 1	-0.379
Native Off-Res and Norm 1	-0.133
Native On-Res and Native Off-Res	0.101

Tables 38(a) to 38(d) give the percentage of items found to contain positive DIP, the percentage showing negative DIP, and the absolute difference in the two percentages, for each pair of groups compared on each GATB sub-test.

TABLE 38(a)

Percentage of Items with DIP on Sub-Test Computation

GROUP PAIR	% ITEMS Δ MH $\geq$ +1.50	% ITEMS Δ MH $\leq$ -1.50	% DIFFERENCE + and -	SIGNIFICANCE (*)	GROUP BIAS AGAINST
Norm 1 and 2	12.2	7.3	4.9		
Chinese 1 and 2	9.8	14.6	4.8		
Chinese 1 and Norm 1	36.6	7.3	29.3	*	Refer. (Norm 1)
Chinese 2 and Norm 1	31.7	9.8	21.0	*	Refer. (Norm 1)
Native On-Res and Norm 1	12.2	7.3	4.9		
Native Off-Res and Norm 1	14.6	14.6	0.0		
Native On-Res and Native Off-Res	12.2	9.8	2.4		

TABLE 38(b)

Percentage of Items with DIP on Sub-Test Three Dimensional Space

GROUP PAIR	% ITEMS ΔMH $\geq$ +1.50	% ITEMS ΔMH $\leq$ -1.50	% DIFFERENCE + and -	SIGNIFICANCE (*)	GROUP BIAS AGAINST
Norm 1 and 2	7.5	0.0	7.5		
Chinese 1 and 2	0.0	7.5	7.5		
Chinese 1 and Norm 1	17.5	7.5	10.0	*	Refer. (Norm 1)
Chinese 2 and Norm 1	7.5	5.0	2.5		
Native On-Res and Norm 1	7.5	2.5	5.0		
Native Off-Res and Norm 1	7.5	0.0	7.5		
Native On-Res and Native Off-Res	7.5	0.0	7.5		

TABLE 38(c)

Percentage of Items with DIP on Sub-Test Vocabulary

GROUP PAIR	% ITEMS Δ MH $\geq$ +1.50	% ITEMS ΔMH $\leq$ -1.50	% DIFFERENCE + and -	SIGNIFICANCE (*)	GROUP BIAS AGAINST
Norm 1 and 2	10.9	6.5	4.4		
Chinese 1 and 2	6.5	8.7	2.2		
Chinese 1 and Norm 1	30.4	19.6	10.8	*	Refer. (Norm 1)
Chinese 2 and Norm 1	28.3	10.9	17.4	*	Refer. (Norm 1)
Native On-Res and Norm 1	4.4	10.9	6.5		
Native Off-Res and Norm 1	6.5	17.4	10.9	*	Focal (Nat. Off-Res)
Native On-Res and Native Off-Res	10.8	8.9	1.9		

TABLE 38(d)

Percentage of Items with DIP on Sub-Test Arithmetic Reasoning

GROUP PAIR	% ITEMS Δ MH $\geq$ +1.50	% ITEMS Δ MH $\leq$ -1.50	% DIFFERENCE + and -	SIGNIFICANCE (*)	GROUP BIAS AGAINST
Norm 1 and 2	5.0	0.0	5.0		
Chinese 1 and 2	5.0	0.0	5.0		
Chinese 1 and Norm 1	10.0	30.0	20.0	*	Focal (Chin 1)
Chinese 2 and Norm 1	25.0	25.0	0.0		
Native On-Res and Norm 1	15.0	15.0	0.0		
Native Off-Res and Norm 1	0.0	5.0	5.0		
Native On-Res and Native Off-Res	10.0	0.0	10.0	*	Refer. (Nat. Off-Res)

Note first that for all four sub-tests a difference between the percentage of items containing positive DIP and those containing negative DIP was not found for both control group pairs (Norm 1 and 2; Chinese 1 and 2). Thus, the technique was considered to be functioning correctly. It was also observed from Tables 38(a) to 38(d) that, when non-equivalence of a sub-test for a group pair was found, the bias detected was not always to the detriment of a culture group. For example, both the Computation and Three Dimensional Space sub-tests appear biased in favour of Chinese. The Vocabulary sub-test is biased in favour of Chinese but to the detriment of Natives (off-reserve). On the other hand, The Arithmetic Reasoning sub-test is biased against Chinese.

It is also interesting to note that for two of the four sub-tests (Computation, Three Dimensional Space), the group against whom a sub-test is found to be biased is the group with the lower mean raw score (Norm 1 group). However, for sub-test Vocabulary, bias is found against the Norm 1 group and in favour of the two Chinese groups, while the two Chinese groups both have a lower mean raw score than the Norm 1 group. For the Vocabulary sub-test then, the ΔMH procedure indicates bias against a group (Norm 1) which has a higher

rather than the usual lower mean raw score. This result provides an indication that the ΔMH procedure is detecting cultural variation and is not simply a reflection of ability (as measured by total test score) differences.

From the contents of Tables 38(a) to 38(d), a culture group by decision (equivalence or non-equivalence) table can be constructed (Table 39).

TABLE 39

Culture Group by Decision (Equivalence/Non-Equivalence) Table for Correct/Incorrect Response Analysis (MH Technique)

	GATB SUB-TEST			
GROUP COMPARISON	COMPUTATION	THREE DIM SPACE	VOCAB	ARITHMETIC REASONING
Chinese 1 vs Chinese 2	1	1	1	1
Norm 1 vs Norm 2	1	1	1	1
Chinese 1 vs Norm 1	0	0	0	0
Chinese 2 vs Norm 1	0	1	0	1
Native (On-Res) vs Norm 1	1	1	1	1
Native (Off-Res) vs Norm 1	1	1	0	1
Native (On-Res) vs Native (Off-Res)	1	1	1	0

Decision Categories: Non-Equivalence = 0
Equivalence = 1

The two measurement techniques of Distractor Pattern Analysis and Correct/Incorrect Response Analysis (ΔMH) were taken, in this study, to represent the same measurement strategy of the modified model (see Table 5). If the two techniques can indeed be used to represent the same measurement strategy, it might be expected that the indices applied with the measurement techniques (a chi-square statistic in the one case and ΔMH in the other) would be correlated. However, because the MH Chi-Square is also a measure of DIP, it seems more appropriate in this case to correlate the chi-square values from both the distractor and MH analysis. Thus, the item by item chi-square values of the Distractor Pattern Analysis technique were correlated (for each GATB sub-test and each group pair compared) with the item-by-item chi-square values of the MH technique. The correlation coefficients are presented in Table 40.

TABLE 40

Correlations of Distractor Pattern Analysis
Chi-Square Values with Correct/Incorrect (MH) Chi-Square Values

SUB-TEST	GROUP PAIR	COEFFICIENT
Computation	Norm 1 and 2	-.0166
	Chinese 1 and 2	-.0101
	Chinese 1 - Norm 1	-.0406
	Chinese 2 - Norm 1	-.1148
	Native (On-Res) - Norm 1	-.1099
	Native (Off-Res) - Norm 1	-.0510
	Native (On-Res) - Native (Off-Res)	-.0012
Three Dimensional Space	Norm 1 and 2	-.0624
	Chinese 1 and 2	-.1733
	Chinese 1 - Norm 1	-.1497
	Chinese 2 - Norm 1	-.0563
	Native (On-Res) - Norm 1	.1219
	Native (Off-Res) - Norm 1	.2781
	Native (On-Res) - Native (Off-Res)	-.1420
Vocabulary	Norm 1 and 2	-.0774
	Chinese 1 and 2	-.1210
	Chinese 1 - Norm 1	.2006
	Chinese 2 - Norm 1	.1811
	Native (On-Res) - Norm 1	-.0011
	Native (Off-Res) - Norm 1	-.1032
	Native (On-Res) - Native (Off-Res)	-.1641
Arithmetic Reasoning	Norm 1 and 2	-.4560
	Chinese 1 and 2	.4123
	Chinese 1 - Norm 1	.3380
	Chinese 2 - Norm 1	.0172
	Native (On-Res) - Norm 1	-.1131
	Native (Off-Res) - Norm 1	.4085
	Native (On-Res) - Native (Off-Res)	-.0222

It is seen in the above table that correlation coefficients for the chi-square statistic of the Distractor Pattern Analysis technique with the chi-square value of the Correct/Incorrect Response Analysis (MH) technique are all fairly small, except for the one sub-test, where four of the seven values are moderately large. Although they are moderately large, there seems to be no apparent pattern. Little consistent relationship (linear) seems to exist between the two indicators of item bias. The two techniques appear to be measuring different things.

Section Two Results

The third hypothesis of the study concerns the consistency in judgements across measurement techniques and the hierarchical nature in the modified model (Figure 2) of the measurement strategies (as represented by specific techniques). This hypothesis can be addressed by examination of the pattern of decisions taken, using the different measurement techniques, about the equivalence or non-equivalence of each sub-test for the four culture groups in relation to the norm group and for the two native groups in relation to each other. To provide for such an examination, the results contained in the "Culture Group By Decision" tables derived for each measurement technique are aggregated into a single table for each GATB sub-test. The aggregated results are presented in Table 41. The measurement strategies of the modified model (Figure 2) which the various measurement techniques represent were given in Table 5. For ease of reference the measurement strategy associated with each measurement technique is included in the left-hand column of Table 41.

From Table 41 it is possible to examine the consistency in judgements (equivalence and non-equivalence) across the pairs of techniques representing the same strategy (e.g., the L_z and ECIZ4 indices representing the Response Pattern - IRT Approach strategy). It is more difficult to observe the pattern of equivalence or non-equivalence judgements across measurement strategies alone, which occur when just one technique is used to represent each strategy. Indeed, a "measurement strategy by decision category" table is needed for each GATB sub-test in order to investigate the hierarchical nature of the model. Thus, a series of "measurement strategy by decision category" tables are constructed for the GATB sub-tests, with the

measurement strategies of Within-Item Analysis - Non-IRT and Response Pattern - IRT Approach, each represented by one of two possible techniques. This results in the construction of four tables for each GATB sub-test, one for each combination of the two pairs of techniques representing the strategies of Within-Item Analysis - Non-IRT and Response Pattern - IRT Approach (e.g., one table for the combination of Distractor Analysis and the L_z index, another table for Distractor Analysis and the ECIZ4 index, and so on). The "measurement strategy by decision" tables are presented as Table 42.

TABLE 41

Decision Table Resulting from Applicaton of the Modified Model

SUB-TEST: COMPUTATION

MEASUREMENT STRATEGY REPRESENTED	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		CHIN 1 VS CHIN 2 (1)	NORM 1 VS NORM 2 (2)	CHIN 1 VS NORM 1 (3)	CHIN 2 VS NORM 1 (4)	NATIVE (ON-RES) VS NORM 1 (5)	NATIVE (OFF-RES) VS NORM 1 (6)	NATIVE (ON-RES) VS NATIVE (OFF-RES)(7)
Internal Structure Congruence	Comparative Factor Analysis (CFA)	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	Modified Caution Index (C*)	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	Distractor Analysis (DA)	1	1	0	0	1	0	1
	Correct/Incorrect Response Analysis (ΔMH)	1	1	0	0	1	1	1
Response Pattern - IRT Approach	Standardized Appropriateness Index (L_z)	1	1	0	0	1	1	1
	Standardized Extended Caution Index (ECIZ4)	1	1	0	0	1	1	1

TABLE 41 (cont'd)

Decision Table Resulting from Applicaton of the Modified Model

SUB-TEST: THREE DIMENSIONAL SPACE

MEASUREMENT STRATEGY REPRESENTED	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		CHIN 1 VS CHIN 2 (1)	NORM 1 VS NORM 2 (2)	CHIN 1 VS NORM 1 (3)	CHIN 2 VS NORM 1 (4)	NATIVE (ON-RES) VS NORM 1 (5)	NATIVE (OFF-RES) VS NORM 1 (6)	NATIVE (ON-RES) VS NATIVE (OFF-RES)(7)
Internal Structure Congruence	Comparative Factor Analysis (CFA)	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	Modified Caution Index (C*)	1	1	1	0	1	1	1
Within-Item Analysis Non-IRT	Distractor Analysis (DA)	1	1	0	1	1	1	0
	Correct/Incorrect Response Analysis (ΔMH)	1	1	0	1	1	1	1
Response Pattern - IRT Approach	Standardized Appropriateness Index (L_z)	1	1	0	0	1	1	1
	Standardized Extended Caution Index (ECIZ4)	1	1	0	0	1	1	1

TABLE 41 (cont'd)

Decision Table Resulting from Applicaton of the Modified Model

SUB-TEST: VOCABULARY

MEASUREMENT STRATEGY REPRESENTED	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		CHIN 1 VS CHIN 2 (1)	NORM 1 VS NORM 2 (2)	CHIN 1 VS NORM 1 (3)	CHIN 2 VS NORM 1 (4)	NATIVE (ON-RES) VS NORM 1 (5)	NATIVE (OFF-RES) VS NORM 1 (6)	NATIVE (ON-RES) VS NATIVE (OFF-RES)(7)
Internal Structure Congruence	Comparative Factor Analysis (CFA)	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	Modified Caution Index (C*)	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	Distractor Analysis (DA)	1	1	0	0	1	0	1
	Correct/Incorrect Response Analysis (ΔMH)	1	1	0	0	1	0	1
Response Pattern - IRT Approach	Standardized Appropriateness Index (L_z)	1	1	0	0	1	0	1
	Standardized Extended Caution Index (ECIZ4)	1	1	0	0	1	0	1

TABLE 41 (cont'd)

Decision Table Resulting from Applicaton of the Modified Model

SUB-TEST: ARITHMETIC REASONING

MEASUREMENT STRATEGY REPRESENTED	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		CHIN 1 VS CHIN 2 (1)	NORM 1 VS NORM 2 (2)	CHIN 1 VS NORM 1 (3)	CHIN 2 VS NORM 1 (4)	NATIVE (ON-RES) VS NORM 1 (5)	NATIVE (OFF-RES) VS NORM 1 (6)	NATIVE (ON-RES) VS NATIVE (OFF-RES)(7)
Internal Structure Congruence	Comparative Factor Analysis (CFA)	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	Modified Caution Index (C*)	1	1	1	1	1	0	1
Within-Item Analysis Non-IRT	Distractor Analysis (DA)	1	1	0	0	1	0	1
	Correct/Incorrect Response Analysis (ΔMH)	1	1	0	1	1	1	0
Response Pattern - IRT Approach	Standardized Appropriateness Index (L_z)	1	1	0	0	0	0	1
	Standardized Extended Caution Index (ECIZ4)	1	1	0	0	1	1	1

TABLE 42

Measurement Strategy by Decision Tables

SUB-TEST: COMPUTATION

(a)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	DA	1	1	0	0	1	0	1
Response Pattern - IRT Approach	L_z	1	1	0	0	1	1	1

(b)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	DA	1	1	0	0	1	0	1
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	1	1

TABLE 42 (cont'd)

SUB-TEST: COMPUTATION

(c)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	Δ MH	1	1	0	0	1	1	1
Response Pattern - IRT Approach	L_z	1	1	0	0	1	1	1

(d)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	ΔMH	1	1	0	0	1	1	1
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	1	1

TABLE 42 (cont'd)

SUB-TEST: THREE DIMENSIONAL SPACE

(e)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	DA	1	1	0	1	1	1	0
Response Pattern - IRT Approach	L_z	1	1	0	0	1	1	1

(f)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	DA	1	1	0	1	1	1	0
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	1	1

TABLE 42 (cont'd)

SUB-TEST: THREE DIMENSIONAL SPACE

(g)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	Δ MH	1	1	0	1	1	1	1
Response Pattern - IRT Approach	L_z	1	1	0	0	1	1	1

(h)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	Δ MH	1	1	0	1	1	1	1
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	1	1

TABLE 42 (cont'd)

SUB-TEST: VOCABULARY

(i)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	DA	1	1	0	0	1	0	1
Response Pattern - IRT Approach	L_z	1	1	0	0	1	0	1

(j)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	DA	1	1	0	0	1	0	1
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	0	1

TABLE 42 (cont'd)

SUB-TEST: VOCABULARY

(k)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	Δ MH	1	1	0	0	1	0	1
Response Pattern - IRT Approach	L_z	1	1	0	0	1	0	1

(l)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	1	1
Within-Item Analysis Non-IRT	ΔMH	1	1	0	0	1	0	1
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	0	1

TABLE 42 (cont'd)

SUB-TEST: ARITHMETIC REASONING

(m)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	0	1
Within-Item Analysis Non-IRT	DA	1	1	0	0	1	0	1
Response Pattern - IRT Approach	L_z	1	1	0	0	0	0	1

(n)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	0	1
Within-Item Analysis Non-IRT	DA	1	1	0	0	1	0	1
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	0	1

TABLE 42 (cont'd)

SUB-TEST: ARITHMETIC REASONING

(o)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	0	1
Within-Item Analysis Non-IRT	Δ MH	1	1	0	1	1	1	0
Response Pattern - IRT Approach	L_z	1	1	0	0	0	0	1

(p)

MEASUREMENT STRATEGY	MEASUREMENT TECHNIQUE	GROUP COMPARISON MADE						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internal Structure Congruence	CFA	1	1	1	1	1	1	1
Response Pattern - Non-IRT Approach	C*	1	1	1	1	1	0	1
Within-Item Analysis Non-IRT	ΔMH	1	1	0	1	1	1	0
Response Pattern - IRT Approach	ECIZ4	1	1	0	0	1	1	1

Discussion

It will be recalled that the purpose of this study was to determine the usefulness of a multi-strategy approach, as portrayed in the modified model of Figure 2 (Chapter Two), for demonstrating the equivalence in performance on a single test of people belonging to different cultural groups in relation to those from a norm group. In Chapter Two several tenets were presented for the modified model, which if found to hold true, would support the validity of the model. Based on the tenets of the model, the three hypotheses of this study, contained in Chapter Two, were stated.

In the first hypothesis it was stated that subjects drawn from the same population should be found equivalent in test performance using the various measurement techniques (Table 5) representing the measurement strategies of the modified model. Under the second hypothesis the expectation was stated that non-equivalence should be found for culture group subjects relative to norm group subjects, using at least some of the measurement techniques representing the measurement strategies of the modified model. The third hypothesis was related to the consistency in judgements of equivalence or non-equivalence for culture groups relative to the norm group expected when two measurement techniques representing the same measurement strategy were applied. Furthermore, under the third hypothesis it was specified that if non-equivalence was found for a culture group relative to the norm using a specific measurement technique representing a measurement strategy higher in the strategies array of the modified model (see Figure 2), it should be found for that group when all measurement techniques lower in the array were applied. This latter expectation was referred to as the hierarchical nature of the model.

In the following pages, the results described earlier under headings of Section One Results and Section Two Results are examined in the light of the three hypotheses of this study. Results related to each hypothesis are discussed under separate headings.

Hypothesis One

The first step in application of each measurement technique presented in Table 5 of Chapter Three was to determine if the technique could be considered to be functioning correctly. As stated in the first hypothesis, no differences should be found in the test performance of the two pairs of control groups if a technique was functioning as it should. From the contents of Section One Results, where the findings from application of each measurement technique were given, it can be seen that the test performance of the two Chinese groups, and of the two norm groups, was found to be equivalent for all six measurement techniques on all four GATB sub-tests. Although the results obtained from analyses of data related to the first hypothesis are possibly a consequence of type II error, it does seem reasonable to assume that the techniques were functioning as expected.

Hypothesis Two

With respect to the second hypothesis, that some of the measurement techniques should detect some test performance differences of culture groups in relation to a norm group, the observations reported in Table 41 of Section Two Results prove revealing. Looking at Table 41 it is seen that of 96 comparisons (taken over the four sub-tests) of culture groups with the

norm group (which excludes the control group and inter-native group comparisons), a decision of non-equivalence is made for 39 of the comparisons. That reflects a finding of non-equivalence in approximately 41% of the culture with norm group comparisons. There is only one measurement technique (Comparative Factor Analysis) with which a decision of non-equivalence was not reached for at least one comparison on one sub-test.

It is not surprising that equivalence was obtained in every case not only for the control group, but also for all culture with norm group comparisons using Comparative Factor Analysis. There seemed to be little likelihood that culture group subjects who share in a major way the "core culture" of all Canadians would be very different in terms of the constructs measured by a well-developed, validated standardized test such as the GATB.

The usefulness of the measurement techniques included in the modified model depends, of course, on the extent to which they can be used to detect differences due to culture. In relation to the study results it could be argued that the differences in test performance detected for the culture groups relative to the norm group were a consequence of one or more extraneous (conditioning) variables. Also, when measurement techniques not conditioned on ability are used, test performance differences could have occurred due to real ability differences among subjects belonging to the different groups. As explained earlier in Chapter Three (Methodology), the confounding effects of several extraneous factors were controlled through sampling procedures. From Table 3 of Chapter Three it was observed that the six groups compared in the study did not differ significantly in terms of the percentage of males and females in the groups. All subjects in all groups were attending secondary or post secondary schools located in the same geographical areas (except for Native Reserve Schools), and were

enrolled in the same types of programs (as discussed in Chapter Three, The Sample). Although the mean educational level of the groups differed somewhat, the maximum difference among means was 1.2 years. This was considerably less than the differences in education of two or more years normally associated with significant differences between groups in the levels of the aptitudes measured by the GATB. The age of study subjects was limited to between 17 and 24 years, younger than the age when a decline in the aptitudes measured by the instrument used in the study (GATB) has been reported to occur. Thus, variables not controlled through sampling and considered to possibly confound study findings were language fluency, degree of cultural exposure, and the real ability of subjects.

To isolate the possible confounding effects of language fluency (in this case, written fluency in English), the sub-tests of the GATB used to obtain dependent measures (test performance) were selected to provide two which are language-related and two which are not language-related. For the two language-related sub-tests (Vocabulary and Arithmetic Reasoning), non-equivalence was found, using the six measurement techniques of Table 5, for 23 of 48 culture groups with norm group comparisons (this excludes control group and inter-native group comparisons). For the two sub-tests not language-related (Computation and Three Dimensional Space), non-equivalence was found for 16 of 48 culture with norm group comparisons (see Table 41). Based on a chi-square statistic obtained from the two by two contingency table of test type (language-related versus non-language-related) by non-equivalence indication category (number of non-equivalence cases versus the number of equivalence cases), the difference in non-equivalence indication for the two test types was observed not to be significant. At the .05 level, a chi-square observed of 2.13 is obtained against a

chi-square critical, one degree of freedom, of 3.84. It would seem, then, that decisions of non-equivalence in test performance for culture groups relative to a norm group are probably not attributable to language fluency. In other words, it would appear that the differences in test performance detected in this study are more probably due to something which has been labelled culture, rather than to the effects of language fluency. However, this area needs to be looked at further.

There was some discussion in Chapter Three (Methodology) concerning the importance of degree of cultural exposure as a factor influencing test performance differences across groups. To get some idea of the possible effect of this variable on the GATB sub-test performance of study subjects, two groups were included in the study design, each assumed to be at one of two (out of four possible) levels of the cultural exposure continuum (Laveault, 1983). There were two groups of natives used in the study: natives living on reserves, who were assumed to represent culture group subjects at the acculturation level of the exposure continuum, and natives living off reserves, assumed to represent culture group subjects at the integration level. For these two groups, a significantly greater incidence of non-equivalence in test performance on the four GATB sub-tests of Native On-Reserve group relative to the norm group, as opposed to Native Off-Reserve group relative to the norm group, would imply an effect due to degree of cultural exposure. It might be expected that culture group subjects at the acculturation level would perform differently from the norm group more often than subjects at the integration level.

In examining the study findings contained in Table 41, it was observed that the test performance of Native On-Reserve group was declared non-equivalent with the performance of norm subjects, using the six

measurement techniques across the four sub-tests, in 1 of 24 instances. The test performance of Native Off-Reserve group was found to be non-equivalent to that of norm subjects in 8 out of 24 cases. It can be seen that non-equivalence indication relative to the norm group was found to be significantly different for the two native groups (at the .05 level with observed chi-square of 6.70 against a critical value of 3.84). It was surprising that the Native Off-Reserve group (integration level) performed differently on the sub-tests more often than Native On-Reserve group (acculturation level), the reverse of what was expected.

Throughout the analyses of performance on the four GATB sub-tests using the six measurement techniques, the two native groups were compared to each other. For those comparisons, it was found (Table 41) that of 24 cases there were just 2 where non-equivalence of the groups with each other was discovered. The implication is that the performance of the two groups on the four sub-tests, as measured by the six measurement techniques, could be considered equivalent.

For the two native groups in this study (on-reserve and off-reserve), it would seem that their performance on the four GATB sub-tests was, in most cases, equivalent to each other (as measured by the techniques applied as part of the modified model for cross-cultural equivalence). However, when the performance of each of the two groups was compared to the norm group, the two groups appeared to be two different cultural entities since one differed from the norm group and the other did not. It is, of course, quite possible to have two groups not be different from each other, yet only one differ from a third group, when mean values are compared. This occurs when the mean values for the two groups which are <u>not different</u> (in this case the

two native groups) are close together in magnitude, while the mean for one but not the other is sufficiently far away in magnitude from the third (the norm group) to be "significantly different". In fact, the expectation stated earlier that the Native On-Reserve group might be non-equivalent in test performance more often than Natives Off-Reserve was not met. The opposite frequently occurred. Such a result might be explained by the fact that the off-reserve natives were more heterogeneous in terms of their schooling than the on-reserve natives. This factor can mask the effect of degree of cultural exposure on test performance. It would seem likely that the two groups of natives selected for this study did not really fit the Laveault (1983) model which was used to explain the effect of degree of cultural exposure. In the end, the effect of degree of cultural exposure on test performance, as measured by the techniques of the modified model, is left unresolved.

It was possible in the study to control for the possible confounding influences of a number of extraneous factors. However, it was more difficult to be certain that the differences among groups in test performance, as detected by some of the measurement techniques, were due to cultural factors and were not just a reflection of real ability differences among subjects in the groups. As explained in Chapter Three (Methodology), several of the measurement techniques supposedly included the means to control for ability differences among groups compared. This was the case with L_z, ECIZ4, and ΔMH. Three of the techniques (C*, Comparative Factor Analysis, Distractor Analysis), did not contain methods for controlling ability differences.

Among the three techniques which employed ability level controls, there was a means of investigating degree of relationship with one indication of ability (total sub-test score) for just two of them (L_z and ECIZ4). By examining the correlation coefficients of Table 33 (obtained for a quadratic relationship of L_z or ECIZ4 with sub-test raw score, broken down by group) it would seem that neither index is related to a large degree to sub- test score.

Among the techniques not including specific procedures aimed at controlling ability differences among groups, a measure of association with one indicator of ability level (total sub-test score) could be obtained for the C* index. For the C* value, it seems to have little or no relationship with sub-test score, when examined at the total group (N=2177) level, or when broken down by groups (four culture and two norm groups).

For two of the techniques used in this study (Correct/Incorrect Response and Distractor Analysis) direct measures of association with an indicator of ability level (total sub-test score) cannot be obtained.

From the above remarks, it would seem that the six measurement techniques used in this study can be used to identify non-equivalence in test performance probably exclusive of the effect of ability level differences among group subjects. The effect of group ability level differences in use of the Comparative Factor Analysis technique could, however, not be directly ascertained.

Hypothesis Three

The third hypothesis of this study was based on the expectations that there would be consistency in judgements made about the equivalence or

non-equivalence of a sub-test for culture groups relative to a norm group when different techniques representing the same strategy were used. Under the third hypothesis it was also suggested that the hierarchical nature of the modified model would be found to hold true. With respect to the first part of the third hypothesis, concerning the consistency in judgements made using different techniques representing the same strategy, it was observed in Table 5 that there were two strategies (Response Pattern - IRT Approach and Within-Item Analysis - Non-IRT) represented by two techniques each. For the two techniques (L_z and ECIZ4 indices) representing the Response Pattern - IRT Approach strategy, it was evident from the results contained in Table 41 that for 16 group comparisons made (four culture group comparisons other than the control and inter-native group comparisons for each of four GATB sub-tests), conflicting findings were obtained in only 2 cases. Those differences were found for the comparisons of Natives (On-Res) and Natives (Off-Res) with Norm 1 on Sub-Test Arithmetic Reasoning. Thus, conflicting results for the two techniques representing the Response Pattern - IRT Approach were found in 12.5% of the comparisons made. For the two techniques representing the Within-Item Analysis - Non-IRT strategy, of 16 comparisons made (four group pairs over four sub-tests) conflicting findings were seen in 3 cases, that is, in 18.8% of the comparisons. Conflicting results occurred when Natives (On-Res) and Natives (Off-Res) were compared with the Norm 1 group on Sub-Test Arithmetic Reasoning and on Sub-Test Computation. Also, conflicting results were found when Chinese 2 was compared with the Norm 1 group on Sub-Test Arithmetic Reasoning. It is suggested here that one probable reason for the inconsistency in finding of equivalence and non-equivalence could be the nature of the sub-test where

this result occurred most often. It was in the Arithmetic Reasoning sub-test that inconsistencies mostly occurred. This sub-test is the shortest of the four (just 20 items), and showed, for the sample, the lowest reliability. In the end, the percentage of cases where inconsistencies occurred was small.

The second part of the third hypothesis is related to the hierarchical nature of the modified model. According to the hypothesis, a finding of non-equivalence on a sub-test for a culture group relative to a norm group using a measurement strategy (represented by a measurement technique) at the top of the array (Figure 2) would mean that non-equivalence should be observed when all strategies (represented by techniques) lower in the array were applied. To see if this hypothesis is valid, it is necessary to examine the equivalence or non-equivalence decisions reached for each pair of groups compared (e.g., Chinese 1 and Norm 1 but not including the inter-native comparisons) using just the four measurement strategies (where each strategy is represented by one technique only). Table 42 (given earlier) provides a display of the measurement strategies by decision categories which can be used for this purpose. Note that there is one table for each combination of measurement techniques (where two techniques represent a single strategy), for each GATB sub-test.

According to Hypothesis Three, the hierarchical nature of the model holds if, for each culture with norm group comparison, a decision of non-equivalence obtained for a strategy higher in the array is followed by decisions of non-equivalence when all strategies lower in the array are applied. The extent to which this condition holds can be examined by identifying the number of incidences of "reversals", or cases where

equivalence is found for a strategy or strategies lower in the array after non-equivalence was stated using a strategy higher in the array. Of course, it is possible that the hierarchical nature of the model might hold when one set of measurement techniques are used to represent the four measurement strategies of the model in Figure 2, but not when a different set of techniques are used. Thus, the incidence of "reversals" should be observed for each possible combination of "techniques representing strategies", and for each GATB sub-test. That is, the occurrence of reversals should be examined in relation to each of the sixteen parts of Table 42 (a to p). A summary of the incidences of reversals is given in Table 43. Note that the number of possible reversals for each column of equivalence (1) or non-equivalence (0) values shown under each group comparison made depends on when a zero value first occurs. For example, in Table 42(a), only one reversal is possible in column 3, only one is possible in column 4 and one in column 6, giving a total number of possible reversals for the table of 3. One reversal occurs in Table 42(a). It is in column 6.

TABLE 43

Summary of Decisions Reached Using the Modified Model

SUB-TEST	TABLE NO.	NUMBER OF POSSIBLE REVERSALS	NUMBER OF REVERSALS
Computation	42(a)	3	1
	42(b)	3	1
	42(c)	2	0
	42(d)	2	0
Three Dimensional Space	42(e)	1	0
	42(f)	1	0
	42(g)	1	0
	42(h)	1	0
Vocabulary	42(i)	3	0
	42(j)	3	0
	42(k)	3	0
	42(l)	3	0
Arithmetic Reasoning	42(m)	4	0
	42(n)	4	1
	42(o)	3	1
	42(p)	3	2
	TOTALS	40	6

Overall, it is seen in Table 43 that there are just 6 instances of a possible 40 when a reversal in decision pattern actually occurs. From the contents of Tables 42(a) to 42(p) and Table 43, it might also be suggested that the best combination of measurement techniques to represent the four strategies, in terms of minimizing the number of reversals, seems to be: (1) Comparative Factor Analysis, C*, Δ MH, L_Z or (2) Comparative Factor Analysis, C*, Distractor Analysis, L_Z. More of this will be said later in the discussion.

The hierarchy of the measurement strategies progresses from the general level (at the top of the array) to the specific level (at the bottom of the array). As the strategy becomes more specific, the equivalence assumptions challenged become more demanding. That is, it becomes more and more difficult to find equivalence of culture groups with a norm group as more specific strategies are used, and as the equivalence assumptions challenged become more demanding. Therefore, it might be expected that non-equivalence would be found more often when more specific, as opposed to more general measurement strategies are applied. Examining Table 42 again, it is seen that non-equivalence is not indicated at all for Internal Structure Congruence but is indicated in 2 of 16 cases for the Response Pattern - Non-IRT Approach (C*). For the Within-Item Analysis - Non-IRT strategy, non-equivalence is indicated in 7 of 16 cases when the Δ MH index is used and in 10 of 16 cases with Distractor Analysis. For the Response Pattern - IRT Approach, non-equivalence is found in 11 of 16 cases when the L_Z index is used, and in 9 of 16 cases when ECIZ4 is applied. Generally then, non-equivalence is seen to be indicated more frequently when specific measurement strategies are applied than when more general strategies are used. This provides further support for the notion of a hierarchy in measurement strategies and equivalence assumptions challenged.

The issue which has not been addressed is the extent to which techniques representing the same strategy are mutually replaceable. It is possible to correlate the indices used with techniques representing the same strategies. For the L_z and ECIZ4 indices, both representing the Response Pattern - IRT Approach strategy, the correlation between them is large for all four GATB sub-tests, as shown in Table 34. For the chi-square values of Distractor Pattern analysis and the chi-square values of Correct/Incorrect Response (MH) Analysis, both techniques representing the strategy of Within-Item Analysis - Non-IRT, just 4 of 28 possible correlations (seven group pairs compared across four sub-tests) is large (with the largest being -.46, according to Table 40). The remainder of the correlations are nearly zero. It would seem that these two techniques are probably not mutually replaceable. Thus, the techniques probably represent separate strategies for a model such as that contained in Figure 2 of Chapter Two. This is not, clearly, a surprising conclusion.

It is also questionable whether the two strategies of Response Pattern - Non-IRT Approach and Response Pattern - IRT Approach should be included in the model as separate strategies. The one (Response Pattern - Non-IRT Approach) can be used to challenge, according to the modified model, the item equivalence assumption only, while the other, Response Pattern - IRT Approach can be used to challenge both the item and the scalar equivalence assumption. Certainly, based on the findings of this study as contained in Table 41, use of techniques (C* on the one hand and L_z or ECIZ4 on the other) representing the two different strategies leads to different conclusions of equivalence or non-equivalence. However, C* is highly correlated with both L_z and ECIZ4 (see Table 34), for all four GATB sub-

tests. The high correlations are not surprising, given the findings of other researchers concerning the rather close relationship (correlations) among the various response pattern indices (Drasgow, 1983; Harnisch, 1983; Harnisch & Linn, 1981). It is, therefore, recommended that either L_z or ECIZ4 be retained in a model such as that in Figure 2, representing a measurement strategy of Response Pattern Analysis (IRT). It will be mentioned shortly why deletion from the model of the Response Pattern Approach - Non-IRT is probably indicated.

Overall Findings

From the results presented and discussed so far, it might be accepted that the six measurement techniques indicate differences among subjects in culture groups relative to those in a norm group due, for the most part, to cultural factors. Then, the next issue which arises is the extent to which the six techniques representing the four strategies of the modified model are mutually replaceable. It can be concluded, based on the patterns of decisions of equivalence or non-equivalence for the sub-tests and pairs of culture group with norm group comparisons used in this study, that applying each of the four strategies, represented by the six techniques, leads to independent judgements of equivalence or non-equivalence being made. However, the large correlation of C* (representing the Response Pattern - Non-IRT Approach strategy) with L_z and ECIZ4 (representing the Response Pattern - IRT Approach strategy) would cause one to believe that these two strategies are, in fact, mutually replaceable. Such a conclusion would be consistent with the evidence of similarity among such indices provided by

other researchers (Drasgow, 1983; Harnisch & Linn, 1981). Thus, it can be argued that a model for cross-cultural equivalence in test performance should include just the one measurement strategy of Response Pattern Analysis. Because Response Pattern Analysis using IRT techniques provides a control for ability level differences among subjects, and can be used to challenge both the item and scalar equivalence assumptions, it is recommended that the single Response Pattern Analysis strategy in the model be one which is based on IRT.

There is also one instance where two techniques representing the same strategy are not correlated. In this case (Distractor Pattern Analysis and Correct/Incorrect Response Analysis) the two techniques are considered, in the end, to represent separate measurement strategies. Thus, the modified model would be one containing four measurement strategies, with the "Within-Item Analysis - Non-IRT" being split into two strategies: Within-Item (Distractor Pattern) Analysis and Within-Item (Correct/Incorrect Response) Analysis, and the other strategies being Internal Structure Congruence and Response Pattern Analysis (IRT). A new model for cross-cultural equivalence in test performance, changed in accordance with the findings of this study, is presented as Figure 3.

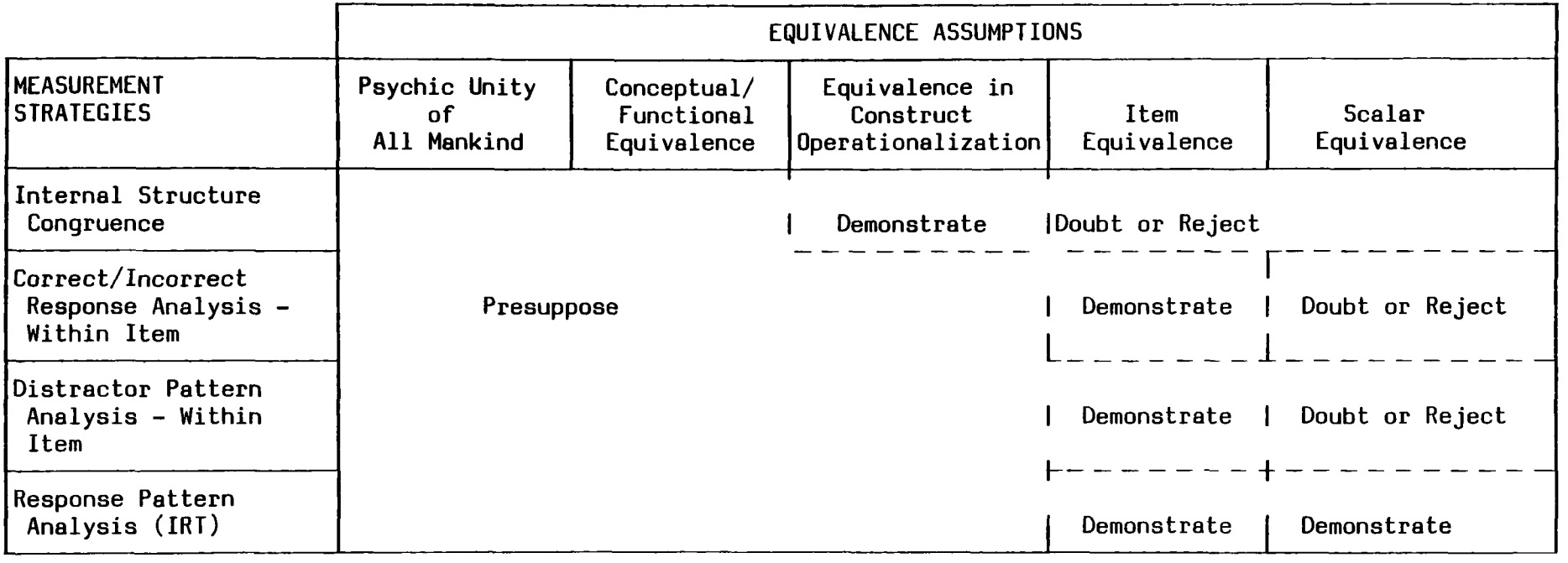
In Figure 3, it should first be observed that the Correct/Incorrect Response Analysis - Within-Item strategy (represented by the Δ MH index) is placed higher in the array than the Distractor Pattern Analysis - Within-Item strategy. The reason for indicating that the first (Δ MH) is more general in application than the second (Distractor Analysis) is because techniques used with the first strategy (like Δ MH) are unidimensional while those used with second (Distractor Analysis) are multidimensional. It might be expected that <u>rejection</u> of the item equivalence assumption would be

easier to attain with a strategy using multidimensional versus unidimensional techniques. Indeed, the findings in this study seem to support the inclusion of the two strategies in the order shown (non-equivalence indication for Δ MH occurs in 7 of 16 cases but for Distractor Analysis in 10 of 16 cases). It is suggested, though not displayed in Figure 3, that the L_z index might produce a higher rate of non-equivalence indication than the ECIZ4 index. To be sure, there are fewer cases of "reversals" in non-equivalence indication (2 compared with 4 cases) when L_z rather than ECIZ4 is used. The L_z index is also more highly correlated with C* than is ECIZ4. Thus, the L_z index is the preferable technique to be used with the Response Pattern Analysis (IRT) strategy.

Concerning the usefulness of the measurement techniques (representing strategies) included in the modified model (Figure 2 of Chapter Two), one final issue needs to be addressed. It might be argued that some or all of the six measurement techniques applied in this study could be used to identify non-equivalence in test performance of only high scorers or low scorers, but not both (Drasgow, 1983; Drasgow, Levine & Williams, 1987; Harnisch, 1983; Harnisch & Linn, 1981). By examining, for each GATB subtest, the mean group raw scores (Table 8) for groups identified as non-equivalent in test performance (Table 41), some interesting findings are revealed. For sub-tests Computation and Three Dimensional Space, the two Chinese groups are found to be non-equivalent, in relation to the Norm 1, using 5 of the 6 measurement techniques. For both of these sub-tests, the two Chinese groups have raw score means significantly higher than the Norm 1 group. For sub-test Vocabulary, both Chinese groups were found to be non-equivalent with the Norm 1 group using 5 of the 6 measurement techniques.

FIGURE 3

A Modified Model of Cross-Cultural Equivalence in Tests
(Based on Study Findings)



For this sub-test, the groups have mean raw scores significantly lower than that of the Norm 1 group. Thus, for the groups and tests of this study all of the measurement techniques except Comparative Factor Analysis can apparently be used to identify non-equivalence in test performance among both high-scoring and low-scoring subjects. As was stated earlier, Comparative Factor Analysis might indeed be used to identify non-equivalence, but was not found to do so in this study. However, the groups of subjects sampled were not really expected to differ much from norm subjects in terms of the constructs measured by the GATB.

Before leaving the subject of the usefulness of the measurement techniques (representing strategies), a few remarks about the strengths and weaknesses of each of the six measurement techniques, based on observations made in this study, should be helpful. Each of the techniques shown in Table 5 are considered in turn.

Comparative factor analysis

The analysis and findings of the present study indicate that this technique is difficult to apply, when the objective is to render a decision of equivalence or non-equivalence of a test across groups. Although criteria are cited in this study for deciding when correlation matrices in factor patterns are equivalent, the criteria are not strongly rooted in the cumulative findings of others, nor are the statistical distributions of the indices (GFI and RMSR) known. Further, a clear theory related to the factoral components of a test is needed <u>before</u> a comparative analysis of factor patterns is made. As argued in Chapter Two, where the modified model

(Figure 2) was presented, the researcher should empirically test the notion, not just assume, that conceptual/functional equivalence, and equivalence in construct operationalization exist for the different cultural groups with whom a test is used.

Modified caution index (C*)

This technique is relatively easy to compute and apply. In this study, it is also shown to be useful in identifying test performance differences between groups due, in large part, to culture. It does not, however, provide a control for differences in ability among subjects in culture groups. But the importance of such control is questionable in the light of the earlier discussion. The index is not, however, standardized. Although not specifically tested in this study, it is expected that high item omit rates among examinees would inflate the value of C*. Standardization alleviates this problem somewhat (Drasgow, Levine & Williams, 1987). High omit rate would most likely be of concern with speeded tests, such as those used in this study.

The standardized appropriateness index (L_z)

This technique requires considerable computing time to apply. It does provide a control for differences in ability level among subjects belonging to different groups. Moreover, the computation of an "expected" pattern of responses over all items on a test is achieved independently, without use of the test data from subjects in the groups eventually to be compared. In

placing examinees' results on a common scale, this index can be used to challenge the scalar as well as the item equivalence assumption (while the C* index cannot). It is also standardized, thus alleviating concern, to some degree, about high item omit rates.

The standardized extended caution index (ECIZ4)

There is little difference in the strengths and weaknesses of this technique relative to the L_z technique. It is also similar in concept and computation to C*, yet is standardized (alleviating concern over high omit rates), and involves the placement of examinees' results on a common scale. Thus, it can be used to challenge the scalar as well item equivalence assumption. However, use of the L_z index generally resulted in a higher rate of indication for, and fewer reversals in the pattern of non-equivalence. Thus, use of L_z is preferred to ECIZ4 in order to represent a Response Pattern Analysis (IRT) strategy.

Distractor analysis technique

The greatest strength of this technique is that it employs examinees' choices of incorrect response options to determine if cultural variation between groups is present. All of the other techniques are based on the use of correct responses to items on a test. If it is accepted that choice of distractor provides clues to the presence of cultural variation, this technique has much to offer. However, the choice of distractor may, as has been discussed, be a reflection of examinees' real ability level. In this study, the empirical link to ability level differences among groups of the

indices used in the technique to establish equivalence or non-equivalence of test performance is not clear. Another potential weakness of the technique, observed through its application in this study, is the absence of standard errors for a key index (Cramer's V) used in the procedure. As well, guessing may affect this technique to a greater extent (since there are multiple categories of response) than the ΔMH technique (which is based on a 1 or 0 dichotomy).

Correct/incorrect response analysis (ΔMH)

This very powerful technique, used in the context of a within-item analysis strategy, provides a control for ability differences among subjects in the groups compared. It is easy and inexpensive to compute. Probably the greatest disadvantage in the practical application of the technique is the absence of clear criteria strongly rooted in the cumulative findings of others, for determining when both items and whole tests are non-equivalent (biased).

From the results and the discussion presented so far, it might be concluded that the six measurement techniques representing the four strategies of the modified model (Figure 2) can all be used to detect differences in test performance of culture group subjects relative to norm group subjects. Further, differences in test performance can be demonstrated using the techniques exclusive of the confounding effects of sex, age, educational level and language fluency at least for the culture groups examined and for the kinds of aptitude tests used in this study. The role of degree of cultural exposure in the detection of test performance differences among groups is also not clarified through the findings of this

study. Further, the implications of ability level differences among subjects in the identification of non-equivalence in test performance using certain of the techniques applied in this study is still not clear.

It should be evident from the results presented that the hierarchical aspect of the measurement strategies, equivalence assumptions, and the correspondence between strategies and assumptions, presented in the model of Figure 2, generally hold true. Thus, it might be concluded that the various measurement strategies included in the model can probably be used, starting from the top of the measurement strategies array and moving down, to demonstrate or deny the equivalence in test performance of culture group people relative to norm group people by challenging equivalence assumptions which are more and more concrete in nature. As one moves down the measurement strategies array of this model (from general to specific), the assumption of equivalence challenged becomes increasingly concrete.

From the original Hui and Triandis model (Figure 1) presented in Chapter Two, specific aspects which relate to the measurement equivalence of a single test for culture group subjects relative to norm subjects were selected for investigation. The model actually investigated, and referred to as the modified model, was displayed in Figure 2 of Chapter Two. Based on the results of empirical study of the modified model, a further change to the model was proposed in this chapter. The modifications included in the model of Figure 3 related to the measurement strategies array only. The equivalence assumptions remained unchanged. It was finally suggested in the earlier discussion that the model, as proposed in Figure 3, might be found more useful in identifying non-equivalence in test performance of culture groups with a norm group than the model presented in Chapter Two.

It is emphasized that this is a descriptive study. Drawing firm conclusions about the usefulness of the Hui and Triandis model for cross-cultural equivalence in test performance is not possible. Instead, it is believed that the data presented in this study support the view that the model has some merit, but that further research in this area is needed. The results obtained in this study do indicate that it is possible to gauge the usefulness of various measurement techniques for determining the cross-cultural equivalence of a test through empirical study using real life data. Of course, the drawback to this approach is evident in this study. That is, it is never really known whether subjects classified in various culture groups are, in truth, culturally different from norm subjects. Strong conclusions about the usefulness of the modified model in identifying non-equivalence across groups are, therefore, difficult to obtain.

CHAPTER V

SUMMARY AND CONCLUSIONS

The purpose of this study was to investigate the usefulness of certain aspects of the Hui and Triandis model for demonstrating the equivalence in test performance of culture group subjects relative to subjects belonging to a norm group.

In the second chapter a model for the cross-cultural comparison of test performance proposed by Hui and Triandis (1985) was presented. It was explained that the model relates, in an hierarchical manner, equivalence assumptions which must be met for equivalence across groups to hold, to various measurement strategies used to challenge the assumptions. The model was shown to account for both measurement and relational equivalence (Drasgow, 1987). However, the specific interest of this study was to articulate a model focused solely on measurement equivalence. It will be recalled that measurement equivalence was said to exist if individuals with equal standing on the trait measured by the test, but drawn from different cultural groups would have equal expected test scores. Relational equivalence, on the other hand, was defined as the extent to which a bivariate measure of association of test scores with criterion scores was identical across relevant cultural groups. Clearly, measurement equivalence referred to the equivalence of groups on a single test, while relational equivalence was applied to equivalence in the association of a test with a criterion external to the test. Based on the notion of measurement equivalence, specific aspects of the Hui and Triandis model were selected for empirical investigation. Those aspects were presented as a modified

model. For both the original Hui and Triandis and the modified model tenets were offered. If the tenets were found to hold true, it was argued, the modified model would, based on the data in this study, appear to be valid. Thus, three hypotheses were stated which would be examined using real-life test data.

In the third chapter the measuring instrument used in the study, the sample selected, and the measurement techniques to be applied in the context of the modified model were described. The instrument chosen to obtain test performance data was the General Aptitude Test Battery (GATB). In fact, just four sub-tests of the battery were used. They were Computation, Three Dimensional Space, Vocabulary, and Arithmetic Reasoning. The first two sub-tests were not language-related while the latter two were. Subjects for the study were selected to represent both "culturally different" and "norm" members of the population. The culture group subjects for the study were Chinese immigrants and native peoples. Two groups of native peoples were included, those living on-reservations and those living off-reservations. These two groups were considered to represent people at two different levels of what was referred to as a cultural exposure continuum (Laveault, 1983). The Chinese subjects were, subsequently, randomly divided into two groups, as were the norm subjects. The reason for having groups drawn at random from the same population was to provide a means for making control group comparisons. That is, differences in test performance of culture group subjects with norm subjects might be expected, but subjects divided at random into two groups from the same population should not differ.

Six measurement techniques were described at length in the third chapter. Those techniques were used to represent the measurement strategies of the modified model (presented in the second chapter). The six

measurement techniques applied in the analyses of test performance data were Comparative Factor Analysis (MacDonald, 1985), Modified Caution Index (C*) (Harnisch, 1983), Standardized Appropriateness Index (L_z) and Standardized Extended Caution Index (ECIZ4) (Drasgow, Levine & Williams, 1987), Distractor Analysis (Veale & Foreman, 1983) and Correct/Incorrect Response Analysis (ΔMH) (Holland & Thayer, 1986).

In this study the first hypothesis, which stated that non-equivalence in test performance of culture group subjects relative to norm subjects would not be found for the control group comparisons using the six measurement techniques, was supported. The second hypothesis stated that non-equivalence should be found for culture group subjects relative to norm subjects using at least some of the measurement techniques. This hypothesis was generally supported, but the design of the study (descriptive and exploratory) makes it difficult to draw firm conclusions. The extent to which findings of non-equivalence in test performance were a result of certain extraneous factors (degree of cultural exposure and ability level differences among subjects) was not clear. Nor was it certain that non-equivalence was not artifactually produced by the methodology employed in the study. Under the third hypothesis it was postulated that there would be consistency in judgements of equivalence or non-equivalence using different measurement techniques representing the same measurement strategy. This hypothesis was only partially supported.

Consistency in judgements was obtained with two techniques (L_z and ECIZ4) representing the Response Pattern Approach - IRT strategy. However, this was not the case with the two techniques (Distractor Analysis and ΔMH) representing the Within-Item Analysis - Non-IRT strategy. It was concluded that the second set of techniques actually represented separate

strategies. In the end, another version of the modified model was proposed for further study. The revised model incorporated changes stemming from the findings of this study, principally the inclusion of four, rather than six, measurement techniques representing four measurement strategies.

Limitations and Future Research

Using real-life data in descriptive, exploratory research, as was done in the present study, does have its drawbacks. First of all, strong conclusions can rarely be obtained in such a study as this one. Secondly, it is never certain whether subjects sampled as part of a particular culture group are, in the end, really different in a cultural sense from those belonging to the norm group. Thus, it is difficult to determine whether test performance differences identified by any of the measurement techniques used in the study were real or artifactual. If the differences are accepted as real, the extent to which the differences can be attributed solely to cultural factors is not known. One solution to this dilemma is the use of simulated data. Future researchers, it is suggested, might conduct a study similar to this one, but use both simulated and real-life data. The incidences of non-equivalence in identification could then be compared for the two data sets, applying all the measurement techniques of the Hui and Triandis model.

In this study support for the second hypothesis was based on the large incidence (39 of 96 cases) of a finding of non-equivalence for culture group subjects relative to norm subjects. Unfortunately, it was not known in what percentage of cases non-equivalence might occur by chance. It is suggested that additional studies of the Hui and Triandis model, using both simulated and real-life data, are needed to establish the level of non-equivalence that might occur randomly.

The generalizability of the findings from this study is also quite limited. Both the culture and norm group subjects obtained for the study were not selected in order to achieve representativeness. Indeed, the geographical areas for sampling, the age and educational levels of subjects, as well as school program types from which subjects were drawn, were all very restricted. Furthermore, just 4 of 12 sub-tests of the GATB were used. Clearly, the Hui and Triandis model or any modified version of it could have a broader application only if subjects were selected to be representative, if groups other than Chinese and native peoples were involved, and if tests other than the GATB were used. It is suggested that an investigation of the version of the Hui and Triandis model derived in the fourth chapter of this study would be profitable. Such a study should include representative sampling from a larger number of culture groups in Canada, and should use several well-known aptitude and achievement tests to obtain test performance data. In particular, studies using non-speeded standardized tests would be profitable. The speededness of the GATB sub-tests used in the present study led to problems of high omit rates which should be avoided in future research. This is evident in the display of item difficulty values for the sub-tests given in Tables A5 to A8. A broader base of sampling may also enable the possible confounding effects of degree of cultural exposure and ability level differences among subjects to be better controlled.

With respect to the various measurement techniques applied in this study, much research remains to be done. Specifically, the statistical distribution of the C*, L_z and ECIZ4 indices are not well established. The effects on these indices of sample size, test length and group homogeneity-heterogeneity need to be examined. For the ΔMH statistic, the effect of interaction of ability level differences with group needs to be

studied further. The relationship of subjects' ability level to choice of distractor is still problematic in the Distractor Analysis technique. Also, for three of the measurement techniques used in this study (Comparative Factor Analysis, Distractor Analysis and Correct/Incorrect Response Analysis [Δ MH]), criteria to judge the equivalence in test performance of groups of subjects are not well-rooted in research findings. A replication of the present study, using data from an even larger "reference" group to establish baseline criteria (for example, to indicate the size of residuals that might be expected using the Comparative Factor Analysis technique) would be profitable. Broader use of these techniques in a variety of studies would also enable the establishment of sound criteria for equivalence or non-equivalence. Finally, broader use of five of the six measurement techniques of this study in cross-cultural settings would be helpful.

Two other possible strategies for challenging both the item and scalar equivalence assumptions of the Hui and Triandis (1985) model surfaced in this study. Drasgow and Guertler (1987) have offered an IRT index for detecting aberrant test performance based on a polychotomous model. They referred to the index, which is standardized and can account for high item omit rates, as Zh. The Zh index could potentially be used in the Hui and Triandis (1985) model, or a modified version of the model, as a replacement for the L_z, ECIZ4, and Distractor Analysis procedures. Another possibility for cross-cultural equivalence study might be the use in the model of structural equation modeling. The latter procedure has the potential for challenging all the equivalence assumptions of the modified model (Figures 2 or 3) simultaneously.

There should be no doubt that this was a descriptive, exploratory study. Neither the Hui and Triandis model nor many of the measurement techniques used to investigate the usefulness of the model, were previously studied in the context of culture group differences in test performance. By using real-life data, the findings of this study were, perhaps, rendered somewhat more applicable to a practical setting than if simulated data alone had been used. Certainly, the findings from this study would indicate that a multi-strategy approach (such as that contained in the Hui and Triandis model) to the determination of cross-cultural equivalence of tests has merit. It is also seen in this study that a multi-strategy approach is not as difficult as might be imagined.

BIBLIOGRAPHY

Angoff, W.H. (1972). A technique for the investigation of cultural differences. Paper presented at the American Psychological Association convention, Honolulu, HI.

Angoff, W.H., & Ford, S.F. (1973). Inter-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Arrasmith, D.G., Dizinno, G.A. (1988). Use of the mantel-haenszel statistic with test data from anglo and native american high school students. Paper presented at the American Educational Research Assocation Annual Meeting, New Orleans, LA, April, 1988.

Berk, R.A. (1982) (Ed). Handbook of methods for detecting test bias. Baltimore, MD: John Hopkins Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, Mass: Addison-Wesley.

Block, N.J., & Dworkin G. (1976). IQ, heritability and inequality. In N.J. Block and G. Dworkin, The IQ controversy, 410-540. New York: Pantheon Books.

Camilli, G. & Shepard, L.A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12(1), 87-99.

Carlson, J.R. & Sarrazin, G. (1984). Technical issues in the study of test bias. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Cleary, J.A., & Hilton, T.L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Cole, M. & Bruner, J.S. (1971). Cultural differences and inferences about psychological processes. American Psychologist, 26, 867-876.

Cox, D.R. (1970). Analysis of Binary Data. London: Methuen and Co.. Ltd.
Holland, P.W. (1985). On the study of differential item performance without IRT. Proceedings of the Military Testing Association, October, 1985.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Darlington, R.D. (1971). Another look at "culture fairness". Journal of Educational Measurement, 8, 71-82.

DeFries, J.C., Vandenberg, S.G., McClearn, G.E., Kuse, A.R., Wibon, J.R., Ashton, G.G., & Johnson, R.C. (1974). Near identity of cognitive structure in two ethnic groups. Science, 183 (1968), 338-339.

Donlon, T.F., & Fischer, F.F. (1968). An index of an individual's agreement with group determined item difficulties. Journal of Educational and Psychological Measurement, 28, 105-113.

Dorans, N.J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. Journal of Educational Measurement, 23(4), 355-368.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72(1), 19-29.

Drasgow, F. (1984). Scrutinizing psychological tests; Measurement equivalence and equivalent relations with external variables are the central issues. Psychological Bulletin, 95, 134-135.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.

Drasgow, F, & Guertler, E., (1987). A decision-theoretic approach to the use of appropriateness for detecting invalid test and scale scores. Journal of Applied Psychology, 72(1), 10-18.

Employment & Immigration Canada, (1987). Employment equity availability data report on designated groups. Ottawa: Employment and Immigration Canada.

Employment & Immigration Canada, (1976). Immigration and the canadian labour market. A study report presented by Research Projects Group, Strategic Planning and Research, Ottawa: Employment and Immigration Canada.

Employment & Immigration Canada, (1986). Immigration statistics, 1984. Published by Minister of Supply and Services Canada, ISBNO-662-54697-0.

Fairwether, G. (1986). The canadian industrial-organizational psychologist and human rights legislation: Present and future strategies for employment testing. Paper presented at the Annual Convention of the Canadian Psychological Association, Toronto, Ontario.

Feuerstein, R. (1979). The dynamic assessment of retarded performers. Baltimore, M.D.: University Park Press.

Frederiksen, N. (1977). How to tell if a test measures the same thing in different cultures. In Y.H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets and Zeitlinger B.V.

Ginsburg, H. (1972). The myth of the deprived child. Englewood Cliffs, N.J.: Prentice-Hall.

Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross classifications. Journal of American Statistical Association, 49, 732-764.

Gordon, M. (1964). Assimilation on american life. New York: Oxford University Press.

Greeley, A.M., & McCready, W.C. (1975). The transmission of cultural heritages: The case of the Irish and Italians. In N. Glazer and D. Moynihan (Eds.), Ethnicity: Theory and experience. Cambridge, Mass: Harvard University Press.

Hambleton, R.K. (1983). Applications of item response theory. Burnaby: Educational Research Institute of British Columbia.

Hambleton, R.K. & Rogers, J. (1988). Detecting biased test items: Comparison of the IRT area and mantel-haenszel methods. Paper presented at the NCME-AERA Joint Session, New Orleans, LA, April 1988.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory principles and applications. Boston: Kluwer-Nijhoff Publishing.

Harman, H.H. (1964). Modern factor analysis. Chicago and London, 256-258: The University of Chicago Press.

Harnisch, D.L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20(2), 191-206.

Harnisch, D.L., & Linn, R.L. (1981a). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18(3), 133-146.

Harnisch, D.L., & Linn, R.L. (1981b). Identification of aberrant response patterns (Final Report. Educational Commission of the States, Denver, Colo., National Assessment of Educational Progress). Illinois University: Champaign.

Harnisch, D.L., Kuo, S., & Torres, R.T. (1983). Student Problem Package User's Guide. Champaign, Illinois: Office of Educational Testing, University of Illinois.

Hennessy, J.J., & Merrifield, P.R. (1976). A comparison of the factor structures of mental abilities in four ethnic groups. Journal of Educational Psychology, 68, 754-759.

Holland, P.W. & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. Program Statistics Research (Technical Report Number 86-69), New Jersey, Educational Testing Sevice.

Hui, H.C., & Triandis, H.C. (1983). Multistrategy approach to cross-cultural research: The case of locus of control. Journal of Cross-cultural Psychology, 14, 65-83.

Hui, H.C., & Triandis, H.C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. Journal of Cross-Cultural Psychology, 16(2), 131-152.

Hunter, J.E., Schmidt, F.L., & Rauschenberger, J.M. (1967). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 62, 245-260.

Insua, A.M. (1983). WAIS-R factor structures in two cultures. Journal of Cross-Cultural Psychology, 14(4), 427-438.

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

Jensen, A.R. (1980). Bias in mental testing. New York: Free Press.

Jensen, A.R. (1984). Test bias: Concepts and criticisms. In C.R. Reynolds and R.T. Brown (Eds), Perspectives on bias in mental testing, 507-586. New York: Plenum Press.

Jöreskog, K.G. & Sörbom, D. (1985). Lisrel VI, analysis of linear structural relationships by the method of maximum likelihood, a users' guide. Mooresville, Ind.: Scientific Software Inc.

Kane, M.T., & Brennan, R.L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, 4, 105-126.

Kluckholn, C. (1967). The study of culture. In P.I. Rose (Ed.), The study of society. New York: Random House.

Kohr, Richard (1977). Cultural validity main program (SR Report No. 8). Pennsylvania Department of Education.

Kok, F.G., Mellenbergh, G.J., & Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. Journal of Educational Measurement, 22(4), 295-303.

Laveault, D. (1983). Problèmes liés à l'évaluation interculturelle. Mesure et évaluation en éducation, 6(3), 79-91.

Levine, M.V., & Drasgow, F. (1983). Appropriateness measurement. Validating studies and variable ability models. In D.J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Levine, M.V. & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple choice test scores. Journal of Educational Statistics, 4, 269-290.

Lindquist, E.F. (1953). Design and analysis of experiments in psychology and education, 78-90, Boston: Houghton Mifflin Company.

Linn, R.L. (1984). Selection bias: Multiple meanings. Journal of Educational Measurement, 21(1), 33-47.

Linn, R.L. (1982). Ability testing: Individual differences, prediction and differential prediction. In A.K. Wigdor and W.R. Gardner (Eds) Ability testing: Uses, consequences and controversies, 335-338. Washington, D.C.: National Academy Press.

Linn, R.L. (1973). Fair test use in selection. Review of Educational Research, 43(2), 139-161.

Lord, F.M. (1977). A study of item bias, using item characteristic curve theory. In Y.H. Poortinga (Ed), Basic problems in cross-cultural psychology, 19-29. Amsterdam: Swets and Zeitlinger.

Lord, F.M. (1980). Applications of item response theory in practical testing problems. Hillsdale, N.J.: Erlbaum.

MacArthur, R.S. (1975). Differential ability patterns: Inuit, Nsenga, and Canadian whites. In J.W. Berry and J. Lonner, Applied cross-cultural psychology. Amsterdam: Swets and Zeitlinger B.V.

Malpass, R. (1977). Theory and method in cross-cultural psychology. American Psychologist, 32, 1069-1079.

Marascuilo, L.A. & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on X^2 statistics. Journal of Educational Measurement, 18(4), 229-248.

Martois, J.S., Rickard, P.L., & Stiles, R.L. (1988). Use of the Mantel-Haenszel statistic with test data from workfair participants. Paper presented at the NCME-AREA Joint Session, New Orleans, LA., April, 1988.

McCauley, D.E., & Colberg, M. (1983). Transportability of deductive measurement across cultures. Journal of Educational Measurement, 20(1), 81-92.

McDonald, R.P. (1985). Factor analysis and related methods. Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

Miller, D.M. (1986). Time allocation and patterns of item response. Journal of Educational Measurement, 23(2), 147-156.

Miller, J., Slomczynski, L.M., & Schoenberg, R.J. (1981). Assessing comparability of measurement in cross-national research: Authoritarian-conservation in different socio-cultural settings. Social Psychology Quarterly, 44, 178-191.

Pedhazar, E.J. (1982). Multiple regression in behavorial research, New York: Holt, Rinehart, and Winston.

Petersen, N.S., & Novick, M.R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13, 3-29.

Reynolds, C.R. (1983). Test bias: In God we trust, all others must have data. Journal of Special Education, 17, 241-260.

Riessman, F. (1974). The hidden IQ.. In A. Gartner, C. Greer, and F. Riessman (Eds), The new assault on equality: IQ and social stratification, 206-223. New York: Harper and Row.

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.

Sato, T., & Kurata, M. (1977). Basic S-P score table characteristics. NEC Research and Development, 47(1), 1977, pp. 64-71.

Sattler, J.M. (1982). Assessment of children's intelligence and special abilities (2nd ed). Boston: Allyn and Bacon.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16(3), 143-152.

Schmidt, F.L., & Hunter, J.E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

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

Sung, Y.H., & Dawis, R.V. (1981). Level and factor structure differences in selected abilities across race and sex groups. Journal of Applied Psychology, 66(5), 613-624.

Tatsuoka, K.K., & Linn, R.L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential application. Applied Psychological Measurement, 7(1), 81-96.

Tatsuoka, K.K., & Tatsuoka, M.M. (1982a.). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7, 215-231.

Tatsuoka, K.K., & Tatsuoka, M.M. (1982b). Standardized extended caution indices and comparison of their rule detection rates (Research Report 82-4-ONR). University of Illinois: Urbana.

Thorndike, R.L. (1982). Applied psychometrics. Boston: Houghton, Mifflin.

Trabin, T.E. & Weiss, D.J. (1983). The person response curve: Fit of individuals to item response theory models. In D.J. Weiss (Ed), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Triandis, H.C. (1978). Some universals of social behaviour. Personality and Social Psychology Bulletin, 4, 1-16.

Triandis, H.C. (1983). Multistrategy approach to cross-cultural research: The case of locus of control. Journal of Cross-Cultural Psychology, 14, 65-83.

United States Department of Labor, Manpower Administration (1970). General Aptitude Test Battery: Section III - Development. Washington, D.C.: United States Government Printing Office.

Van der Flier, H., Mellenbergh, G.J., Adler, H.J. & Wijn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21(2), 131-145.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. Journal of Cross-Cultural Psychology, 13, 267-298.

Van der Flier, H. (1977). Environmental factors and deviate response patterns. In Y.H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets and Zeitlinger, B.V.

Veale, J.R. (1988). Cultural variation in foil selection I: A mathematical model for cultural bias in multiple choice items and some empirical data relating to its validity (SR Report No. 12). Pennsylvania Department of Education.

Veale, J.R., & Foreman, D.I. (1983). Assessing cultural bias using foil response data: Cultural variation. Journal of Educational Measurement, 20(3), 249-258.

Welch, C.J., Ackerman, T.A., Doolittle, A.E., Hurley, J. (1987). An examination of statistical procedures for detecting cross-cultural differential item performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April 1987.

Werner, O., & Campbell, D.T. (1970). Translating, working through interpreters, and the problem of decentering. In R. Narroll & R. Cohens (Eds), A Handbook of Methods in Cross-Cultural Methodology, 398-420, New York: Natural History Press.

Williams, R.L. (1971). Abuses and misuses in testing black children. Counselling Psychology, 2, 62-67.

Winer, B.J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

Wingersky, M.S., Barton, M.A., & Lord, F.M. Logist User's Guide. Princeton, N.J.: Educational Testing Service.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

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 in San Francisco, April 1986.

APPENDIX

ADDITIONAL TABLES

TABLE A1

Estimated Item Parameters For Sub-Test Computation

ITEM NO	DISCRIMINATION(A)	STD ERR	DIFFICULTY(B)	STD ERR	GUESSING(C)	STD ERR
1	0.38321	0.421	-4.65717	0.555	0.05152	0.214
2	0.37799	0.132	-4.66823	0.625	0.05162	0.166
3	0.78453	0.405	-4.16002	3.421	0.05162	0.110
4	0.93554	0.078	-3.33369	1.006	0.05162	0.102
5	0.73559	0.116	-1.93114	0.266	0.05162	0.024
6	0.67427	0.445	-3.37121	0.701	0.05162	0.010
7	1.15915	0.406	-1.14932	0.454	0.05162	0.006
8	0.39070	0.123	-5.51016	0.233	0.05162	1.015
9	0.86041	0.147	-2.66316	0.124	0.05162	0.013
10	0.97036	0.076	-1.66151	0.142	0.05162	0.058
11	0.50750	0.063	-2.71684	0.065	0.05162	0.049
12	0.98771	0.063	-1.74106	0.099	0.05162	0.006
13	0.98080	0.079	-0.88751	0.061	0.05162	0.006
14	0.90464	0.133	-2.13897	0.086	0.05162	0.011
15	1.79103	0.156	-1.41298	0.042	0.05162	0.094
16	0.92756	0.159	-1.72010	0.056	0.07924	0.075
17	1.09126	0.069	-1.20915	0.046	0.10138	0.044
18	0.89166	0.144	-1.03988	0.063	0.06005	0.025
19	1.05896	0.066	-0.95462	0.061	0.02368	1.025
20	0.95012	0.052	-0.65162	0.090	0.01420	0.006
21	1.48056	0.075	-0.69855	0.074	0.00933	0.006
22	2.13156	0.401	-0.31952	0.062	0.06649	0.022
23	1.80747	0.399	-0.09177	0.177	0.03774	0.011
24	1.47851	0.244	0.26898	0.025	0.04794	0.015
25	1.24370	0.301	0.78782	0.239	0.02164	0.004
26	2.17440	0.412	0.68922	0.087	0.06062	0.006
27	1.84378	0.402	1.09449	0.082	0.03179	0.004
28	2.06452	0.220	1.29335	0.044	0.03325	0.003
29	2.17440	0.179	1.57275	0.068	0.01294	0.011
30	1.90974	0.395	1.93579	0.031	0.01038	0.022
31	2.17440	0.420	2.42357	0.083	0.03750	0.010
32	2.17440	0.417	2.71944	0.110	0.03052	0.061
33	2.17440	0.445	3.49019	0.107	0.00626	0.055
34	2.15803	0.395	3.83036	0.125	0.00872	0.020
35	2.17440	0.436	3.91659	0.119	0.00361	0.045
36	1.47996	0.478	4.46431	0.135	0.00364	0.033
37	2.17440	1.410	4.43894	0.147	0.00177	0.083
38	1.25385	0.482	4.94052	0.191	0.00261	0.041
39	Not used in estimation procedure					
40	Not used in estimation procedure					
41	Not used in estimation procedure					

TABLE A2

Estimated Item Parameters For Sub-Test Three Dimensional Space

ITEM NO	DISCRIMINATION(A)	STD ERR	DIFFICULTY(B)	STD ERR	GUESSING(C)	STD ERR
1	0.56277	0.097	-3.87366	0.617	0.14567	0.123
2	0.73313	0.082	-2.03051	0.240	0.14567	0.091
3	1.04056	0.108	-1.50608	0.149	0.14567	0.073
4	0.79941	0.089	-2.08711	0.227	0.14567	0.091
5	1.17962	0.125	-1.63542	0.141	0.14567	0.073
6	0.45447	0.062	-2.23850	0.416	0.14567	0.109
7	1.63240	0.197	-1.17906	0.111	0.30902	0.065
8	0.58239	0.069	-1.27349	0.256	0.14567	0.087
9	0.65082	0.076	-0.80796	0.203	0.14567	0.075
10	0.67064	0.076	-1.86060	0.249	0.14567	0.091
11	0.60800	0.072	-2.22620	0.298	0.14567	0.097
12	0.89530	0.064	-0.96528	0.065	0.00000	0.002
13	0.42520	0.060	-1.75160	0.413	0.14567	0.107
14	0.78203	0.084	-0.88762	0.165	0.14567	0.069
15	1.47754	0.148	-0.81409	0.085	0.19019	0.050
16	0.96955	0.097	-0.78570	0.122	0.14567	0.058
17	0.96035	0.097	-1.02912	0.135	0.14567	0.064
18	0.60800	0.074	-0.65733	0.215	0.14567	0.075
19	0.93599	0.095	-0.87080	0.132	0.14567	0.061
20	0.82869	0.086	-0.23457	0.115	0.10996	0.049
21	1.07498	0.098	-0.18611	0.077	0.09210	0.037
22	0.89620	0.097	0.15655	0.096	0.12113	0.040
23	2.00000	0.180	-0.10537	0.043	0.13101	0.026
24	1.87928	0.185	0.33565	0.045	0.15042	0.022
25	2.00000	0.188	0.27908	0.041	0.12999	0.021
26	2.00000	0.185	0.44835	0.038	0.09013	0.017
27	1.71866	0.181	0.73839	0.045	0.09975	0.017
28	2.00000	0.237	0.78648	0.044	0.15324	0.018
29	1.49516	0.182	1.50613	0.067	0.03545	0.009
30	2.00000	0.262	1.40160	0.054	0.06889	0.010
31	2.00000	0.267	1.64114	0.062	0.03099	0.007
32	2.00000	0.302	1.66465	0.065	0.06306	0.009
33	2.00000	0.294	1.59260	0.066	0.04822	0.008
34	2.00000	0.352	2.01062	0.085	0.03601	0.007
35	1.32017	0.476	3.05624	0.365	0.02319	0.006
36	1.32017	0.465	3.19301	0.413	0.00947	0.004
37	0.98599	1.579	4.64205	3.652	0.02134	0.006
38	2.00000	1.191	3.21364	0.419	0.01523	0.004
39	Not used in estimation procedure					
40	Not used in estimation procedure					

TABLE A3

Estimated Item Parameters For Sub-Test Vocabulary

ITEM NO	DISCRIMINATION(A)	STD ERR	DIFFICULTY(B)	STD ERR	GUESSING(C)	STD ERR
1	0.62434	0.104	-3.91085	0.628	0.05720	0.202
2	0.22406	0.087	-6.12668	6.334	0.05720	1.403
3	0.58407	0.097	-3.87813	0.656	0.05720	0.212
4	0.94752	0.121	-2.62890	0.258	0.05720	0.112
5	0.60932	0.090	-3.37798	0.508	0.05720	0.175
6	0.71905	0.076	-1.73092	0.223	0.05720	0.099
7	0.66088	0.075	-2.02420	0.274	0.05720	0.115
8	1.26724	0.139	-2.02262	0.144	0.05720	0.080
9	0.60761	0.068	-1.11696	0.234	0.05720	0.093
10	0.42487	0.062	-1.38822	0.455	0.05720	0.139
11	0.39547	0.063	-2.61032	0.705	0.05720	0.210
12	0.62125	0.069	-1.21395	0.232	0.05720	0.094
13	0.76671	0.079	-0.44976	0.134	0.05720	0.059
14	0.66856	0.073	-1.68947	0.241	0.05720	0.103
15	1.37546	0.142	0.14168	0.062	0.14736	0.031
16	0.72619	0.074	-1.05281	0.176	0.05720	0.078
17	0.52092	0.069	-0.30168	0.241	0.05730	0.082
18	1.16517	0.095	-0.73731	0.075	0.03186	0.039
19	1.26605	0.124	0.01939	0.066	0.11308	0.033
20	1.34059	0.132	0.10033	0.061	0.11898	0.031
21	1.12420	0.093	-0.90840	0.084	0.03265	0.044
22	1.45051	0.116	-0.20069	0.048	0.03728	0.024
23	1.41952	0.127	-0.16966	0.057	0.08533	0.031
24	0.98175	0.064	-0.14087	0.044	0.00000	0.002
25	0.64172	0.055	1.15538	0.094	0.00000	0.006
26	1.81634	0.159	0.04793	0.041	0.07905	0.022
27	0.88396	0.078	0.42406	0.063	0.01030	0.021
28	1.55921	0.157	0.96617	0.050	0.05878	0.013
29	1.18249	0.122	1.32931	0.068	0.02379	0.011
30	1.39780	0.148	1.37786	0.063	0.02982	0.010
31	1.46888	0.156	1.26563	0.059	0.04290	0.011
32	1.41509	0.124	0.94873	0.049	0.01695	0.009
33	1.22331	0.130	1.65997	0.080	0.00935	0.007
34	1.87959	0.239	1.67251	0.061	0.02304	0.006
35	1.49635	0.188	1.84854	0.080	0.01422	0.006
36	2.00000	0.405	2.14665	0.089	0.04806	0.007
37	2.00000	0.439	2.33996	0.106	0.03408	0.006
38	2.00000	0.445	2.40862	0.113	0.02621	0.005
39	2.00000	0.403	2.37094	0.107	0.01381	0.004
40	2.00000	0.463	2.44067	0.117	0.02971	0.006
41	2.00000	0.478	2.60167	0.136	0.01332	0.004
42	1.80103	0.406	2.79129	0.174	0.00156	0.001
43	Not used in estimation procedure					
44	Not used in estimation procedure					
45	Not used in estimation procedure					
46	Not used in estimation procedure					

TABLE A4

Estimated Item Parameters For Arithmetic Reasoning

ITEM NO	DISCRIMINATION(A)	STD ERR	DIFFICULTY(B)	STD ERR	GUESSING(C)	STD ERR
1	0.47847	0.087	-3.75047	0.757	0.12339	0.191
2	0.77277	0.090	-1.80074	0.243	0.12339	0.111
3	1.14971	0.160	-2.67029	0.250	0.12339	0.123
4	1.06991	0.118	-1.68130	0.170	0.12339	0.094
5	0.41622	0.065	-2.52362	0.593	0.12339	0.165
6	1.22565	0.137	-1.71481	0.152	0.12339	0.090
7	1.19407	0.110	-0.58844	0.079	0.07499	0.045
8	0.78786	0.088	-0.94633	0.180	0.12339	0.083
9	0.61210	0.081	-0.05123	0.200	0.12339	0.072
10	1.31977	0.142	-0.58884	0.093	0.19090	0.055
11	1.11384	0.099	-0.19078	0.066	0.04664	0.033
12	1.28183	0.127	-0.12557	0.068	0.11942	0.037
13	2.00000	0.183	0.16598	0.037	0.08738	0.020
14	1.58052	0.180	1.25360	0.059	0.06098	0.012
15	2.00000	0.253	1.33990	0.053	0.06790	0.010
16	1.09804	0.185	2.55974	0.147	0.02509	0.007
17	2.00000	0.515	2.75186	0.119	0.04800	0.007
18	2.00000	0.627	3.19655	0.164	0.02335	0.005
19	2.00000	0.738	3.43462	0.203	0.01605	0.004
20	2.00000	0.905	3.71291	0.272	0.00982	0.003

TABLE A5

Item Difficulty (P) Values For Sub-Test Computation

ITEM NO	P	STD DEV
1	.9850	.1215
2	.9956	.0663
3	.9921	.0887
4	.9868	.1142
5	.8811	.3239
6	.9674	.1777
7	.8211	.3834
8	.9709	.1681
9	.9595	.1973
10	.8863	.3175
11	.9225	.2676
12	.8987	.3019
13	.7445	.4363
14	.9295	.2561
15	.9101	.2861
16	.8881	.3154
17	.8264	.3789
18	.7674	.4227
19	.7683	.4221
20	.6837	.4652
21	.7295	.4444
22	.6370	.4811
23	.5463	.4981
24	.4370	.4962
25	.2890	.4535
26	.2960	.4567
27	.1938	.3955
28	.1515	.3587
29	.1031	.3042
30	.0731	.2605
31	.0319	.1910
32	.0308	.1730
33	.0159	.1250
34	.0088	.0935
35	.0088	.0935
36	.0062	.0783
37	.0035	.0593
38	.0026	.0514
39	.0000	.0000
40	.0009	.0297
41	.0018	.0420

TABLE A6

Item Difficulty (P) Values For Sub-Test Three Dimensional Space

ITEM NO	P	STD DEV
1	.9568	.2033
2	.8855	.3186
3	.8661	.3407
4	.9040	.2948
5	.8916	.3110
6	.8388	.3679
7	.8837	.3207
8	.7727	.4193
9	.7119	.4531
10	.8599	.3472
11	.8846	.3197
12	.7313	.4435
13	.7850	.4110
14	.7463	.4353
15	.7921	.4060
16	.7410	.4383
17	.7850	.4110
18	.6828	.4656
19	.7559	.4297
20	.6026	.4896
21	.5903	.4920
22	.5189	.4999
23	.5938	.4913
24	.4714	.4994
25	.4740	.4995
26	.3982	.4898
27	.3313	.4709
28	.3533	.4782
29	.1366	.3435
30	.1648	.3711
31	.0996	.2995
32	.1269	.3330
33	.1101	.3132
34	.0722	.2590
35	.0335	.1800
36	.0167	.1284
37	.0229	.1497
38	.0176	.1316
39	.0132	.1142
40	.0176	.1316

TABLE A7

Item Difficulty (P) Values For Sub-Test Vocabulary

ITEM NO	P	STD DEV
1	.9692	.1730
2	.9084	.2886
3	.9630	.1889
4	.9533	.2111
5	.9498	.2185
6	.8449	.3621
7	.8670	.3398
8	.9348	.2470
9	.7304	.4439
10	.7216	.4484
11	.8370	.3695
12	.7507	.4328
13	.6264	.4840
14	.8282	.3774
15	.5269	.4995
16	.7410	.4383
17	.5753	.4945
18	.7145	.4518
19	.5436	.4983
20	.5233	.4997
21	.7524	.4318
22	.5744	.4946
23	.5859	.4928
24	.5322	.4992
25	.2608	.4393
26	.5154	.5000
27	.3885	.4876
28	.2493	.4328
29	.1753	.3804
30	.1586	.3655
31	.1841	.3878
32	.2273	.4193
33	.1119	.3154
34	.0987	.2972
35	.0819	.2744
36	.0485	.2148
37	.0344	.1822
38	.0264	.1605
39	.0361	.1867
40	.0300	.1705
41	.0273	.1631
42	.0123	.1104
43	.0053	.0725
44	.0035	.0593
45	.0053	.0725
46	.0079	.0887

TABLE A8

Item Difficulty (P) Values For Sub-Test Arithmetic Reasoning

ITEM NO	P	STD DEV
1	.9410	.2358
2	.8722	.3340
3	.9674	.1777
4	.8960	.3053
5	.8493	.3579
6	.9101	.2861
7	.6899	.4628
8	.7471	.4348
9	.5648	.4960
10	.7339	.4421
11	.5648	.4960
12	.5806	.4937
13	.4643	.4989
14	.2044	.4034
15	.1877	.3906
16	.0696	.2546
17	.0678	.2516
18	.0335	.1800
19	.0229	.1497
20	.0141	.1179

TABLE A9

GATB Sub-test Reliability Coefficients Broken Down by Group

SUB-TEST COMPUTATION		
GROUP	NUMBER	KR-20 RELIABILITY
Chinese 1	333	0.84
Chinese 2	347	0.84
Native (On-Res)	386	0.85
Native (Off-Res)	371	0.89
Norm 1	366	0.85
Norm 2	374	0.84

SUB-TEST THREE DIMENSIONAL SPACE		
GROUP	NUMBER	KR-20 RELIABILITY
Chinese 1	333	0.87
Chinese 2	347	0.89
Native (On-Res)	386	0.82
Native (Off-Res)	371	0.87
Norm 1	366	0.87
Norm 2	374	0.87

SUB-TEST VOCABULARY		
GROUP	NUMBER	KR-20 RELIABILITY
Chinese 1	333	0.80
Chinese 2	347	0.81
Native (On-Res)	386	0.83
Native (Off-Res)	371	0.88
Norm 1	366	0.86
Norm 2	374	0.86

SUB-TEST ARITHMETIC REASONING		
GROUP	NUMBER	KR-20 RELIABILITY
Chinese 1	333	0.69
Chinese 2	347	0.63
Native (On-Res)	386	0.73
Native (Off-Res)	371	0.75
Norm 1	366	0.71
Norm 2	374	0.71