INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

University of Alberta

ASSESSING THE EFFECT OF THE MULTIDIMENSIONALITY-BASED DIF ANALYSIS PARADIGM ON THE COMMON-ITEM NONEQUIVALENT GROUP EQUATING DESIGN USING TRANSLATED TESTS

by

YUEN YEE LI 

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

in

Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta

Fall 2005

DEDICATION

I dedicate this work

to the memory of my grandmother Chun Ng-Li for your beautiful spirit,

to my parents Chu Koon Li and Pik Chun Leung, my dear brothers Shun Lap,

Shun Tak, Shun Pong, Shun Cheong for your unconditional love and faith in me

while I had to be away from home for such a long period of time, and

to my beloved husband Patrick Baltazart for your infinitive support and

being denied some of our precious time together.

Abstract

The primary purpose of this study was to evaluate the effect of using the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996) to select the common items to be used for a common-item nonequivalent group equating design to equate translated achievement tests. Three different DIF conditions were created for comparisons: No DIF, Exploratory DIF, and Confirmatory DIF. Four equating methods were used: Tucker linear equating, Levine observed score equating, equipercentile equating, and item response theory observed score equating. Six achievement tests in three different languages and four subject areas were analyzed. Evaluations were focused on the common-item sets, the equated scores of the translated tests, and the re-classification of examinees who wrote the translated tests and met the standards.

The results revealed that the exploratory DIF condition tended to identify more DIF items than the confirmatory DIF condition, which lead to the development of different sets of common items for these two DIF conditions in five of the six tests. All the common-item sets for the six tests fulfilled the evaluation criterion for equating. The equating results revealed that there were important changes in the equated scores in four tests and that these scores affected the re-classification of examinees in three of the four tests, regardless of whether the tests were translated from English into French or Chinese. Differences in the equating results were found between the DIF and No DIF conditions, between the exploratory and confirmatory DIF conditions, between different subject areas, and between different equating methods. Two of the six tests showed no differences in the equated scores and the re-classification of examinees in all DIF conditions and across all equating methods.

To conclude, the multidimensionality-based DIF analysis paradigm is a suitable approach for DIF detection and equating translated tests. It processes two advantages over the other DIF procedures. First, the paradigm tends to identify more common items than the traditional statistical DIF analyses, which enhances the content and statistical representativeness of the common-item sets for equating. Second, the paradigm helps to identify the sources of DIF which enhances the interpretation of the equated translated tests scores in contrast to the traditional statistical DIF analyses. Discussion and implications of the present research have been included.

Acknowledgements

I would like to thank each of the following people whose help and guidance has made the completion of this dissertation possible.

To Dr. Mark Gierl, my dissertation supervisor, for your enthusiasm, excellence, and insightful vision to maintain research in a high standard. I am grateful to have you as my supervisor and mentor throughout the program. You have provided me the guidance whenever I needed.

To Dr. Todd Rogers, for your perfectionism, infinitive patient, and kindness. I appreciated your generosity to accept students as your own family members. My husband and I will never forget your unconditional supports that help us to overcome many hardships in both academic and career developments.

To Dr. Ollie Triska, for your wisdom, invaluable comments, and support. I am grateful to have you broaden my insights in applied measurement and evaluation in medicine and public health areas.

To Dr. Judy A. Cameron and Dr. Leonard L. Stewin, for your interest and insightful suggestions as members of my dissertation committee.

To all the CRAMErs, for your friendships and encouragements throughout the program. I am thankful to have a wonderful experience in CRAME.

Table of Contents

List of Tables

CHAPTER I INTRODUCTION

Educational and psychological tests are frequently translated to allow comparable multi-cultural and multilingual testing. Translated tests can facilitate comparative studies across national, ethnic, and cultural groups both at the national and international levels (Cook, Schmitt-Cascallar & Brown, 2005; Feuer & Fulton, 1994; Hambleton & Kanjee, 1995; Hambleton & Patsula, 1998; International Association for the Evaluation of Educational Achievement, 1994). Making comparisons about student achievement across different language groups using different language forms may range from easy to difficult. When items are perfectly translated from the source language to the target language, it is assumed that different language forms are equivalent and comparisons are straightforward. However, translation is rarely perfect (Allalouf, Hambleton, & Sireci, 1999; Choi & McCall, 2002, Ercikan, 1999, Ercikan, Gierl, McCreith, Puhan, & Koh, 2002; Gierl & Khaliq, 2001; Gierl, Rogers & Klinger, 1999; Rogers, Gierl, Tardif, Lin, & Rinaldi, 2003; Sireci, 1997; van de Vijver & Tanzer, 1997; Wainer, 1999). Therefore, it is often a challenge to compare student achievement across different language groups. This chapter provides an overview of how to improve the comparison of student achievement across different language groups by identifying and describing issues surrounding translation errors, differential item functioning, and equating. This chapter is concluded by stating the purpose of the present research.

Items on translated tests may not have equal difficulty due to the unintended effects of translation (Sireci, 1997; Wainer, 1999). For example, an English mathematics item concerning "the length of time between 11am to 2pm" is translated to French as "le temps qu'il faut de 11h00 à 14h00". This example illustrates the English-French cultural

difference between the English item with a 12-hour clock using AM and PM and the French translation that uses a 24-hour clock (Gierl & Khaliq, 2001). As a result, the French item may be less difficult than the English item because the time metric differs in the two language groups. That is, the French students need only to take the 24-hour clock time difference while the English students need to understand the AM-PM time difference. Therefore, it is important to evaluate the equivalence of items on the translated tests before making comparisons between different language groups.

Differential item functioning (DIF) is often used to evaluate item equivalence on different language forms (Hambleton, 1994). DIF is present in an item when examinees from different groups have different probabilities or likelihoods of answering an item correctly after conditioning on ability (Shepard, Camilli, & Averill, 1981). Continuing with the time difference mathematics item, if the French students have a higher probability of answering the item correctly than the English students independent of the overall performance on the test, then this item is considered a DIF item (Gierl & Khaliq, 2001).

Statistical DIF analyses are used commonly on national and international assessments and many DIF items have been detected on translated achievement tests (Allalouf et al., 1999; Ercikan, 1999, Ercikan et al., 2002; Gierl & Khaliq, 2001; Gierl et al, 1999; Puhan, 2003). Therefore, translated tests cannot be assumed to be equivalent across different language groups. It is a challenge to compare student achievement across different language groups as the performance differences may be due to test differences (different language forms of tests and incorrect translations) or group differences (ability which the test is intended to measure) (Hambleton, 1994; Sireci, 1997).

To help disentangle test differences from group differences, the scores of the two groups need to be placed on a common scale. When the scores from both groups are placed on the same scale, test differences between the two groups are removed and, thus, performance differences between the groups can be attributed to group differences. Rather than making the assumption of item equivalence between different language forms, a statistical process is required to place different language forms on to a common scale. This process is called equating.

Equating can be used to place two groups of examinees who write different language test forms on to a common scale. Equating is a statistical process that is used to adjust scores on test forms so that the adjusted scores on the forms can be used interchangeably (Kolen & Brennan, 2004, p.2). Thus, equating translated tests may enhance the interpretability of comparisons between different language groups. A common-item nonequivalent group equating design can be used to equate different language forms (Kolen & Brennan, 2004; Sireci, 1997). This equating design is used when two forms of a test have a set of items in common and different groups of examinees are administered the two forms. The two groups of examinees are considered to be nonequivalent but the common items should behave similarly in both test forms (Kolen & Brennan, 2004, pp.18-19). Therefore, item equivalence across different language forms needs to be evaluated and only the items considered invariant across the language groups should be used to form the common set of items for equating tests (Sireci, 1997).

Traditionally, the outcomes of statistical DIF analyses have been used to identify the common items for equating across different language forms (Angoff & Cook, 1988;

Choi & McMall, 2002; Marais & Gierl, 2002; Rapp & Allalouf, 2002, 2003). Items identified with DIF reveal that student performance differs between the language groups, and DIF items are considered to be variable across the groups. Alternatively, items that do not display DIF (non-DIF items) are considered to be invariant across the groups. After excluding the DIF items, the non-DIF items are used as the common items. Therefore, statistical DIF analyses can be used to identify non-DIF items which can then be used as common items in equating.

However, before the removal of the items displaying DIF, it is necessary to determine why there is DIF. Items flagged as DIF may be due to poor translation (item bias) or to differences between groups (item impact) (Hambleton, 1994). There are two possible outcomes. First, when DIF items are attributed either to poor item translation or to ability differences between groups, using the outcome of statistical DIF analyses to identify the common items for equating tests may result in the exclusion of some useful items namely DIF items due to group ability differences. Second, when all of the DIF items are attributed to group ability differences and the translated test does not contain translation errors, the translated test is comparable to the original test. As different language test forms are already on a common scale, equating is not necessary. However, statistical DIF analyses cannot be used to explain why items are functioning differentially between the groups. As a result, the use of outcomes from statistical DIF analyses in selecting common items without identifying the source of DIF may adversely affect the results of equating. The interpretability of the equating results using translated tests is often related to the selection of the common-item set. Thus, it is important to identify the sources of DIF when selecting common items.

Substantive DIF analyses that involve the judgmental review process to interpret the nature of the DIF items have been used to identify the sources of DIF (Hambleton, 1994; Puhan, 2003). However, the use of substantive analyses to explain the statistically identified DIF items has generally not been successful at helping researchers and practitioners understand why DIF occurs (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Camilli & Shepard, 1994; Engelhard, Hansche, & Rutledge, 1990; Gierl, Bisanz, Bisanz, & Boughton, 2003; Gierl & Khaliq, 2001; Hambleton & Jones, 1994; Plake, 1980; Roussos & Stout, 1996). That is, once the DIF items have been flagged statistically, it is difficult to specify the reasons for the differential performance between groups (AERA et al., 1999). Without identifying all the sources of DIF, the use of the traditional DIF analyses outcomes may affect the selection of the common items and the interpretability of equating results with translated tests.

In an attempt to overcome this problem, Roussos and Stout (1996) proposed the multidimensionality-based DIF analysis paradigm to bridge the gap between the substantive and statistical DIF analyses. It is a two-stage confirmatory approach: the substantive analysis is used to generate DIF hypotheses and the statistical analysis is used to test the DIF hypotheses. Gierl (2005) noted three strengths of the multidimensionality-based DIF analysis paradigm: (1) it is guided by a multidimensional model for understanding how DIF occurs, (2) it provides better Type I error control than single-item statistical DIF analyses, and (3) it can be used to evaluate single items and bundles of items. In some cases, the multidimensionality-based DIF analysis can also be used to understand why DIF occurs (Bolt & Stout, 1996; Gierl et al., 2003; Gierl & Bolt, 2003;

Gierl & Khaliq, 2001). However, no study has been conducted in which the outcomes from the multidimensionality-based DIF analysis paradigm are used to select the common items to be used to equate translated tests and to evaluate the interpretability of the equated translated test scores.

Purpose of the Present Research

Thus, the purpose of the present research was to evaluate the effect of using the multidimensionality-based DIF analysis paradigm to select the common items to be used in the common-item nonequivalent group equating design with translated achievement tests. Three DIF conditions were created for comparison: No DIF, the Exploratory DIF condition (the use of statistical DIF analyses), and the Confirmatory DIF condition (the use of the multidimensionality-based DIF analysis paradigm). Four equating methods were used: Tucker linear equating, Levine observed score equating, equipercentile equating, and item response theory observed score equating. Six data sets in four subjects areas with either French or Chinese translated test forms were included.

Three research questions were addressed in the present research:

1. By comparing the DIF (exploratory and confirmatory) and the No DIF conditions, are there any differences between the equated and non-equated scores? Do these scores affect the classification of examinees in meeting the standards of the translated tests?
2. By comparing the equating results from the exploratory and confirmatory DIF conditions, are there any differences in the common-item sets, equated scores, and the classification of examinees?

3. By comparing the equating results across different languages of translation and subject areas, are there differences in the common-item sets, equated scores, and the classification of examinees?

Three characteristics distinguished the present research from previous studies. First, the use of three different DIF conditions provides an evaluation of whether the presence of DIF items affects the equating of translated tests. Second, this research uses the outcomes of the multidimensionality-based DIF analysis paradigm (confirmatory DIF condition) to select common items when equating the translated tests whereas previous studies only used the outcome of statistical DIF analyses (exploratory DIF). The comparisons of the exploratory and confirmatory DIF conditions provide an evaluation of the use of the corresponding outcomes and their effects on common items selection and the equating results interpretation. Third, this research was conducted using data from achievement tests which were translated from English to French (Social Studies and Mathematics) and English to Chinese (Economics and Physics). Thus, the effects of using the multidimensionality-based DIF analysis paradigm for equating translated tests were evaluated in different languages and different subject areas.

Definition of Terms

- *Common-item nonequivalent group equating design:* Equating design when two forms of a test have a set of items in common, and different groups of examinees are administered the two forms. The two groups of examinees are considered to be nonequivalent but the common items should behave similarly in both test forms (Kolen & Brennan, 2004, pp.18-19).

- *Confirmatory DIF condition:* The use of the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996) to detect DIF items in the present research. In the first stage substantive analyses, the four sources of translation DIF described by Gierl and Khaliq (2001) were applied to interpret the items and generate the DIF hypotheses. Then, in the second stage statistical analyses, SIBTEST (Shealy & Stout, 1993) was used to test the DIF hypotheses.
- *Differential item functioning (DIF):* DIF is present in an item when examinees from different groups have a different probability or likelihood of answering an item correctly, after conditioning for ability (Shepard et al., 1981).
- *Equating:* Statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen & Brennan, 2004, p.2).
- *Exploratory DIF condition:* The use of statistical DIF analyses SIBTEST (Shealy & Stout, 1993) to identify DIF items.
- *Item bias:* Invalidity or systematic error in how a test item measures a construct for the members of a particular group. When a test item unfairly favors one group of examinees compared to another, the item is biased (Camilli & Shepard, 1994, p.8).
- *Item impact:* Constant item differences for the members of a particular group where these performance differences reflect actual knowledge and experience differences that the test is intended to measure (Clauser & Mazor, 1998).
- *Multidimensionality-based DIF analysis:* A two-stage confirmatory approach that is based on the Multidimensionality-based DIF analysis paradigm developed by Roussos and Stout (1996). The first stage substantive analysis is used to

generate DIF hypotheses and the second stage statistical analysis is used to test the DIF hypotheses.

- *No DIF condition*: A condition used in the present study that is based on two assumptions: the test was translated perfectly into different language forms, as in the ideal situation, or DIF items were ignored in the translated tests, and thus, items on different language forms were considered to be equivalent and the tests should display no translation DIF. Thus, the translated tests are assumed to be comparable to the original test without equating. The No DIF condition was used in the present research as a control condition.
- *Statistical DIF analyses:* Statistical procedure used to identify DIF items.
- *Substantive DIF analyses:* Judgmental review process used to interpret the sources of DIF items (i.e., reasons why DIF occurs).

CHAPTER II LITERATURE REVIEW

This chapter is organized into four sections. The first section provides an overview of the key issues that arise when tests are translated from one language to another. It includes explanations for the increase in the use of translated tests, concerns about comparisons of student achievement across different language groups, and sources of translation biases found on translated tests. The second section reviews different strategies designed to enhance comparisons of student achievement across different language groups. It includes the use of differential item functioning (DIF) to evaluate item equivalence, the use of equating to place different language test forms on to a common scale, and the use of traditional DIF analyses to select the common-item set to be used to equate translated tests. The third section reviews previous research on the use of traditional DIF analyses when equating translated achievement tests. This section provides a critical review of previous studies highlighting the limitations of the traditional approach to DIF detection. The fourth section introduces the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996). It includes an explanation of how four sources of translation DIF (Gierl & Khaliq, 2001) can be used in substantive analyses, and how SIBTEST (Shealy & Stout, 1993) can be used in statistical analyses. It provides an overview of how the multidimensionality-based DIF analysis paradigm was used in the present study.

Section I: Translated Tests

Increase Use of Translated Tests

Globalization of economics, tourism, migration, and related political changes have contributed to an increase in cross-cultural comparisons (Cook et al., 2005; van de

Vijver & Leung, 1997; van de Vijver & Looner, 1995). Educational and psychological tests are frequently adapted and translated to enhance multicultural and multilingual testing. There are four reasons to explain the increased use of translated tests (Choi, 1999; Cook et al., 2005; Feuer & Fulton, 1994; Hambleton, 1993; Hambleton and Kanjee, 1995; Hambleton & Patsula, 1998; IEA, 1994; Sireci, 1997). First, the use of translated tests can enhance fairness in assessment by allowing examinees to be assessed in their native language. Second, the use of translated tests can facilitate comparative studies across national, ethnic and cultural groups at the provincial, national, and international levels. Third, the use of translated tests helps reduce costs and save time relative to developing new tests in each language. Fourth, the use of translated tests is a requirement in some multilingual countries and cities (e.g., Canada and Hong Kong). Moreover, comparisons of student achievement across different language groups are popular in multilingual countries where translated achievement tests are used.

Student Achievement across Different Language Groups

For some countries with bilingual educational policies, it is necessary to administer educational achievement tests in both official languages. These tests are usually high-stakes educational achievement tests. For example, test scores for graduating high school students serve as an important outcome for university admissions or career placement decisions. Two examples are reviewed to demonstrate the challenges that arise when comparing student achievement across different language groups. These examples include the Alberta Education Provincial Achievement Testing program in Canada and the Hong Kong Certificate Education Examinations program in Hong Kong.

Alberta Education Achievement Testing Program

Canada is known throughout the world as a bilingual country. The practice of promoting two official languages, English and French, is enforced by the Official Languages Act enacted in 1969 (Churchill, 1998). To guarantee the educational rights of each language group, both English and French educational systems are offered in each province, and all the provincial and national exams are administered in both official languages (Churchill, 1998; Puhan, 2003).

In the Canadian province of Alberta, Alberta Education administers provincial achievement tests annually in four content areas (Language Arts, Mathematics, Science, and Social Studies) to students in Grades 6 and 9. Both English and French forms are provided for Mathematics, Science, and Social Studies, while in Language Arts there are separate tests for each language. The purposes of these achievement tests are to evaluate student performance and monitor the assessment standards set by the province over time (Alberta Education, 2003). Students are evaluated according to whether they achieve the "Acceptable Standard" or the "Standard of Excellence." Each year, it is expected that at least 85% of the examinees should achieve the Acceptable Standard and 15% of the examinees should achieve the Standard of Excellence.

Students' results on the annual provincial achievement tests have several implications (Alberta Education, 2003). First, students' performances on achievement tests reflect their cumulative growth and achievement. Second, school principals and school district authorities report achievement results to parents and other members of the communities, and submit annual reports to Alberta Education. Third, the results help

teachers and parents to encourage a continuous improvement process in schools that supports high standards of achievement across the province.

The tests are typically developed in English and then translated into French, except for Language Arts. The item translation process can be described in four steps. First, all the items on the achievement tests are developed in English and then translated into French by a translator with reference to the Program of Studies (Alberta Education, 1989, 1996) and approved textbooks for the grade level. Second, a committee with at least two French teachers and a bilingual test developer validates the French version test. The committee reviews the test items and provides comments and feedback to the translator and test developer. Third, the test developer makes decisions about suggested changes and the translator modifies the French test form. Fourth, the test developer reviews the revised form and, if acceptable, approves the French form for use.

Concerns have been raised about the accuracy of the Alberta Education test translation process. Researchers have reported that a large number of DIF items have been found on the Social Studies and Mathematics Achievement tests across the English and French examinee groups (Gierl & Khaliq, 2001; Gierl et al., 1999). However, these performance differences may be due to differences in item difficulty (item bias), real group ability (item impact), or a combination of both factors (Hambleton, 1994; Sireci, 1997). Because item equivalence between the English and French test forms is questionable, it is possible that the English and French groups may have different cut-scores for the Acceptable Standard and the Standard of Excellence. If the cut-scores differ across language groups, comparisons between the English and French examinees are compromised.

Hong Kong Certificate of Education Examinations (HKCEE) Program

Another example of a high-stakes achievement test that requires different language forms is the Hong Kong Certificate of Education Examinations (HKCEE). The HKCEE offers 42 content-specific tests. Apart from the language-related content areas, all tests are offered in both English and Chinese. These tests are administered annually at the end of the Secondary 5 academic year (Grade 12). The purpose of the HKCEE is to serve as a determinant for admission to the Secondary 6 and a basic qualification for employment. The results of the tests are reported in six grades from A to F (E is regarded as the minimum passing grade), and all students are then evaluated according to the grade they received. For example, Grade C or better is recognized as equivalent to an O-level pass on an overseas GCE O-level examination conducted by the UK Examination Board (Choi, 1999).

Students' results on the HKCEE have important implications (Choi, 1999). For instance, the test scores serve as the determinant for admission to the Secondary 6 and a basic qualification for employment. Also, in the Chinese culture, academic credentials are considered to be superior to other qualifications. Therefore, parents and school officials are highly concerned about student performance.

The tests are usually developed in English and then translated into Chinese except for the language-related content areas. Unfortunately, the translation process is not well documented. However, Choi (1999) stated that items in both language test forms have identical content, the scripts are marked using the same marking schemes, and the results are graded by the same cut-scores for standard setting. All the HKCEE exam papers are also sent to the UK Examination Board for moderation to maintain standards with the

GCE O-level exam. A sample of scripts is sent to the Board for marking and grading in accordance with the UK marking and grading standards. This policy is to ensure that Grade C or better (HKCEE) is equivalent to an O-level pass in the GCE O-level examination conducted by the UK Examination Board (Choi, 1999).

Concerns have been raised about the accuracy of the HKCEE test translation for two reasons. First, samples of the English test forms are sent to the UK Examination Board for marking and grading in accordance with the UK standard. However, only the English test forms are evaluated by this arrangement. This may imply that the English and Chinese test forms are equivalent, but no evidence is provided to substantiate this assumption. Second, Gierl, Cheng, Rogers, Gotzmann and Vandenberge (2000) reported that a large number of DIF items were found on the Computer Studies, Economics, Geography, History, Mathematics, and Physics tests across the English and Chinese examinee groups. These performance differences between the two language groups may be due to differences in test difficulty, real group ability, or a combination of both factors (Hambleton, 1994; Sireci, 1997). Because the item equivalency between the two language test forms is questionable, it is possible that the English cut-scores for setting the grades from A to F may not represent the same standards when the same cut-scores are applied to the Chinese groups.

Therefore, the Alberta Education Provincial Achievement Testing program and the HKCEE face many challenges that need to be satisfactorily addressed before comparisons are made between student performances across language groups. The present research uses data from the Alberta Education achievement tests and the HKCEE achievement tests.

Translation Bias on Achievement Tests

As indicated for the Alberta Education and the HKCEE achievement tests, concerns have been raised about the accuracy of the test translation. When items are translated perfectly from the source language to the target language, the different language forms are equivalent and group comparisons are straightforward. However, translation is rarely perfect and translation errors often occur (Allalouf et al., 1999; Choi & McCall, 2002, Ercikan, 1999; Gierl & Khaliq, 2001; Gierl et al., 1999; Puhan, 2003; Sireci, 1997; van de Vijver & Tanzer, 1997; Wainer, 1999).

Items on translated tests may not have equal difficulty levels due to unintended effects of translation (Sireci, 1997; Wainer, 1999). Hambleton and Patsula (1999) provide an example from an international comparative study of reading achievement. The English version of the test item is "Determine whether these two words are similar or different: Pessimistic versus sanguine" and about 54 % of the American students correctly answered the item. In the foreign language test form, there is no equivalent word for "sanguine" and another word was chosen that had a different meaning than pessimistic. However, when the item was translated back into English, the back translated form was "pessimistic versus optimistic." Almost 100 % of the examinees in the foreign language form answered the item correctly. This example illustrates that translation differences can affect the interpretation of the test score in a multilingual comparison.

To understand translation errors, van de Vijver and Tanzer (1997) identified three kinds of bias that could affect translated tests: construct bias, method bias, and item bias. Construct bias occurs when the construct measured is not identical across cultural groups

(van de Vijver & Tanzer, 1997). That is, the same construct is interpreted and understood in a different way in two language or cultural groups (Hambleton, 1994). For example, while many Western intelligence tests measure speed in problem solving as one important feature of intelligence, speed is not crucial in the oriental culture because depth of knowledge is considered a better measure of intelligence than speed of processing (Hambleton & Patsula, 1999; Lonner, 1990; Puhan, 2003). As a result, members of the oriental culture group often score lower on Western intelligence tests because of their failure to perform quickly (Hambleton & Patsula, 1999).

Method bias refers to problems arising from the method employed in empirical studies such as the presence of sampling, instrument, and administration bias (van de Vijver & Tanzer, 1997). Sampling bias refers to sample differences that may confound group comparisons because the groups are not equivalent. For example, some groups who are frequently exposed to psychological tests show less motivation than the groups for whom the instrument or the test situation is highly novel (van de Vijver & Tanzer, 1997). Instrument bias refers to stimulus familiarity and response style. An example of stimulus familiarity can be found in the items from the Information subtest in the WISC (van de Vijver & Poortinga, 1997). Exposure of children to the items in the Information subtest differ across cultures and this difference influences the test results. An example of a response style difference is that the Hispanic examinees tend to choose extreme points (1 or 5) on a five-point Likert scale more often than the White examinees (Hui & Triandis, 1989). Thus, performance differences between the Hispanics and the White Americans can be confounded by response style. Administration bias occurs when tests include item formats (e.g., multiple choice, essay) and the use of graphics and

presentation modes (paper-and-pencil, computer) that are less familiar to examinees in one group than the examinees in another group (Hambleton & Patsula, 1998; Puhan, 2003).

Item bias refers to an item that is invalid or results in systematic error in what it measures for members of a particular group but not for another group (Camilli & Shepard, 1994). Item bias includes poor item translation, nuisance factors, and cultural specifics (Gierl & Khaliq, 2001; Hambleton, 1994; Puhan, 2003; van de Vijver & Tanzer, 1997). Poor item translation may occur when the concepts, expressions, and ideas used in the source language do not have equivalents in the target language (Gierl & Khaliq, 2001). For example, the word "rollerblader" in the English phrase "most rollerbladers favor a helmet bylaw" does not have an equivalent word in the French, and thus, it is difficult to translate this phrase from English to French so as to maintain the same meaning (Gierl & Khaliq, 2001).

A translated item may contain a nuisance factor that provides an instance of construct-irrelevant easiness by providing extraneous cues in the item that are irrelevant to the construct being measured for one language group but not another language group. Consequently, the item will be relatively easy for one language group compared to another language group (Hambleton, 1994; Messick, 1989). As a result, the translated item has different item difficulties across different language forms (Angoff & Cook, 1988; Sireci, 1997; Wainer, 1999). Hambleton (1994) provides an example of an English item translated into Swedish. In the item "Where is a bird with webbed feet most likely to live?" there is no direct translation of the phrase "webbed feet" in the Swedish translation. Thus, the Swedish item translated into English was equivalent to "swimming

feet". As such, the translation provides a clue to the Swedish examinees, which leads to the construct-irrelevant easiness in the Swedish form of the item.

Item bias is culturally specific with respect to cultural distance and subject area (Rogers, 2002, cited in Puhan, 2003; van de Vijver & Tanzer, 1997; Wainer, 1999). The greater the cultural distance between language and ethnic groups, the more adverse the effects may be on test equivalence (Rogers, 2002, cited in Puhan, 2003). For instance, translated tests may show more comparability across similar language groups such as French-Spanish groups than the English-French groups (Puhan, 2003). Since French and Spanish have the same language root in Latin, French-Spanish groups are closer in their culture distance than English-French groups. Also, when making comparisons on translated tests across different subject areas, non-verbal items (e.g., Mathematics) do not change as dramatically as verbal items (e.g. Social Studies) (Wainer, 1999). In this case, the non-verbal items are less affected by culture and language during translation (Wainer, 1999). Therefore, the amount of item bias present on the translated test may vary according to different subject areas.

In sum, the presence of construct, method, and item biases may adversely affect the accuracy of translated tests. As translated achievement tests are often used to make comparisons between students in different language groups at the provincial, national and international levels, translated tests cannot simply be assumed to be equivalent (Cook et al., 2005; Feuer & Fulton, 1994; Hambleton, 1993, 1994; IEA, 1994; Sireci, 1997; Wainer, 1999; Puhan, 2003). It is important to evaluate item equivalence on translated tests before making comparisons across different language groups and to identify the source of nonequivalence.

Section II: Comparisons across Different Language Groups Using Translated Tests

The second section reviews different strategies for enhancing the comparisons of student achievement across different language groups. It includes the use of differential item functioning (DIF) to evaluate item equivalence, the use of equating to place different language test forms on to a common scale, and the use of traditional DIF analyses to select common items to be used when equating translated tests.

Differential Item Functioning on Translated Tests

Differential item functioning (DIF) is used to evaluate item equivalence on different language forms (Hambleton, 1994). DIF is present in an item when examinees from different groups have a different probability or likelihood of answering an item correctly after conditioning on ability (Shepard et al., 1981). DIF may occur because of item bias or item impact. Using the example of the time-differences mathematics item, the French group had a higher probability of answering the item correctly than the English group after conditioning on ability, and thus, this item is considered a DIF item. DIF occurs in this item because the translation provides construct-irrelevant easiness for the French group. This outcome is referred to as item bias. However, DIF may also occur in an item if one group learns more than the other language group. For example, students in one country may learn algebra at an earlier stage than students in another country because of differences in their education curriculums. Consequently, performance differences between the two groups on algebra items may occur due to curricular differences. This outcome is referred to as item impact (Clauser & Mazor, 1998).

Statistical DIF analyses are used commonly and many DIF items have been detected on translated achievement tests (Allalouf et al., 1999; Ercikan, 1999, Ercikan et al., 2002; Gierl & Khaliq, 2001; Gierl et al., 1999; Puhan, 2003). For example, Allalouf et al. (1999) reported that 39.6% (42 out of 125 items) of the verbal items on the Israeli Psychometric Entrance Test (Hebrew and Russian test forms) displayed DIF. Gierl et al. (1999) reported 55.1% (27 out of 49 items) of the items in a Canadian Social Studies achievement test and 12.0% (6 out of 50) of the items in a Canadian Mathematics achievement test (English and French test forms) displayed DIF. Ercikan (1999) reported that 41.4% (58 out of 140) science items and 18.4% (29 out of 158) mathematics displayed DIF on the Third International Mathematics and Science Study when comparing Canadian English and French examinees. Results from these studies reveal that translated tests display many DIF items, and that verbal items (e.g., Social studies) display more DIF than non-verbal items (e.g., Mathematics). Therefore, translated tests cannot be assumed to be equivalent across language groups or subject areas.

Often, it is a challenge to compare student achievement across different language groups as the performance differences may be due to test differences (different language test forms) or group differences (ability which the test is intended to measure) (Hambleton, 1994; Sireci, 1997). To help disentangle test differences from group differences, the two language test forms need to be placed on to a common scale. When items are equivalent in both language forms, test differences between the two groups are removed, and performance differences between the groups can be attributed to group differences. Rather than assuming that translated tests are equivalent, a statistical process

called equating, is required to place different language forms on to a common scale when comparing students' achievement across different language groups.

Equating Translated Tests

Equating can be used to place the two groups who write different language test forms on to a common scale (Angoff & Cook, 1988; Marais & Gierl, 2002; Rapp & Allalouf, 2002, 2003). Equating is a statistical process that is used to adjust scores on test forms so that the scores on the forms can be used interchangeably (Kolen & Brennan, 2004, p.2). There are statistical processes that are similar to equating. For example, the term "linking" has been used broadly to refer to a relationship between two tests' scores (Kolen, 2004; Linn, 1993; Mislevy, 1992). The two tests to be linked can be measures of the same construct, similar (but not identical) constructs, or different constructs, whereas two tests to be equated are measures of the same construct (Kolen, 2004; Linn, 1993).

Equating is used in the present research because the translated achievement tests are assumed to measure the same construct and the translated test scores are used interchangeably with the original test scores in standard setting and in university entrance or placement decisions (Cook & Schmitt-Cascallar, 2005; Linn, 1993; Lord, 1980). Thus, equating translated tests may enhance the interpretability of comparisons among students across different language groups. After equating different language tests, test differences between the two groups are removed, and performance differences between the groups can be attributed solely to group differences.

Sireci (1997) suggested three research designs that could be used to equate different language tests on to a common scale: separate monolingual group design,

bilingual group design, and matched monolingual group design. The separate monolingual group design is applied when the original and translated tests are separately administered to their respective language groups. The bilingual group design is used when a group of bilingual examinees assumed to be equally proficient in both languages are tested with the two different language forms. The matched monolingual group design uses separate monolingual groups which are matched on particular criteria that might affect the equating results (e.g., socioeconomic status, education).

Among these three research designs, the separate monolingual group design is used most commonly for different language tests (Angoff & Cook, 1988; Choi & McCall, 2002; Marais & Gierl, 2002; Rapp & Allalouf, 2002, 2003). Three advantages of the separate monolingual group design over the other designs are that the source and translated language test forms are administered separately to their respective language groups; the design does not need a bilingual group that is equally proficient in both languages; and different language groups do not need to be matched on any criteria. For the purpose of the present research, the separate monolingual group design is selected because examinees from different language groups only take one language form of a test, and the samples are not matched on any criteria.

The separate monolingual group design is a variation of the common-item nonequivalent groups design for same language equating (Angoff, 1971; Kolen & Brennan, 2004). The common-item nonequivalent groups equating design is used when two forms of a test have a set of items in common and different groups of examinees are administered the two forms. Kolen and Brennan (2004) suggested that the common items used in this equating design should behave similarly on both test forms. The common

items in same language equating are identical, whereas in different language test equating, the common items are chosen from the translated items. These items are treated as if they were identical, and they are assumed to measure the same construct and have the same psychometric characteristics (Rapp & Allalouf, 2003). The common-item nonequivalent group equating design was used in the present research.

There are two requirements for equating different language tests using the common-item nonequivalent groups designs: the different language tests measure the same construct and the translated items used as common items across different language tests are equivalent (i.e., they retain the same meaning and psychometric properties following translation) (Cook et al., 2005; Rapp & Allalouf, 2003; Sireci, 1997). When evaluating the requirements of construct equivalence and common-item equivalence, the last requirement is the most difficult to satisfy and evaluate in practice (Sireci, 1997).

Construct Equivalence

The International Testing Committee (ITC) was formed in 1992 to develop technical standards or guidelines for test translation. The ITC guideline C.2 states, "The amount of overlap in the constructs in the populations of interest should be assessed" (Hambleton, 2001). Therefore, before attempting to equate different language tests on to a common scale, it must be demonstrated that the constructs measured by the tests are comparable (Hambleton, 2001; Sireci, 1997). Construct equivalence across translated educational achievement tests can be evaluated using many theoretical and empirical methods (Cook et al., 2005; Poortinga, 1983, 1989; van de Vijver & Leung, 1997; van de Vijver & Tanzer, 1997) and achieved in practice (e.g., Gierl, 2000; Marais & Gierl, 2002; Zumbo, 2003). However, construct equivalence does not guarantee item equivalence

(Sireci, 1997; Zumbo, 2003). The data used in the present research included translated tests with the same test specifications and same test administration procedures as with the original tests. Moreover, as construct equivalences of these data have been evaluated in previous studies (e.g., Gierl, 2000; Marais & Gierl, 2002), the present research focuses on the evaluation of the common items.

Common-item Equivalence

Once the construct equivalence has been established, attention needs to be paid to common-item equivalence in equating. Four characteristics of common-item equivalence have been recommended (e.g., Kolen & Brennan, 2004; Sireci, 1997). These characteristics include the representativeness of the common-item set relative to the full-length test forms to be equated, the number of items to be included in the common-item set, the invariance of the common items between different groups, and the use of nonverbal items as common items.

Representativeness. The set of common items should represent the full-length (original) test forms in both content and statistical characteristics (Kolen & Brennan, 2004). The content characteristics are maintained when the common-item sets have the same test specification proportions as the full-length tests. The statistical characteristics are usually based on classical statistics such as mean, standard deviations, and distributions of item difficulties and discriminations for a particular group of examinees (Kolen & Brennan, 2004).

Researchers reported that the effectiveness of the common-item set depends on the content and statistical representativeness of the common-item sets, which are crucial when the two groups differ in ability (Cook & Peterson, 1987; Petersen, Marco, &

Steward, 1982). Klein and Jarjoura (1985) reported that a content representative common-item set functioned better than a longer, non-representative common-item set. In fact, the accuracy of equating depends on the content representativeness of the common-item set when using Tucker linear equating and IRT equating methods (Yang, 1997).

Other researchers have reported that the content and/or statistical representativeness are not important under some conditions (Beguin, 2002; Budescu, 1985; Harris, 1991; Rapp & Allalouf, 2002; Rogers, 2002, cited in Marais & Gierl, 2002). Rogers (2002, cited in Marais & Gierl, 2002) suggested that if the test forms contain a dominant factor, the need for the common-item set to mirror the test specifications becomes less severe, whereas the similarity of the statistical characteristics becomes more important. Harris (1991) found that content representativeness did not greatly influence equating results. However, if the common-item set was not statistically representative, a content representative common-item set may produce less equating error than a content non-representative common-item set.

Rapp and Allalouf (2002) studied the effect of using a non-representative common item set on multilingual equating. Two language versions of the Psychometric Entrance Test (PET) verbal domain subtest for admission to Israeli universities were used. Equated examinee scores determined using representative and non-representative common-items sets were compared. The results showed that score differences were about one-fifth of a standard deviation. Rapp and Allalouf (2002) concluded that the adverse effects of non-representative common-item set have been over-emphasized in the literature or that these effects are test or situation specific. Moreover, Budescu (1985)

and Beguin (2002) found that the unidimensional equating methods were fairly robust to the violations of the assumption of common-item representativeness, and that a high correlation between the common-item set and the full-length test was an important determinant of the accuracy of the equating process.

Length. The appropriate number of common items to be included in the common-item sets has been studied (Angoff, 1971; Budescu, 1985; Kolen & Brennan, 2004; Raju, Bode, Larsen, & Steinhaus; 1986; Skaggs, 1990; Skaggs & Lissitz; 1986; Wainer, 1999; Wingersky, Cook, & Eignor, 1987; Wingersky & Lord, 1984; Yang & Houang, 1996). Wingersky and Lord (1984) reported that the use of two common items with small standard errors worked as well as a set of 25 common items. Raju et al. (1986) found that a five-item common-item set was as effective as a longer common-item set. Skaggs (1990) reported that fewer than 10 common items were sufficient with the IRT equating method. Skaggs and Lissitz (1986) found that 15 items were adequate for the equipercentile equating.

Budescu (1985) and Wingersky, Cook, and Eignor (1987) reported that a larger number of common items led to less random equating error than a smaller number of items. However, when the number of common items increased beyond 20% of the total number of items in the test, the improvements on equating accuracy were not practically significant (Yang & Houang, 1996). Two rules of thumb have been suggested for determining the number of common items sufficient for equating purpose: A common-item set should be at least 20% of the full-length test (test with 40 or more items), or have 15 to 20 common items (Angoff, 1971; Kolen & Brennan, 2004; Wainer, 1999).

Invariance. Before equating the different language tests using the common-item nonequivalent group design, common items need to be selected. To do so, the translated items presented in the translated tests are evaluated for invariance across the different language tests (Sireci, 1997). Statistical DIF analyses are often used to identify the items that are functioning differently in the two language groups (Angoff & Cook, 1988; Harris, 1993; Rapp & Allalouf, 2003; Sireci, 1997). Items that do not display DIF are considered to be invariant across the language groups. Sireci (1997) suggested that the items to be used in the common-item sets should be selected from items considered invariant across different language tests. Therefore, the common items used in the equating process are translated items selected according to the outcome of statistical DIF analyses, meaning the invariant items contain psychometric characteristics that have not significantly changed following translation (Allalouf et al., 1999; Rapp & Allalouf, 2003; Sireci, 1997).

The ITC guideline D.10 states, "Nonequivalent items between versions intended for different populations should not be used in preparing a common scale or in comparing these populations" (Hambleton, 2001). Typically, DIF items that are not statistically equivalent are considered to be nonequivalent items. DIF items should not be used in equating. However, these items may be considered unique to the different language forms for the assessment of the construct of interest and, therefore, may be retained to increase test reliability (Hambleton, 2001; Sireci, 1997). Moreover, DIF items attributed to item impact and not item bias may be used as common items (this point is discussed in more detail in the next section).

Nonverbal items as common items. Nonverbal items, or items minimally associated with linguistic content, provide a theoretically appealing source of invariant

common items (Sireci, 1997, Wainer, 1999). For example, a mathematics item such as "3 + 5 x 22 =?" is considered as a non-verbal item. The assumption that the difficulty of a translated item does not change radically may hold for nonverbal items (e.g., mathematics) but not necessarily for verbal items (Wainer, 1999). Therefore, the equivalence of nonverbal items across languages is likely to be defensible irrespective of a statistical evaluation (Sireci, 1997).

However, in educational achievement tests, it is difficult to have items free of linguistic elements. Many educational achievement tests measure verbal skills that cannot be measured in a manner that is independent of linguistic context. When nonverbal items are used in lieu of other items, unintended effects of translation may occur as with the verbal items, and changes in the item difficulty may be found after translation (Sireci, 1997). However, researchers reported that the non-verbal items (e.g., in the Mathematics achievement test) tended to display less DIF than the verbal items (e.g., Social Studies) (Allalouf et al., 1999; Ercikan, 1999; Gierl et al., 1999). Although it is difficult to find real nonverbal items in translated educational achievement tests, nonverbal items may be a better source of common items than the verbal items.

Traditional DIF Analyses in Selecting Common Items

To fulfill the common-item equivalence requirement for the common-item nonequivalent groups equating design, item equivalence across different languages forms needs to be evaluated, and only items considered invariant across the language groups should be used as the common items for equating tests (Sireci, 1997). The outcomes of statistical DIF analyses have been used to identify the common items across the language forms (Angoff & Cook, 1988; Choi & McMall, 2002; Marais & Gierl, 2002; Rapp &

Allalouf, 2002, 2003). Traditionally, the statistically identified DIF items are not used as common items and only the non-DIF items are to be included as the common items in equating translated tests.

However, statistical DIF analyses yield results cannot be used to explain *why* items are functioning differentially between the groups. DIF may occur because of item bias or item impact. Substantive DIF analyses, a judgmental review process used to determine why DIF items show DIF, can be used to identify the sources of DIF (Allalouf et al., 1999; Camilli & Shepard, 1994; Engelhard et al., 1990; Engelhard, Davis, & Hansche, 1999; Gierl & Khaliq, 2001; Hambleton, 1993, 1994; Puhan, 2003). DIF item may reflect item bias or item impact. The previous example of an English item for which the phrase "webbed feet" was translated as "swimming feet" in Swedish provided a clue to the Swedish examinees (Hambleton, 1994). This example reflects item bias and this item should be excluded from the common-item set. Alternatively, a mathematic item such as "3 + 5 x 22 =?" may be detected as a DIF item in different language groups. In this example, there is no translation bias in the item and performance differences between groups are attributed to real group differences. This is item impact and the item can be used as a common item in equating. Thus, it is important to interpret the source of the DIF.

The use of substantive analyses to interpret statistically identified DIF items has generally not been successful at providing explanations for why DIF occurs (AERA et al., 1999; Camilli & Shepard, 1994; Engelhard et al., 1990; Gierl et al., 2003; Gierl & Khaliq, 2001; Hambleton & Jones, 1994; Plake, 1980; Roussos & Stout, 1996). Once the DIF items have been flagged statistically, it is difficult to specify the reasons for the

differential performance between groups (AERA et al., 1999). However, without identifying the sources of DIF, it is difficult to interpret the statistically identified DIF items.

Using the outcomes from traditional DIF analyses (either statistical DIF analyses alone or statistical analyses followed by substantive analyses) may result in the exclusion of all DIF items from the common-item set and, thus, affect the interpretation of the equated scores on translated tests. Items may be flagged as DIF because of poor translation (bias) or because of actual group differences (impact) (Hambleton, 1994). DIF items that occur due to actual group difference are useful for common-item development and they should be retained on the test to enhance validity and reliability. However, the use of traditional DIF analyses to identify common items for equating may result in excluding some useful items (DIF items due to item impact) and thus, affect the interpretability of the results from equating. The difficulties encountered when using the outcome of traditional DIF analyses in equating are reviewed in the next section.

Section III: Traditional DIF Analyses and Equating

The success of equating requires construct and common-item equivalences between the different language forms of a test. However, it is often the second requirement that is difficult to achieve (Sireci, 1997). Since DIF analyses have been used to identify common items for equating, it is important to evaluate the consequences of using the outcome from traditional DIF analyses. Previous studies have investigated the use of traditional DIF analyses in equating same language forms and different language forms. Given the importance of these studies for the present research, it is necessary to review these studies.

Equating Same Language Tests

Two studies have been conducted to evaluate the effect of DIF analyses on equating same language test forms (Hanson & Feinstein, 1997; Zhang, Matthews-Lopez & Dorans, 2003). Unfortunately, the results from these studies are inconclusive. Zhang et al. (2003) studied the effect of deleting the DIF items on the reported scores by comparing the original scores with all items included and the re-equated scores after the DIF items had been deleted. Data were obtained from the Reasoning section of the Scholastic Assessment Test (SAT) and the section contained 78 items. Examinees (N=9,517) were classified into ten subgroups: African American Females, African American Males, Asian Females, Asian Males, Hispanic Females, Hispanic Males, White Females, White Males, All Other Females, and All Other Males.

The standardization method for DIF detection (Dorans & Kulick, 1986) was used and three DIF items (out of 78 items) were identified. The DIF items were deleted and the test scores from the 75 item test were re-equated on to the original score scale using equipercentile equating. The mean scores of each subgroup after equating were calculated. Zhang et al. (2003) reported that after the deletion of the DIF items, the mean test scores of the subgroups disadvantaged by the DIF items increased, whereas the mean test scores of the advantaged subgroup decreased after equating. However, the mean score differences amounted to less than one scale-point on a 20 to 80 point scale.

Results from the Zhang et al. (2003) study revealed that the presence of DIF items in the test affect the equated scores, and the impact varied in different subgroups. However, the study was limited by the small sample size (n = 230) for some of the ethnic groups (disadvantaged group). Consequently, the DIF result may be unstable. Moreover,

only statistical DIF analyses were used in the study and the source of DIF in the three DIF items was neither known nor considered. These DIF items were simply deleted from the test. However, if the source of DIF can be identified by substantive analyses, then additional information is available to make the decision about what to do with the DIF items. DIF items attributed to item bias should be deleted and items attributed to item impact should be retained in the test. Therefore, it is important to identify the source of DIF using substantive analyses before re-equating.

Hanson and Feinstein (1997) evaluated the effect of DIF items present in the common-item set on the equated scores using the common-item nonequivalent groups equating design. Data were obtained from a 150-item multiple-choice test in 1991 (n=1,521) and 1993 (n=1,375). All items were dichotomously scored. The 1993 test form was equated on to the 1991 score scale using 38 common items. The two test forms were in the same subject area but additional information about the forms and the sample was not documented. Hanson and Feinstein (1997) developed a log-linear model to identify DIF items. Results revealed that 15 of the 38 common items displayed DIF. Substantive analyses were then used to interpret the sources of DIF. Two of the 15 items were found to have format differences between the two forms; there was no known source for the remaining 13 items. Hanson and Feinstein (1997) reported that after excluding the two DIF items from the common-item set, the scores of 23 out of 1,375 examinees using Tucker linear equating and 310 examinees using Levine observed score equating either increased or decreased by one score point.

Despite the large sample size (n=1,375) and substantive DIF analyses, Hanson and Feinstein (1997) only provided limited information about the test forms and

information about the sample characteristics (e.g., age or grade) was not available. The study was limited because the source of DIF for only two items was interpretable and the sources of DIF for the other DIF items could not be identified. The results of this study demonstrate that the use of traditional DIF analyses, statistical analyses followed by substantive analyses, is often not successful at explaining why DIF occurs. Moreover, Hanson and Feinstein (1997) and Zhang et al. (2003) only identified a small number of DIF items, and in both studies, the different subgroups used the same language test form.

Equating Different Language Tests

Five studies have been conducted using traditional DIF analyses to select common items to equate different language tests using the common-item nonequivalent group equating design (Angoff & Cook, 1988; Choi & McCall, 2002; Marais & Gierl, 2002; Rapp & Allalouf, 2002, 2003). Angoff and Cook (1988) established equivalent scores between the SAT (English form) and the Prueba de Aptitud Academica (PAA, Spanish form) using the common-item nonequivalent group design to enhance comparisons between United States and Puerto Rican students. Data were obtained from the Verbal and Mathematics sections of the SAT and PAA. Each test was developed in its own language, and there were no translated items from one form to the other. The SAT contained 110 Verbal and 62 Mathematics items in English whereas the PAA contained 105 Verbal and 62 Mathematics items in Spanish. The sample included 2,000 examinees.

As the SAT and PAA tests were not translated test forms of one other, three steps were used to develop the common-item set for equating: develop translated items from the original tests, administer the translated tests to a sample of examinees, and identify

DIF and non-DIF (common) items between different language tests. First, four translators were used: two for translating the Spanish items into English, and two for translating the English items into Spanish. Then the two sets of translators back translated each other's work into the original languages. The test developers reviewed all the items and items that were judged inadequate were dropped out. Consequently, the two language sets of items included 160 verbal and 100 mathematics items considered as comparable in both languages. Second, the tests were administered. The English items were administered to the SAT examinees and the Spanish items were administered to the PAA examinees. All the SAT and PAA examinees samples consisted of 2,000 cases but the sample size for each language group was not known. Third, statistical DIF analyses were used to identify the DIF items between the groups. The item response theory (IRT) method was used and the item characteristic curves and estimates of item parameters were compared for the two groups. The non-DIF items, which included 39 verbal and 25 mathematical translated items, formed the common-item set.

After the common-item set was established, IRT true score equating was used to equate the SAT and PAA using the common-item nonequivalent group equating design. Angoff and Cook (1988) reported that items in the PAA were easier than the SAT in both the Verbal and Mathematics tests sections.

Angoff and Cook's (1988) study provided an early example of using statistical DIF analyses to identify common items in equating different language tests using the common-item nonequivalent group equating design. However, there are two limitations in this study. First, the item parameter invariance properties of IRT may not hold over samples derived from different language examinee groups (Sireci, 1997; Wainer, 1993,

Cook et al., 2005). Thus, the use of IRT to identify DIF items in different language tests may affect the validity of the results obtained by Angoff and Cook. Alternative methods other than IRT may be needed for DIF detection as well as common items selection. Second, only statistical DIF analyses were used in the study and the source of DIF was neither known nor considered.

Marais and Gierl (2002) evaluated the impact of DIF items on the equating function with translated tests. Data were obtained from the Grade 9 Alberta Learning Provincial Achievement Test in Social Studies (55 items) and Mathematics (44 items) for French and English. Five thousand examinees wrote the English forms and 2,000 examinees wrote the French forms in each test. Statistical DIF analyses were conducted using SIBTEST (Shealy & Stout, 1993). Fifteen DIF items in Social Studies and four DIF items in Mathematics were found. DIF items were deleted from the test. The classical test theory (CTT) approach using *p*-values and the IRT method using the item parameter estimates were used to identify common items. Then, the French translated test scores were equated on to the English scale. Tucker linear and equipercentile methods were used with the common-item nonequivalent groups equating design under two situations: the use of the full-length tests and the shortened test with the deletion of all DIF items. Equated scores of the French examinees from these two situations were compared. Marais and Gierl reported that the CTT approach and the IRT method resulted in the selection of similar sets of common items, and that the presence of DIF items had no significant impact on the equated scores.

Marais and Gierl (2002) provided an example using both CTT and IRT methods to select common items, and applied the common-item nonequivalent groups equating

design for equating translated tests. The limitation of their study was that only statistical DIF analyses were used in the study and the source of DIF was neither known nor considered, as with Angoff and Cook (1988). Moreover, the deletion of statistical DIF items from the test without identifying the source of DIF resulted in a shortened test form that may affect the test reliability and content representativeness of the shortened test. Therefore, it is important to identify the source of DIF using substantive analyses before re-equating.

Rapp and Allalouf (2002) evaluated the effect of group ability differences and the use of a nonrepresentative common-item set on equating using multilingual tests. Data were obtained from the Israeli University Psychometric Entrance Test (PET), which is written in Hebrew and then translated into five target languages: English, French, Spanish, Russian, and Arabic. The Hebrew (n=3,733) and Arabic (n=3,706) Verbal test forms were used. The Verbal test was composed of 40 translated items and 20 non-translated items that were written in the target language (i.e., Hebrew or Arabic).

The Mantel-Haenszel (MH) DIF detection method (Holland & Thayer, 1988) was used to identify DIF items among the 40 translated items. Sixteen of the translated items were considered as the non-DIF items and were included in the common-item set. Because the common items were selected from only the translated items, a non-representative common-item set was used in the common-item nonequivalent groups equating design.

Levine observed score equating and Tucker linear equating methods were applied to equate the translated Arabic test on to the Hebrew score scale. Equated scores were compared between the examinees with similar ability and examinees with dissimilar

abilities from the two language groups. Rapp and Allalouf reported that relatively small differences were found between the ability groups and the results were not influenced by the representativeness of the common-item set. Larger differences were found between the Levine and Tucker equating methods.

Rapp and Allalouf (2002) used a multi-method approach in equating to demonstrate the effect of ability differences between the groups on equating using different language tests. As different equating methods have different assumptions, it is important to use multiple equating methods and compare the results when equating translated tests. However, there are two limitations in the study. First, only statistical DIF analyses were used and the source of DIF was neither known nor considered, as with Angoff and Cook (1988) and Marais and Gierl (2002). Without identifying the source of DIF, the DIF items were excluded from the common-item set. However, items identified as item impact may be retained in the test as common items if substantive analyses are used to explain why DIF occurs. Second, the study was limited by the use of only two linear equating methods (Levine and Tucker methods). Thus, the effect of using non-linear equating, such as equipercentile equating and IRT equating, is not known.

Rapp and Allalouf (2003), in a second study, evaluated the stability of equating in large-scale multilingual testing. Data were obtained from the PET across six language forms (Hebrew and five other languages). Each language form was composed of 40 translated items and 20 non-translated items. The sample included 7,000 examinees in the Hebrew form and 2,500 examinees in each of the five language forms.

The common-item set was composed of translated and non-translated items. First, the MH DIF detection method (Holland & Thayer, 1988) was used to identify DIF items

from the translated items between the Hebrew and each translated test form. The non-DIF items were included as part of the common items. Second, practical considerations regarding the test specifications were made to select common items from the non-translated items. Non-translated items represented the same content and measured the same psychological construct were considered as the common items. Consequently, the common-item sets contained non-DIF items and other content relevant items, and the number of common items varied from eight to 22 in different language forms.

The common-item nonequivalent group equating design was used to equate each translated test on to the Hebrew scale. A linear equating method was conducted but the equating procedure was not specified. Rapp and Allalouf reported that the average difference between the equating results indicated the overall instability of the equating process, which was about 1.9 score points at the extreme ends of the score scale and 1.2 score points at the medial score. They claimed that the results might be related to the use of an overly short or a non-representative common-item set to equate the groups that were dissimilar in ability.

Rapp and Allalouf (2003) provided an example of the difficulties in selecting common-items in practice when non-translated items were included in the test. Rather than solely using statistical DIF analyses to select the common-item set from the translated items, practical considerations were also used to select the common items from the non-translated items. However, this study might be limited by the inclusion of non-translated items in the common-item set and this use might have resulted in the instability of the equating process. Again, Rapp and Allalouf used statistical DIF analyses to screen common items and, thus, the results of these studies are inconclusive and difficult

to interpret (Angoff & Cook, 1988; Marais & Gierl, 2002; Rapp & Allalouf, 2002, 2003). It is necessary to evaluate the use of statistical and substantive DIF analyses to screen common items on the equating process using translated tests.

Choi and McCall (2002) evaluated the application of IRT equating in the common-item nonequivalent groups equating design to equate a translated Spanish mathematics test on to a English score scale. Data were obtained from the Grade 3 and Grade 5 Oregon Statewide Mathematics tests, English and Spanish forms. There were 40 items in the Grade 3 and 60 items in the Grade 5 Mathematics tests. The Grade 3 sample included 8,895 examinees for the English form and 314 examinees for the Spanish form, whereas the Grade 5 sample included 8,762 examinees for the English form and 308 examinees for the Spanish form.

Two sets of common items were developed using the statistical and substantive DIF analyses. First, statistical DIF analyses were used to identify the DIF items. The IRT likelihood ratio procedure (Thissen, Steinberg, & Wainer, 1993) was used to identify the DIF items and non-DIF items were selected as one common-item set. Second, substantive analyses were used to select another common-item set. A bilingual mathematics content specialist and a bilingual mathematics teacher were asked to choose items most amenable to clear translation and least dependent on language and culture across the two language groups. These items formed the second common-item set.

Two common-item sets were established in each data set. The Grade 3 Mathematics test included 28 common items from the statistical analyses and 17 common items from the substantive analyses. The Grade 5 test included 39 common items from the statistical analyses and 16 common items from the substantive analyses. The IRT

true score equating method was used to equate the translated Spanish test on to the English scale. Choi and McCall reported that despite the use of two different common-item sets, there was little difference on the overall equated scores or percentage of examinees meeting the standard for the Spanish group.

However, there were limitations in Choi and McCall (2002) study. For example, Mathematics tests are more amenable to translation than subject area tests with heavier verbal components. Thus, the researchers concluded that the results of their study did not necessarily generalize to other subject areas or other equating methods. Also, substantive analyses in this study involved the use of the experts to identify DIF items. However, they did not explain why DIF occurred. A larger panel of math curriculum experts and expert translators would be needed to identify the source of DIF for the DIF items. It is important to have the substantive analyses that are effective and successful to interpret the source of DIF in translation tests, which in turn will enhance the interpretability of the equating results.

In sum, the use of traditional DIF analyses, either statistical analyses alone or followed by substantive analyses, in selecting common items for equating translated tests has many limitations. Statistical analyses are useful to detect DIF items and identify common items. However, statistical analyses alone are limited because the sources of the DIF are often not identified once the items are flagged (AERA et al., 1999; Hambleton, 1994; Roussos & Stout, 1996).

Section IV: Multidimensionality-Based DIF analysis Paradigm

In view of the shortcomings associated with the traditional DIF analyses used to identify common items for equating, the multidimensionality-based DIF analysis

paradigm (Roussos & Stout, 1996) can be used to help bridge the gap between statistical and substantive DIF analyses in order to identify common items. This section provides an overview of how to apply the two-stage multidimensionality-based DIF analysis paradigm in the present research. It includes an explanation for adopting the four sources of translation DIF identified by Gierl and Khaliq (2001) for the substantive analyses as the first stage, and the choice of SIBTEST (Shealy & Stout, 1993) to statistically identify items in the second stage.

The Multidimensionality-Based DIF Analysis Paradigm

Roussos and Stout (1996) developed the multidimensionality-based DIF analysis paradigm. It provides a theoretical basis for understanding how DIF occurs. It is a two-stage confirmatory approach: the substantive analysis is used to generate DIF hypotheses and statistical analysis is used to test DIF hypotheses. By combining substantive and statistical analyses, the multidimensionality-based DIF analysis paradigm can provide a systematic way to identify and study the sources of DIF.

Gierl (2005) noted three strengths of the multidimensionality-based DIF analysis paradigm: (1) it is guided by a multidimensional model for understanding how DIF occurs, (2) it provides better Type I error control than single-item statistical DIF analyses, and allows researchers to identify the sources of DIF, and (3) it can be used to evaluate single items and bundles of items. Thus, the multidimensionality-based DIF paradigm is useful for identifying and interpreting DIF items (Bolt & Stout, 1996; Gierl et al., 2003; Gierl & Bolt, 2003; Gierl & Khaliq, 2001). However, no study has been conducted to apply the multidimensionality-based DIF paradigm to select common-items in equated translated tests and, therefore, the consequences of applying these analyses for equating

scores on translated tests are unknown. The following section explains the application of the multidimensionality-based DIF analysis paradigm in the present research using a two-stage confirmatory approach with substantive and statistical analyses.

Stage 1: Substantive Analyses

The first stage of the Roussos and Stout (1996) multidimensionality-based DIF analysis paradigm involves conducting a substantive analysis to generate the DIF hypotheses. A DIF hypothesis is a description of whether a particular combination of substantive characteristics will elicit to DIF. To understand the multidimensional nature of DIF, a dimension is defined as the substantive characteristic of an item that can affect the probability of a correct response (Shealy & Stout, 1993). The primary dimension is the main construct that the test is intended to measure. The secondary dimension represents other constructs the test may or may not be intended to measure.

To generate the DIF hypotheses, organizing principles are used to identify items believed to measure secondary dimensions with specific characteristics. Gierl, Bisanz, Bisanz, Boughton and Khaliq (2001) described four organizing principles used commonly in substantive analyses. These principles include content-related properties (e.g., content categories), psychological characteristics (e.g., particular problem-solving strategies), test specifications (e.g., categories in test development), or empirical outcomes (e.g. statistical outcomes). Among these principles, researchers have suggested different content-related organizing principles to guide the substantive analyses in identifying the sources of DIF on translated tests (Allalouf et al., 1999; Engelhard, et al., 1990; Engelhard et al., 1999; Gierl & Khaliq, 2001).

The organizing principle developed by Gierl and Khaliq (2001) is effective for interpreting DIF on translated tests. Data were obtained from the Grade 6 and Grade 9 Social Studies and Mathematics Achievement tests in English and French. There were 3,000 examinees in the English form and 2,115 examinees in the French form in each data set. Gierl and Khaliq used an expert committee composed of test translators, editors, analysts, and test developers to identify four sources of translation DIF. These four sources included (1) omissions or additions of words, phrases, (2) differences in words, expressions, or sentence structure of items, (3) differences in words, expressions, or sentence structure of items that is not inherent to language and /or culture, and (4) differences in punctuation, capitalization, item structure, typeface, and other formatting usages (Gierl & Khaliq, 2001). Each source is described below.

The first source is omissions or additions of words, phrases, or expressions that affect meaning and are likely to affect the performance of one group of examinees. For example, on an item with a contour relief map, the English form contained the phrase "cross section cut along a line" while the French form contains the phrase "une coupe transversale qui montre le relief". The idea of "relief" is excluded from the English form.

The second source is differences in words, expressions, or sentence structure of items that are inherent to the language and/or culture and are likely to affect the performance of one group of examinees. One example is to illustrate the language difference between an English sentence, "most rollerbladers favor a helmet bylaw " and the French translation, "La plupart des pesonnes qui ne font pas de patin a roulettes sont pour un reglement municipal en faveur du port du casque protecteur." The English expressions "rollerblader" and "helmet bylaw" have no directly parallel French

equivalent. Another example to illustrate the cultural difference includes an English item with a 12-hour clock using AM and PM while the French translation uses a 24-hour clock.

The third source is differences in words, expressions, or sentence structure of items that are not inherent to language and/or culture and are likely to affect the performance of one group of examinees. For example, the phrase in English "basic needs met" versus the phrase in French "les services offerts" focus on "needs" in English and "services" in French. Hence, this example is categorized as differences not inherent to language and/or culture.

The fourth source is differences in punctuation, capitalization, item structure, typeface, and other formatting usage and is likely to affect the performance of one group of examinees. For example, an item contained a title in capital letters in one test form but not in the other test form is a difference in typeface. If these differences provide a clue to the correct answer for one group of examinees, then the item may not be comparable across language groups.

These four sources of translation errors reveal that differences in words, expressions, ideas, and format across different languages or cultures can affect the performance for one group of examinees. Using this organizing principle for the substantive analysis, sources of DIF items in translated tests can be interpreted (Gierl & Khaliq, 2001). However, the outcomes of this substantive analyses have not been used to select common items for equating. The present research applied the organizing principle developed by Gierl and Khaliq (2001) to generate DIF hypotheses and identify the common items for equating tests written in different languages.

Gierl and Khaliq (2001) provide the following example. The phrase in English mathematics item "the length of time between 11am to 2pm" was translated to French as "le temps qu'il faut de 11h00 à 14h00". Gierl and Khaliq's (2001) organizing principle was used to generate DIF hypotheses as the first stage substantive analyses: A hypothesis is generated that DIF would be found in the item in favor of the French examinees due to the differences in words, expressions, or sentence structure of items that are inherent to the culture (i.e., the second source of translation DIF). This example illustrates the English-French cultural difference includes an English item with a 12-hour clock using AM and PM while the French translation uses a 24-hour clock. Moreover, the organizing principle can also be used to generate DIF hypotheses for a single item as well as bundles of items.

Stage 2: Statistical Analyses

The second stage in the multidimensionality-based DIF analysis paradigm involves statistically testing the DIF hypotheses. It is a confirmatory statistical test of each DIF hypothesis resulting from the substantive analysis in the first stage. For example, the second stage statistical analyses are used to test the DIF hypothesis for the English mathematics item involving time. If the statistical analyses results reveal that the DIF hypothesis is tenable, then it follows that the item is sensitive to the second source of translation DIF and that the item favours the French examinees.

The most commonly used statistical procedures for detecting DIF include the Mantel-Haenszel (MH) (Mantel & Haenszel, 1959; Holland & Thayer, 1988), Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993), Logistic Regression (LR) (Swaminathan & Rogers, 1990), and the item response theory methods (IRT)

(Hambleton, Swaminathan & Rogers, 1991; Lord, 1980). Among these procedures, SIBTEST is effective in DIF detection and it provides three advantages over the other procedures.

First, SIBTEST was developed using the multidimensional model of DIF (Shealy & Stout, 1993). An item or a bundle of items is flagged as DIF if it is measuring a secondary dimension in addition to the primary dimension. SIBTEST can also be used to statistically test DIF hypotheses and quantify the size of DIF item. The advantage of having an effect size measure can be of great importance where Type I error inflation can pose serious threats to the validity of results derived from a statistical test. Second, SIBTEST can detect DIF for a bundle of items simultaneously; in contrast, MH, LR, and IRT test only one item at a time. Consequently, the bundle analyses leads to a decrease in Type I error rates. Third, SIBTEST tends to identify more DIF items than either MH or LR, hence it is a more liberal test (e.g., Ercikan et al., 2002; Gierl et al., 1999; Roussos & Stout, 1996; Shealy & Stout, 1993). A more thorough analysis of the test items will result when more DIF items are identified, and this result may lead to a more comprehensive evaluation of the test. Therefore, SIBTEST was selected to identify the DIF items in the second stage of the multidimensionality-based DIF analysis.

The statistical hypothesis tested by SIBTEST is

$$H_0: B(T) = P_R(T) - P_F(T) = 0,$$

Versus

$$H_1: B(T) = P_R(T) - P_F(T) \neq 0,$$

where $B(T)$ is the difference in probabilities of correct response on the studied item for examinees from the reference and focal groups, $P_R(T)$ is the probability of a correct

response on the studied item for examinees from the reference group with true score T, and $P_F(T)$ is the probability of a correct response on the studied item for examinees from the focal group with true score T.

With SIBTEST, items are divided into two subsets: the studied subtest and the matching subtest. The studied subtest contains items that potentially contain DIF due to the influence of both the primary and secondary dimensions, whereas the matching subtest contains only items that measure the primary dimension. Matching subtest scores at each score level for the reference and focal group are then used to categorize examinees into k subgroups and the groups are compared. $B(T)$ is estimated using $\hat{\beta}_{UNI}$, which is the weighted sum of differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels. The weighted mean difference between the reference and focal groups on the studied subtest item or bundle across the k subgroups is given by

$$\hat{\beta}_{UNI} = \sum_{k=0}^{k} p_k d_k \quad ,$$

where p_k is the proportion of focal group examinees in subgroup k, and d_k is the difference in the means on the studied subtest item for the reference and focal groups across k subgroups.

After testing the DIF hypotheses, SIBTEST provides both a statistical significance test and an effect size for each DIF item. $\hat{\beta}_{UNI}$ is interpreted as the amount of DIF for each item. It has a standard normal distribution with mean 0 and standard deviation 1 under the null hypothesis of no DIF. A statistically significant positive value

of $\hat{\beta}_{UNI}$ indicates DIF favoring the reference group and a negative value indicates DIF favoring the focal group.

Roussos and Stout (1996, p. 220) also proposed guidelines for classifying DIF item based on the $\hat{\beta}_{UNI}$ values. When the null hypothesis is rejected and $\left|\hat{\beta}_{UNI}\right| < 0.059$, DIF is classified as negligible or A-level; when the null hypothesis is rejected and $0.059 \leq \left|\hat{\beta}_{UNI}\right| < 0.088$, DIF is classified as moderate or B-level; and when the null hypothesis is rejected and $\left|\hat{\beta}_{UNI}\right| \geq 0.088$, DIF is classified as large or C-level. However, these guidelines are applicable only for individual DIF items. There is no equivalent guidelines available for classifying bundles of items. Thus, only statistical tests are used for evaluating the DIF hypotheses for bundles of items. Therefore, if only one item in included in the DIF hypothesis, both the significance test and the effect size are used for evaluation. If a bundle of items is included in the same DIF hypothesis, then only the statistical test is used.

To summarize, the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996) combines the use of statistical and substantive DIF analyses to detect DIF. It can also provide information about why DIF occurs. No study has been conducted to apply the outcomes of the multidimensionality-based DIF analysis paradigm in selecting common items when equating translated tests. The present research is designed to evaluate the application of the multidimensionality-based DIF analysis paradigm in selecting common items using the common-item nonequivalent groups design for equating translated tests.

CHAPTER III METHOD

The data, samples, and analyses that were used in the present research are described in this chapter. This chapter is organized into three sections. The data sets are described in the first section. It includes a description of the Alberta Education Achievement Tests (English and French test forms) and Hong Kong Certificate of Educational Examination (English and Chinese test forms). The second section includes a description of the sample size for each test, the standard setting procedure used to set the cut-scores, and the grouping of the sample according to the standards. The third section describes the data analysis procedures. Three DIF conditions were used for comparison: No DIF, Exploratory DIF, and Confirmatory DIF.

Section I: Data

A total of six data sets in three languages (English, French, and Chinese), four content areas (Mathematics, Social Studies, Economics and History), and three grades [Grades 6, Grade 9, and Secondary 5 (Grade 12)] were used in the present study. Four data sets correspond to the English and French test forms for the Grade 6 and Grade 9 Mathematics and Social Studies from the Alberta Education Achievement testing program administered in 1997. Two data sets correspond to the English and Chinese test forms for Secondary 5 in the content areas of Economics and History from the Hong Kong Certificate of Education Examination Program (HKCEE) administered in 1999.

The Grade 6 Mathematics Achievement test has 50 multiple-choice (MC) items categorized into five content areas: Number Relations, Fractions, Computation and Operations, Measurement and Geometry, and Data Analysis (Alberta Education, 1996). The Grade 9 Mathematics Achievement test has 45 MC items and 10 numeric response

items categorized into four content areas: Number, Patterns and Relations, Shape and Space, and Statistics and Probability (Alberta Education, 1996). The Grade 6 Social Studies Achievement test has 50 MC items. These items were categorized into four content areas: Local Government, Ancient Greek Civilization, China, and Geography and Mapping (Alberta Education, 1989). The Grade 9 Social Studies Achievement test has 55 MC items categorized into four content areas: Technology and Change, Economic Systems, Quality of Life available from Different Economic Systems, and the Former USSR (Alberta Education, 1989). The Secondary 5 Economics HKCEE test has 54 MC items categorized into nine content areas: Basic Economic Problems, Demand, Supply and Price, Production, Units of Production, Market Structure, National Income, Money and Banking, Public Finance, and International Trade (Hong Kong Examinations Authority, 1999). The Secondary 5 History HKCEE test has 40 MC items categorized into eight content areas: Rise of Nation-States in Europe, China from Self-Strengthening Movement, the Rise of Japan as a World Power, Russian Revolutions, First World War, the Developments in Major Countries, Second World War, and the Contemporary World (Hong Kong Examinations Authority, 1999).

Section II: Sample

In general, IRT equating requires a larger sample than equipercentile equating, which requires a larger sample than linear equating methods (Harris, 1993). In the present research, the sample size for each language group from each data set was controlled at a minimum of 2,000 examinees, which is adequate to produce stable equating results across all the equating methods used in the present study (Harris, 1993; Harris & Crouse, 1993; Kolen & Brennan, 2004). Samples of examinees were randomly

selected from each data set. For the Grade 6 and Grade 9 Mathematics and Social Studies Achievement tests, 3,000 English examinees (out of 38,200) and 2,115 French examinees (out of 3,000) were randomly selected in each data set. For the Economics HKCEE test, 3,000 English examinees (out of 27,600) and 3,000 Chinese examinees (out of 10,730) were randomly selected. For the History HKCEE test, 3,000 English (out of 11,720) and 3,000 Chinese examinees (out of 7,710) were randomly selected.

Section III: Analyses

Three DIF conditions were compared: No DIF, Exploratory DIF, and Confirmatory DIF. To make these comparisons, the data analyses were completed in three steps: (1) identify DIF items in each condition, (2) equate the translated achievement tests on to a common scale using the common-item nonequivalent groups equating design, and (3) evaluate the effects of DIF on the equated translated tests.

Step 1: Identify DIF Items

The first step was to identify the DIF items for the three DIF conditions. The non-DIF items were used as the common-items in the common-item nonequivalent groups equating design. A process chart for three DIF conditions in identifying the DIF and non-DIF items is provided in Appendix A and the process used for each condition is described below.

Condition 1: No DIF

The No DIF condition was based on one of two assumptions: the test was translated perfectly into different language forms, as in the ideal situation, or DIF items were ignored in the translated tests. Thus, items on different language forms were considered to be equivalent and the tests should display no translation DIF. Often, in

many practical situations, the presence or absence of DIF is not evaluated on translated tests. Further, equating is rarely performed between different language tests. That is, the translated tests are assumed comparable to the original test without equating. The No DIF condition was used in the present research as a control condition against which the DIF conditions (exploratory and confirmatory DIF conditions) were compared.

Condition 2: Exploratory DIF Condition

In the Exploratory DIF condition, SIBTEST (Shealy & Stout, 1993) was used to conduct the statistical DIF analyses in the present research to identify DIF items. The English examinees were the reference group and the French or Chinese examinees were the focal group. Single-item DIF analyses were conducted using SIBTEST to identify the DIF items in each data set. To evaluate the DIF items, both the statistical significance test with an alpha-level of 0.05 and the guidelines developed by Roussos and Stout (1996) were used to classify DIF items based on their $\hat{\beta}_{UNI}$ values. When the null hypothesis is rejected and $\left|\hat{\beta}_{UNI}\right| < 0.059$, DIF is classified as negligible or A-level; when the null hypothesis is rejected and $0.059 \leq \left|\hat{\beta}_{UNI}\right| < 0.088$, DIF is classified as moderate or B-level; and when the null hypothesis is rejected and $\left|\hat{\beta}_{UNI}\right| \geq 0.088$, DIF is classified as large or C-level. The B- or C-level DIF items were considered as DIF in the present research because these DIF items are typically scrutinized for potential bias in tests reviews (Zieky, 1993). The remaining items were considered to be non-DIF items and formed the common-item sets for equating.

Condition 3: Confirmatory DIF Condition

In the Confirmatory DIF condition, the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996) was used to detect DIF items. The first stage uses substantive analysis to interpret the items and generate the DIF hypotheses, and the second stage uses statistical analysis to test the DIF hypotheses and confirm the results.

Substantive analysis. The four sources of translation DIF described by Gierl and Khaliq (2001) were applied as the substantive analysis to interpret the items. These four sources of translation DIF are (1) omissions or additions of words, phrases, (2) differences in words, expressions, or sentence structure of items, (3) differences in words, expressions, or sentence structure of items that is not inherent to language and/or culture, and (4) differences in punctuation, capitalization, item structure, typeface, and other formatting usages.

In the translation review process, four translators used the four sources of translation DIF developed by Gierl and Khaliq (2001) to identify sources of DIF in all the data sets. Two translators, bilingual in English and French and who had extensive experience in translating educational tests, texts, and documents, used the four sources of translation DIF to identify sources of DIF in the Grade 6 and Grade 9 Mathematics and Social Studies achievement tests (Gierl & Khaliq, 2001). Two different translators, bilingual in English and Chinese, identified the sources of DIF in HKCEE Economics and History tests (Gierl et al., 2000). In the review process, the translators first identified DIF items separately in each test. They then specified for each DIF item which language group would be favored, identified the reason(s) for the difference, and categorized the reason(s) into the four sources of translation errors identified by Gierl and Khaliq (2001).

Once the reviews were completed, the translators met to discuss their decisions. The translators reached consensus on the source of DIF for all the DIF items they identified in the six data sets. Items attributed to translation DIF (bias) were considered as DIF items and items attributed to group differences (impact) were retained in the tests. These substantive analyses have been completed by Gierl and Khaliq (2001) and Gierl et al. (2000), and their results were used in the present research as to identify DIF items in the confirmatory DIF condition.

Statistical analysis. The second stage of the confirmatory DIF condition used SIBTEST (Shealy & Stout, 1993) to confirm the results statistically. Single items or bundles of items categorized by the sources of translation error were analyzed. If only a single item was presented in the DIF hypothesis, the statistical test and the effect sizes were used to evaluate the presence or absence of DIF. If bundles of items were presented in the DIF hypothesis, only the statistical test was used to evaluate the DIF bundles. Only the items flagged in the substantive analyses and confirmed by the statistical analyses were considered as DIF items in the confirmatory condition. That is, items not identified in substantive analyses and items identified in the substantive analyses but not in the statistical analyses were included in the common-item set for equating.

As indicated earlier, three DIF conditions were considered. The No DIF condition assumed the absence of DIF items and thus, no equating was followed. The exploratory DIF condition used statistical DIF analyses to identify DIF items. The non-DIF items were included in the common-item set for equating. The confirmatory DIF condition used the multidimensionality-based DIF analysis paradigm to identify DIF items. Items identified in the substantive analyses and confirmed in the statistical

analyses were considered as DIF items and not included in the common-item sets. The common-item sets from the exploratory and the confirmatory DIF conditions were used to equate the translated tests.

Step 2: Equate the Translated Tests

The second step was to equate the translated tests (French or Chinese forms) on to the original scale (English form) using the common-item nonequivalent groups equating design according for the exploratory and confirmatory DIF conditions. Equated results from the two DIF conditions were compared with the No DIF condition. A multi-method approach was applied to evaluate the consistency among the equating results. Linear and non-linear observed score equating methods were used so that outcomes across different equating methods could be compared. Four equating methods were selected in the present research. These included two linear - Tucker linear and Levine observed score equating- and two non-linear - equipercentile equating and IRT observed score equating - equating methods. The assumptions, equating functions, and computer programs used for each equating method are described.

Tucker linear equating method. Gulliksen (1950) described the Tucker linear equating method, which transforms the observed scores on Form X to the observed scores on the scale of Form Y. Form X is the translated test taken by the focal group selected from population 1 and the random variable score is X, and Form Y is the source language test taken by the reference group from population 2 and the random variable score is Y.

In equating, populations 1 and 2 must be combined to obtain a single population for defining an equating relationship. To obtain a single population, Braun and Holland

(1982) introduced the concept of a synthetic population where populations 1 and 2 are weighted by w_1 and w_2, respectively, such that

$$w_1 + w_2 = 1, \text{ and } w_1, w_2 \geq 0.$$

For the purpose of the present study, w_1=1 and w_2=0. To compare population 1 and 2 on a common scale, the scores of Form X (focal group) are equated and transformed on to Form Y (reference group).

There are two assumptions for the Tucker linear equating method (Gulliksen, 1950). First, the regression of X (total score) on V (common-item scores) and the regression of Y on V are assumed to be the same linear function for both populations 1 and 2. Second, the conditional variance of X given V and the conditional variance of Y given V are assumed to be the same for populations 1 and 2. According to these assumptions, the equation function for the synthetic population is given by

$$l_{Ys}(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - \mu_s(X)] + \mu_s(Y),$$

where the synthetic population means and variances are

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1[\mu_1(V) - \mu_2(V)],$$
$$\mu_s(Y) = \mu_1(Y) + w_1\gamma_2[\mu_1(V) - \mu_2(V)],$$
$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_1^2[\mu_1(V) - \mu_2(V)]^2,$$
$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2[\sigma_1^2(V) - \sigma_2^2(V)] + w_1w_2\gamma_2^2[\mu_1(V) - \mu_2(V)]^2,$$

where s refers to the synthetic population. The regression slopes are

$$\gamma_1 = \frac{\sigma_1(X,V)}{\sigma_1^2(V)},$$
$$\gamma_2 = \frac{\sigma_2(Y,V)}{\sigma_2^2(V)}$$

The Tucker linear equating was performed with the program CIPE (Kolen, 2003).

Levine observed score method. Levine (1955) developed the Levine observed score method. This is an observed score equating method based on the assumptions about true scores (Kolen & Brennan, 2004). Three assumptions underline the Levine observed score equating method. First, it is assumed that X, Y, and V are all measuring the same true score such that T_X (true score for X) and T_V as well as T_Y and T_V are perfectly correlated in both populations 1 and 2. Second, the regression of T_X on T_V and the regression of T_Y on T_V are assumed to be the same linear function for both populations 1 and 2. Third, the measurement error variance for X is assumed to be the same for populations 1 and 2. A similar assumption is made for Y and V.

According to the above assumptions, the synthetic population equating function and the synthetic population means and variances are the same as with the Tucker Linear Equating Method. The regression slopes are given by

$$\gamma_1 = \frac{\sigma_1(T_X)}{\sigma_1(T_V)} = \frac{\sigma_1(X)\sqrt{\rho_1(X,X')}}{\sigma_1(V)\sqrt{\rho_1(V,V')}},$$

$$\gamma_2 = \frac{\sigma_2(T_Y)}{\sigma_2(T_V)} = \frac{\sigma_2(Y)\sqrt{\rho_2(Y,Y')}}{\sigma_2(V)\sqrt{\rho_2(V,V')}}.$$

The Levine observed score equating was also performed with the program CIPE (Kolen, 2003).

Equipercentile equating. Angoff (1971) and Braun and Holland (1982) developed frequency estimation equipercentile equating that estimates the cumulative distributions of scores on Form X and Form Y for a synthetic population using the common-item nonequivalent groups design. The equipercentile equating method assumes that the conditional distribution of scores, given the common-item scores (V), is the same in both populations for Forms X and Y (Kolen & Brennan, 2004).

Equipercentile equating can be viewed as a two-stage process (Kolen, 1984). First, the relative cumulative frequency distributions are tabulated for the two forms to be equated. Second, equated scores are obtained from these relative cumulative frequency distributions. The equating function transforms the distribution of scores of Form X to Form Y when the percentile ranks on both forms are set to be equal (Kolen & Brennan, 2004).

In the first stage, the frequency estimation equipercentile equating formula are given by

$$f_s(x) = w_1 f_1(x) + w_2 \sum_v f_1(x|v) h_2(v),$$
$$g_s(x) = w_1 \sum_v g_2(y|v) h_1(v) + w_2 g_2(y)$$

where f refers to the distribution of X for Form X, g refers to the distribution of Y for Form Y, h refers to the distribution of scores, s refers to the synthetic population, and w_1 and w_2 are used to weight populations 1 and 2 to form the synthetic population, and $w_1 + w_2 = 1$. For the synthetic population, $f_s(x)$ can be cumulated over values of x to produce the cumulative frequency distribution $F_s(x)$. The cumulative frequency distribution $G_s(y)$ is similarly derived for value of y.

In the second stage, the equipercentile equating function for equating X to Y on the synthetic population is given by

$$ey_s(x) = Q_s^{-1}[P_s(x)],$$

where $P_s(x)$ is the percentile rank for Form X, and Q_s^{-1} is the inverse of the percentile rank function for Form Y.

The percentile ranks on both Forms X and Y are set to be equal. The equated scores are obtained from these relative cumulative frequency distributions. The equipercentile equating was performed with the program CIPE (Kolen, 2003).

IRT observed score equating. Kolen (1981) and Han (1993) developed the IRT observed score equating that uses the IRT model to produce an estimated item and ability (theta) parameters for each form to generate an estimated observed score distribution of number-correct scores for each form, which then are equated using equipercentile methods (Han, Kolen & Pohlmann, 1997).

The IRT observed score equating was conducted in three stages. At the first stage, the 2PL IRT item parameters were estimated and theta distributions were produced using BILOG 3.11 (Mislevy & Bock, 1997). Among the three IRT models - 1PL, 2PL, and 3PL models- the 2PL model was selected in the present study for three reasons. The 1PL model only encounters the item difficulties and not the item discriminations. However, discrimination indices ranged from 0.37 to 0.54 for the Grade 6 and Grade 9 Social Studies and Mathematics achievement tests (Gierl & Khaliq, 2001), and from 0.28 to 0.52 for the Economics and History HKCEE tests (Gierl, et al., 1999). Hence, the 1PL model was not adequate for the present study. The 3PL model was not chosen because the item parameters could not be estimated due to the non-convergence of the conditional marginal maximum likelihood estimates (Puhan, 2003). Consequently, the 2PL model was selected. These analyses were performed separately for each test form.

The second stage put the parameter estimates of the translated tests on to the same scale as the parameter estimates of the original English tests, using a rescaling function to transform the translated tests parameter estimates to the English scale. The Stocking and

Lord's (1983) parameter rescaling equation was used to transform the parameter estimates with the program ST (Zeng & Hanson, 1995). At the third stage is to compute the equated scores of the translated tests on to the English scale were computed. The item parameter estimates of the English tests, the rescaled item parameter estimates of the translated tests, and the estimated theta distributions of the English and translated tests were used to compute the IRT observed equated scores with the program PIE (Hanson & Zeng, 1995).

Step 3: Evaluate the Effect of DIF on Equating

Outcomes across the equating methods were used to evaluate the effect of DIF on equating. The evaluation focused on the common-item sets, the equated scores, and the classification of examinees according to the standards set for the translated tests in French or Chinese.

Evaluate the Common-item Sets

Common-item equivalence is a requirement for equating different language tests using the common-item nonequivalent groups design. It is recommended that the common-item set should have 15 to 20 common items or be at least 20% of the full-length test (test with 40 or more items) to be sufficient for equating purposes (Angoff, 1971; Kolen & Brennan, 2004; Wainer, 1999). The content characteristics should be remained when the common-item sets have the same test specifications as the full-length tests (Kolen & Brenan, 2004). Moreover, a high correlation between the common-item set and the full-length test is recommended as an important determinant for assessing the accuracy of the equating process (Beguin, 2002; Budescu, 1985).

In order to evaluate the common-item sets from the Exploratory and Confirmatory DIF conditions in the present research, three criteria were used: The number of common items, the proportions (percentages) of the common items as compared to the full-length test specifications, and the correlations of the examinees scores between the common-item sets and the full-length tests.

Evaluate the Equated Scores

Once the common-item equivalence requirement was fulfilled, the common-items were used for equating. After equating the translated achievement tests onto their corresponding original test scales using the common-item nonequivalent groups equating design, the new test scores of the translated tests were considered equivalent to their original test scores, and they were labeled as the equated scores. The equated scores were unrounded and the equating results from the exploratory and confirmatory conditions were evaluated.

To evaluate the relative magnitude for the unrounded equated scores, Dorans, Holland, Thayer, and Tateneni (2003) and Dorans (2004) suggested considering a score "difference that matters" (DTM), which is half of a reported score unit (Kolen & Brennan, 2004). The DTM logic is as follows: If the score differences between the equated and the non-equated scores are within half a reported score unit, then the score differences are ignorable. As the reported scores from all the data sets in the present research were reported as an integer, a 0.50 score was used as the DTM in the present study. That is, if the equated mean score and the original mean score differences are less than 0.50, the score differences are considered to be unimportant. Alternatively, if the

mean score differences are equal to or greater than 0.50, the score differences are considered to be important.

To evaluate further the effect of equating along the score scale across the three DIF conditions, student performance at three cut-scores in the observed score distribution for the students who wrote the translated tests was determined. These three points were at minus one standard deviation (-1SD), mean, and plus one standard deviation (+1 SD). These cut-scores were chosen to reflect the similarity of the Achievement Tests that classify examinees' performances into the Acceptable Standard and the Standard of Excellence (Alberta Education, 2003). The samples in each data set were then categorized into four groups according to their original test score: below 1 SD, between minus 1 SD and mean, between mean and plus 1 SD, and above 1 SD.

The equated mean scores of the total sample and the four subgroups using the translated tests were calculated. By comparing the unrounded equated mean scores of the subgroups from the DIF condition (either exploratory or confirmatory) and the No-DIF condition, the differences in the mean equated score were calculated. A difference between the equated mean scores which was greater than or equal to 0.50, the DTM score, was considered an important change. For example, if the mean score difference of the subgroup (above 1SD) using the Tucker linear equating method in the exploratory condition in one of the data sets is 0.75, the difference is greater than the DTM score at 0.50, and it is considered as an important change in the equated scores on the translated test. The number of changes occurring in each equating method, in each DIF condition, and in each data set were calculated and evaluated according to the DTM logic.

To evaluate the effect of DIF on equating translated tests, the unrounded equated scores seems preferable because they do not incorporate the added "noise" that results from the rounding scores (Kolen & Brennan, 2004). However, in many practical circumstances, the reported scores that are actually used to make decisions about examinees are always rounded to integers. If the rounded equated scores increase or decrease a reported score point which located around the cut-scores, that may affect the classification of examinees in meeting the standards, and thus, misclassification of examinees may occur. Kolen and Brennan (2004) suggested that the extent to which such differences in the equated scores "matter" depends on the nature of the decisions that are made and where along the score scale that decisions are made. Therefore, the rounded equated scores were used to evaluate the re-classification of examinees in meeting the standards after equating.

Evaluate the Classification of Examinees

The classification of examinees was evaluated by comparing the number and proportion of examinees using the translated tests in meeting each standard before and after equating. In the present research, the standards were determined by using three cut-scores at -1 standard deviation, mean, and +1 standard deviation from each translated test. Based on the cut-off scores before equating, the number and proportion of examinees meeting each standard using the translated tests were calculated using the rounded equated scores, and comparisons were made between the DIF condition (either Exploratory or Confirmatory) and the No DIF condition. The differences in the number and proportion of examinees in meeting each standard before and after equating were evaluated.

In sum, the purpose of the present research was to evaluate the effect of using the multidimensionality-based DIF analysis paradigm on the common-item nonequivalent group equating design using translated achievement tests. Three DIF conditions were created for comparison: (1) No DIF, (2) Exploratory DIF, and (3) Confirmatory DIF. This research was conducted using four equating methods on six data sets from achievement tests that were translated from English to French (Social studies and Mathematics) and English to Chinese (Economics and Physics). Evaluation focused on the common-item sets, equated scores, and the classification of the examinees in the translated tests. Results of the data analyses are presented in the next chapter.

CHAPTER IV RESULTS

This chapter is organized into three sections. The first section describes the summary statistics of the six data sets used in the present research. It includes a description of the number of examinees, the total number of test items, the mean score, standard deviation, and the reliability of the original and translated tests. The second section contains the results of the evaluations, which are focused on the common-item sets, the equated scores of the translated tests, and the classification of examinees in meeting the standards before and after equating. The third section contains a summary in which the results from these six data sets are summarized.

Section I: Summary Statistics

The summary statistics for the six data sets considered in the research are presented in Table 1. These data sets include the Grade 6 Social Studies Achievement Test (English-French), the Grade 9 Social Studies Achievement Test (English-French), the Grade 6 Mathematics Achievement Test (English-French), the Grade 9 Mathematics Achievement Test (English-French), the Economics HKCEE (English-Chinese), and the History HKCEE (English-Chinese). The number of examinees, the total number of test items, the mean test score, the standard deviation, and the reliability (internal consistency) for both the original and the translated tests are presented.

The numbers of examinees in each test are listed in Table 1. These sample sizes were of at least 2,000 and they are adequate to produce stable equating outcomes (Harris, 1993; Harris & Crouse, 1993; Kolen & Brennan, 2004). Kolen and Brennan (2004) suggested that when the mean difference between the two groups to be equated was less than 0.5 standard deviation units, and the ratios of the groups' standard deviations was

Table 1

Summary Statistics for all Six Data Sets: Original and Translated Tests

Data set	Original Test	Translated Test
Grade 6 Social Studies (English-French)		
Number of examinees	3000	2115
Number of items	50	50
Mean	33.68	32.17
SD	8.34	7.77
Reliability index	0.87	0.84
Grade 9 Social Studies (English-French)		
Number of examinees	3000	2115
Number of items	55	55
Mean	38.31	39.04
SD	8.98	7.98
Reliability index	0.88	0.86
Grade 6 Mathematics (English-French)		
Number of examinees	3000	2115
Number of items	50	50
Mean	35.30	36.56
SD	8.46	7.45
Reliability index	0.89	0.86
Grade 9 Mathematics (English-French)		
Number of examinees	3000	2115
Number of items	49	49
Mean	29.04	33.52
SD	10.24	8.72
Reliability index	0.92	0.89
Economics (English-Chinese)		
Number of examinees	2000	2000
Number of items	54	54
Mean	25.78	22.82
SD	8.73	6.79
Reliability index	0.85	0.87
History (English-Chinese)		
Number of examinees	2000	2000
Number of items	38	38
Mean	21.88	21.25
SD	8.00	7.50
Reliability index	0.89	0.88

Note. Reliability index is calculated using Cronbach's alpha (1951).

smaller than 1.2, then the two groups were similar in ability. Comparison of the mean scores of the examinees taking the original and translated tests indicated that the

differences were less than 0.50. The ratios of the standard deviation for all were smaller than 1.2 except for the Grade 9 Mathematics test (ratio = 1.23) and the Economics test (ratio = 1.28). These results indicated the ability differences between the two groups of examinees were small for the six tests. Moreover, the reliabilities (internal consistency) of pairs of the original and the translated tests were similar. Further, Cronbach's alpha ranged between 0.84 and 0.92, suggesting that the tests were internally consistent. The similar values for the means, standard deviations, and internal consistencies suggested that the two groups performed similarly in the original and the translated tests (Kolen & Brennan, 2004; Puhan, 2003). There are two advantages when the two groups were in similar ability levels: The accuracy for DIF detection is enhanced (Hambleton, 1993), and the results are suitable for common-item nonequivalent group equating (Kolen & Brennan, 1995; Raff & Allalouf, 2002).

Results of the Evaluation

Three steps were conducted in the data analyses: (1) identify DIF and common items in each DIF condition, (2) equate the translated achievement tests onto a common scale using the common-item nonequivalent groups equating design, and (3) evaluate the effect of DIF on the equated translated tests. The evaluations are focused on the common-item sets, the equated scores of the translated tests, and the classification of examinees in meeting the standards of performance.

The Common-item Sets Evaluation

Three criteria were used to evaluate whether or not the common-item sets were sufficient for equating: (1) the common-item set should have 15 to 20 common items and be at least 20% of the full-length test (test with 40 or more items) (Angoff, 1971; Kolen

& Brennan, 2004; Wainer, 1999), (2) the common-item sets should have test specification proportions that are approximately equal to the corresponding proportions for the full-length tests (Kolen & Brennan, 2004), and (3) the scores of common-item sets should have a high correlation with the full-length (total) test scores (Beguin, 2002; Budescu, 1985). To evaluate the first criterion, the numbers of common items in the six tests are presented in Table 2. As shown, the No DIF condition contained all of the test items in the common-item sets, while the numbers of common items in the exploratory condition ranged from 21 to 46, and the numbers of common items in the confirmatory DIF condition ranged from 33 to 47. The corresponding percentages in terms of the total number of common items in each test were 100 % for the No DIF condition, 44.4 % to 86.0 % in the exploratory DIF condition, and 61.1 % to 97.4 % in the confirmatory DIF condition. Given that there are more than 20 items in the common-item sets and that the common-item sets composed at least 20 % of the full-length tests for each DIF condition, the first evaluation criterion for the common-item set was met.

Table 2

Evaluation of the Common-item Sets in Six Data Sets: Number of Common Items

	No DIF		**Exploratory DIF**		**Confirmatory DIF**	
Data set	Number of Common items	%	Number of Common items	%	Number of Common items	%
Grade 6 Social Studies	50	100.0	21	42.0	30	60.0
Grade 9 Social Studies	55	100.0	35	63.6	44	80.0
Grade 6 Mathematics	50	100.0	43	86.0	47	94.0
Grade 9 Mathematics	55	100.0	46	83.6	46	83.6
Economics	54	100.0	24	44.4	33	61.1
History	38	100.0	23	60.5	37	97.4

Note. Percentage (%) is calculated using the number of common items divided by the number of total items in the corresponding full-length tests.

The second criterion is that the common-item sets should have test specification proportions that approximate the corresponding proportions for the full-length tests. The proportions (percentages) of the common items in each content area corresponding to their full-length test specifications are presented in Table 3. As expected, the six common-item sets in the No DIF condition have the same test specification proportions as the full-length test. The common-item sets test specification proportions differed from the proportions for the full-length test from 0.7 % to 10.2 % in the exploratory condition, and from 0 % to 10.7 % from the confirmatory condition across the six data sets. The common-item sets in the exploratory and confirmatory DIF conditions approximate the test specification of the full-length tests. Therefore, the second criterion was met.

The third criterion is that there be a high correlation between scores obtained from the common-item set and the scores obtained from the full-length test scores of the translated tests. The correlations between the common-item sets scores and the full-length test scores are presented in Table 4. The No DIF condition, which was used as a control condition, has the same values in the mean and SD as the corresponding full-length tests across the six data sets before equating. The correlations were 1.00 for the No DIF condition, and ranged from 0.86 to 0.99 in the exploratory DIF condition and from 0.93 to 0.99 in the confirmatory DIF condition. The results reveal that the third criterion was also met.

Consequently, all common-item sets in the six data sets fulfilled the three evaluation criteria and thus, the common-item sets are adequate for equating. It should be noted that all the common items in the exploratory DIF condition were also identified in the confirmatory DIF condition in the data sets. Therefore, the common items in the

Table 3

Evaluation of the Common-item Sets in Six Data Sets: Proportions of Common Items in Each Content Areas Corresponding to Their Full-length Test Specifications

Content area	No DIF		Exploratory DIF		Confirmatory DIF	
	Number of Common items	%	Number of Common items	%	Number of Common items	%
Grade 6 Social Studies						
1 Local Government	17	34.0	7	33.3	11	36.7
2 Greece	17	34.0	6	28.6	7	23.3
3 China	16	32.0	8	38.1	12	40.0
Grade 9 Social Studies						
1 Technology & Change	21	38.2	11	34.4	13	29.6
2 Economic Systems	21	38.2	15	42.9	19	43.2
3 Quality of Life	8	14.6	5	14.3	8	18.2
4 Former USSR	5	9.1	4	11.4	4	9.1
Grade 6 Mathematics						
1 Number & Fraction	15	30.0	12	27.9	13	27.7
2 Numbers & Decimals	11	22.0	10	23.3	11	23.4
3 Time, Area & Volume	8	16.0	6	14.0	7	14.9
4 2D Figures	10	20.0	9	21.0	10	21.3
5 Interpretations of data	6	12.0	6	14.0	6	12.8
Grade 9 Mathematics						
1 Number Concepts	16	29.1	14	30.4	14	30.4
2 Patterns & Relations	17	31.0	15	32.6	15	32.6
3 Shape & Space	15	27.3	13	28.3	13	28.3
4 Statistics & Probability	7	12.7	4	8.7	4	8.7
Economics						
1 Basic Economics	4	7.4	1	4.2	1	3.0
2 Demand & Supply	8	14.8	3	12.5	5	15.2
3 Production	8	14.8	6	25.0	8	24.2
4 Units of Production	6	11.1	2	8.3	4	12.1
5 Market Structure	1	1.6	0	0.0	0	0.0
6 National Income	10	18.5	2	8.3	3	9.1
7 Money and Banking	6	11.1	4	16.7	5	15.2
8 Public Finance	5	9.3	3	12.5	3	9.1
9 International Trade	6	11.1	3	12.5	4	12.1
History						
1 Rise of Europe	5	13.2	1	4.4	5	13.5
2 China	4	10.6	1	4.4	4	10.8
3 Rise of Japan	4	10.6	2	8.7	4	10.8
4 Russian Revolution	5	13.2	3	13.0	5	13.5
5 First World War	5	13.2	5	21.7	5	13.5
6 Major countries	5	13.2	2	8.7	4	10.8
7 Second World War	5	13.2	5	21.7	5	13.5
8 Contemporary World	5	13.2	4	17.4	5	13.5

Note. Percentage (%) is calculated using the number of common items in each content area divided by the total number of common items in each DIF condition.

Table 4

Evaluation of the Common-item Sets in Six Data Sets: Correlations between the Common-item Sets Scores and the Full-length Translated Test Scores

Translated Test	No DIF	Exploratory DIF	Confirmatory DIF
Grade 6 Social Studies			
Mean of common items	32.17	14.91	20.14
SD	7.77	3.61	4.98
Correlation	1.00	0.91	0.95
Grade 9 Social Studies			
Mean of common items	39.04	25.37	31.48
SD	7.98	5.02	6.61
Correlation	1.00	0.96	0.98
Grade 6 Mathematics			
Mean of common items	36.56	32.24	37.84
SD	7.54	6.37	7.17
Correlation	1.00	0.99	0.99
Grade 9 Mathematics			
Mean of common items	35.52	28.16	28.16
SD	8.72	7.15	7.15
Correlation	1.00	0.98	0.98
Economics			
Mean of common items	22.82	9.35	13.33
SD	6.79	3.18	4.19
Correlation	1.00	0.86	0.93
History			
Mean of common items	21.25	12.88	20.63
SD	7.50	4.67	7.23
Correlation	1.00	0.96	0.99

Note. Correlation is calculated using the examinees' common-item scores and their full-length test scores.

exploratory condition overlapped with the common items in confirmatory condition. The percentage overlaps were 70 % in the Grade 6 Social Studies test, 79.6 % in the Grade 9 Social Studies test, 91.5 % in the Grade 6 Mathematics test, 100 % in the Grade 9 Mathematics test, 72.8 % in the Economics test, and the 62.2 % in the History test. The high degree of overlap between the two common-item sets suggests the similarity in the common items across these conditions.

Given the common-item sets satisfied the requirements established for equating, the next step involved equating and evaluating the equating results. The No DIF

condition assumed that the original and the translated test forms were comparable, and thus, no equating was conducted in the condition. In the exploratory and the confirmatory DIF conditions, the translated tests were equated onto the corresponding original test forms using the common-item nonequivalent groups equating design. Four equating methods were used: Tucker linear equating, Levine observed score equating, equipercentile equating, and IRT observed score equating.

The Equated Scores Evaluation

To evaluate the relative magnitude for the unrounded equated scores, the "Difference that matters" (DTM) score at 0.50, which is half of a reported score unit in the present research, was used as the criterion (Kolen & Brennan, 2004; Dorans, 2004; Dorans et al., 2003). A difference between the unrounded equated mean score and the non-equated mean score which was greater than or equal to 0.50 was considered an important change in the equating results for the examinees who wrote the translated tests.

Further, to evaluate the equating results along the score scale across the three DIF conditions, the samples who wrote the translated tests were then categorized into four sub-samples based on their test scores on the translated tests. Three cut-scores in the translated tests before equating were determined at 1 standard deviation (SD) below the mean, the mean, and 1 SD above the mean. The differences in the equated mean scores for the total sample and the four sub-samples formed using the three cut-scores are presented in Table 5. Mean score differences are calculated using either the mean score from the exploratory or the confirmatory DIF condition (after equating) minus the mean score from the No DIF condition (before equating).

Table 5

Evaluation of the Unrounded Equated Scores in Six Translated Tests

Translated Test	Mean Score Differences							
	Exploratory DIF – No DIF				Confirmatory DIF – No DIF			
	Tu	Eq	Le	IRT	Tu	Eq	Le	IRT
Grade 6 Social Studies								
Total sample	0.50*	0.46	0.30	0.24	0.54*	0.53*	0.45	0.43
Sample below -1SD	0.14	0.10	0.20	0.30	0.35	0.34	0.41	0.53*
Sample between -1SD and mean	0.38	0.43	0.27	0.24	0.47	0.55*	0.44	0.47
Sample between mean and 1SD	0.59*	0.44	0.33	0.22	0.59*	0.50*	0.47	0.40
Sample above 1SD	0.78*	0.65*	0.38	0.24	0.69*	0.53*	0.49	0.35
Grade 9 Social Studies								
Total sample	0.15	0.13	0.22	0.19	0.08	0.07	0.10	0.10
Sample below -1SD	-0.05	0.00	0.28	0.21	-0.33	-0.29	-0.22	-0.16
Sample between -1SD and mean	0.09	-0.04	0.23	0.15	-0.05	-0.06	0.00	-0.02
Sample between mean and 1SD	0.20	0.12	0.20	0.16	0.18	0.10	0.18	0.16
Sample above 1SD	0.29	0.47	0.17	0.28	0.38	0.46	0.33	0.35
Grade 6 Mathematics								
Total sample	0.19	0.18	0.22	0.21	0.30	0.29	0.31	0.31
Sample below -1SD	-0.02	0.09	0.07	0.11	0.02	0.14	0.05	0.14
Sample between -1SD and mean	0.13	-0.05	0.18	0.09	0.22	0.11	0.24	0.19
Sample between mean and 1SD	0.25	0.22	0.26	0.23	0.38	0.33	0.38	0.36
Sample above 1SD	0.34	0.53*	0.33	0.46	0.50*	0.64*	0.49	0.54*
Grade 9 Mathematics								
Total sample	0.24	0.23	0.33	0.30	0.24	0.23	0.33	0.30
Sample below -1SD	0.34	0.15	0.53*	0.28	0.34	0.15	0.53*	0.28
Sample between -1SD and mean	0.27	0.19	0.39	0.36	0.27	0.19	0.39	0.36
Sample between mean and 1SD	0.21	0.35	0.27	0.34	0.21	0.35	0.27	0.34
Sample above 1SD	0.16	0.18	0.17	0.22	0.16	0.18	0.17	0.22
Economics								
Total sample	0.20	0.14	-0.33	-0.47	0.43	0.41	0.17	0.09
Sample below -1SD	-0.51*	-0.90*	-0.01	-0.18	0.04	-0.48	0.38	0.20
Sample between -1SD and mean	-0.06	-0.20	-0.21	-0.41	0.29	0.16	0.25	0.11
Sample between mean and 1SD	0.41	0.81*	-0.42	-0.59*	0.54*	1.01*	0.11	0.02
Sample above 1SD	0.98*	0.58*	-0.68*	-0.58*	0.86*	0.61*	-0.07	0.08
History								
Total sample	0.28	0.25	0.25	0.26	0.16	0.17	0.16	0.16
Sample below -1SD	0.11	-0.14	0.18	0.17	0.28	0.21	0.28	0.22
Sample between -1SD and mean	0.21	0.25	0.22	0.25	0.21	0.29	0.21	0.20
Sample between mean and 1SD	0.33	0.42	0.28	0.27	0.12	0.13	0.12	0.09
Sample above 1SD	0.45	0.31	0.33	0.35	0.03	0.01	0.03	0.16

Note. Mean score differences are calculated using either the mean score from the exploratory or confirmatory DIF condition minus the mean score from the No DIF condition.

Tu= Tucker linear equating, Eq= equipercentile equating, Le= Levine observed score equating, and IRT=IRT observed score equating.

* The mean score difference ≥ 0.50 of a score unit.

Grade 6 Social Studies

The results revealed that the mean score difference for the total sample between the exploratory and the No DIF conditions in the Grade 6 Social Studies test using Tucker linear equating method is 0.50 (see Table 5). As the mean score difference is equal to the DTM score at 0.50, the difference between the mean equated scores is considered to be an important change. Moreover, the results revealed that mean score differences in the sub-samples - between the mean and 1SD (0.59) and sample above 1SD (0.78) - using Tucker linear equating, and in the sample above 1SD (0.65) using equipercentile equating are also greater than the DTM score (see Table 5). Consequently, the results indicate that important changes occurred three times when using Tucker linear equating, once when using equipercentile equating, and none when using Levine or IRT observed score equating for the exploratory DIF condition. In the case of the confirmatory versus the No DIF comparisons, the results from the equated scores indicate that important changes occurred three times when using Tucker linear equating, four times when using equipercentile equating, and once when using IRT equating.

Grade 9 Social Studies

For the Grade 9 Social Studies test, the results revealed that all the mean score differences between the unrounded mean equated score were less than 0.50 for both the exploratory and confirmatory DIF conditions. Thus, important changes in the mean equated scores did not occur in this test.

Grade 6 Mathematics

The results for the Grade 6 Mathematics test revealed that important changes in the equated mean scores occurred once in the sample above 1SD (0.53) using

equipercentile equating for the exploratory DIF condition (see Table 5). Important changes also occurred three times for the confirmatory DIF condition for the same sub-sample using Tucker linear equating (0.50), equipercentile equating (0.64) and IRT observed score equating (0.54).

Grade 9 Mathematics

The results for the Grade 9 Mathematics test revealed that important changes in the equated mean scores occurred once in the exploratory DIF condition and once in the confirmatory DIF condition. Changes occurred in the sample below minus 1SD using Levine observed score equating (0.53) in both the exploratory and confirmatory condition versus No DIF comparisons. Since the exploratory and confirmatory DIF conditions have the same set of common items, the equating results are the same in both conditions.

Economics

The results for the Economics test revealed that important changes in the equated mean scores occurred eight times in the exploratory DIF condition comparison. These differences included the sample below -1SD (-0.51) and the sample above 1SD (0.98) using Tucker linear equating; the sample below -1SD (-0.90), the sample between the mean and 1SD (0.81), and the sample above 1SD (0.58) using equipercentile equating; the sample above 1SD (-0.68) using Levine observed score equating; the sample between the mean and 1SD (-0.59) and the sample above 1SD (-0.58) using IRT observed score equating (see Table 3). Important changes occurred four times in the confirmatory DIF condition. These changes include the sample between the mean and 1SD (0.54) and the sample above 1SD (0.86) using Tucker linear equating; and the sample between the mean and 1SD (1.01) and the sample above 1SD (0.61) using IRT observed score equating.

History

The results for the History test revealed that for both exploratory and confirmatory DIF conditions: all the mean score differences of the unrounded equated mean score were less than 0.50. Thus, the equated scores showed no important changes.

In sum, the equated mean scores of all six translated tests were evaluated. The results revealed that in four of the six tests, the equated mean scores in the exploratory and confirmatory DIF conditions contained important changes relative to the new equated scores. These four tests included the Grade 6 Social Studies Achievement Test, the Grade 6 Mathematics Achievement Test, the Grade 9 Mathematics Achievement Test, and the Economics HKCEE. For the other two translated tests, Grade 9 Social Studies Achievement Test and the History HKCEE, the equated mean scores in the exploratory and confirmatory DIF conditions revealed no important changes in either DIF condition.

The Classification of Examinees Evaluation

The accuracy of the classification of examinees was evaluated by comparing the number of examinees who wrote the translated tests and who were placed in each classified category in meeting the standards before and after equating the translated tests. Any difference in the number of examinees placed in each classification category is considered an important change. In the present research, the standards are determined by using three cut-scores at -1 standard deviation, mean, and +1 standard deviation in the initial observed score distribution of each translated test before equating. The numbers of examinees meeting each standard after equating the translated tests are calculated using the rounded equated scores. The results of the comparisons; which are reported in Table 6, are made between the DIF conditions (either exploratory or confirmatory) and

the No DIF condition. Differences in the number of examinees classification were calculated using either the number of examinees classified in Exploratory or Confirmatory DIF condition minus the number of examinees classified in the No DIF condition. The total number of re-classifications represents the number of examinees in each test who were re-classified into a different standard after equating.

Grade 6 Social Studies

The three cut-scores before equating were at the score values of 24 (-1SD), 32 (mean), and 40 (+1SD) for the Grade 6 Social Studies test. Thus, examinees are classified into four sub-samples: sample below -1SD (score ranges from 0 to 23), sample between -1SD and the mean (score 24 to 31), sample between the mean and 1SD (score 32 to 39), and sample above 1SD (score 40 to 55). The number of examinees placed in each standard in the exploratory or confirmatory DIF conditions after equating is compared to the number of examinees in the No DIF condition.

In the exploratory condition, the results revealed that there are differences in the number of examinees in two sub-samples when Tucker linear equating method was used: the sample between the mean and 1SD (n = -96) and the sample above 1SD (n = +96) (see Table 6). The examinees (n = 96) with a reported score of 39 before equating, which is one-score point below the 1SD cut-score at 40, were initially categorized in the sample between the mean and 1SD before equating. After equating, these examinees obtained an equated score of 39.68 which was rounded to the next integer 40, and thus, reaching the 1SD cut-score (40). That is, the 96 examinees with the reported score at 39 before equating have increased one score-point after equating the translated test. These examinees were then re-classified into the sample above 1SD, and thus, there is an

Table 6

Evaluation of the Classification of Examinees in Meeting the Standards in the Six Translated Tests

Translated Test	Differences in the Number of Examinees Classification							
	Exploratory DIF – No DIF				Confirmatory DIF – No DIF			
	Tu	Eq	Le	IRT	Tu	Eq	Le	IRT
Grade 6 Social Studies (N=2,115)								
Sample below -1SD	0	0	0	0	0	-49	0	-49
Sample between -1SD and mean	0	0	0	0	-102	-53	0	+49
Sample between mean and 1SD	-96	-96	0	0	+6	+6	0	0
Sample above 1SD	+96	+96	0	0	+96	+96	0	0
Total number of re-classifications	**96**	**96**	**0**	**0**	**198**	**247**	**0**	**49**
Grade 9 Social Studies (N=2,115)								
Sample below -1SD	0	0	0	0	0	0	0	0
Sample between -1SD and mean	0	0	0	0	0	0	0	0
Sample between mean and 1SD	0	0	0	0	0	0	0	0
Sample above 1SD	0	0	0	0	0	0	0	0
Total number of re-classifications	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Grade 6 Mathematics (N=2,115)								
Sample below -1SD	0	0	0	0	0	0	0	0
Sample between -1SD and mean	0	0	0	0	0	0	0	0
Sample between mean and 1SD	0	-118	0	0	0	-118	0	0
Sample above 1SD	0	+118	0	0	0	+118	0	0
Total number of re-classifications	**0**	**118**	**0**	**0**	**0**	**118**	**0**	**0**
Grade 9 Mathematics (N=2,115)								
Sample below -1SD	0	0	0	0	0	0	0	0
Sample between -1SD and mean	0	0	0	0	0	0	0	0
Sample between mean and 1SD	0	0	0	0	0	0	0	0
Sample above 1SD	0	0	0	0	0	0	0	0
Total number of re-classifications	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Economics (N=2,000)								
Sample below -1SD	0	+96	0	0	0	+96	0	0
Sample between -1SD and mean	0	-217	0	+107	0	-217	0	0
Sample between mean and 1SD	-67	+54	+57	-50	-67	+54	0	0
Sample above 1SD	+67	+67	-57	-57	+67	+67	0	0
Total number of re-classifications	**67**	**284**	**57**	**164**	**67**	**284**	**0**	**0**
History (N=2,000)								
Sample below -1SD	0	0	0	0	0	0	0	0
Sample between -1SD and mean	0	0	0	0	0	0	0	0
Sample between mean and 1SD	0	0	0	0	0	0	0	0
Sample above 1SD	0	0	0	0	0	0	0	0
Total number of re-classifications	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Note. Differences in the number of examinees classification were calculated using either the number of examinees classified in Exploratory or Confirmatory DIF condition minus the number of examinees classified in the No DIF condition.

Total number of re-classifications is the number of examinees who were re-classified into a different standard of performance after equating.

Tu= Tucker linear equating, Eq= equipercentile equating, Le= Levine observed score equating, and IRT=IRT observed score equating.

increase in 96 examinees in the sub-sample after equating. Consequently, 96 out of 211 (4.5 %) French examinees were re-classified into a different standard after Tucker linear equating in the exploratory DIF condition.

A similar situation occurred for the equipercentile equating in the exploratory DIF condition. Examinees (n = 96) with a reported score of 39 before equating were initially classified into the sample between the mean and 1SD. These examinees obtained an equated score of 39.90 in equipercentile equating and gained one score-point after the equated score was rounded up to 40 in the translated test, and thus, they were re-classified into the sample above 1SD. Therefore, 96 out of 2115 (4.5 %) French examinees were re-classified into a different standard after the equipercentile equating in the exploratory DIF condition.

The examination of the classifications of examinees in the confirmatory DIF condition in the Grade 6 Social Studies Test revealed differences in the number of examinees in three sub-samples using Tucker linear equating method: the sample between minus 1SD and mean (n = -102), the sample between mean and 1SD (n = +6), and the sample above 1SD (n = +96) (see Table 6). Examinees (n = 102) with a reported score of 31 before equating (one-score point below the mean at 32) obtained an equated score of 31.57 and gained one score-point after the equated score is rounded up to 32 in the translated test. Consequently, the 102 examinees were re-classified into the sample between the mean and 1SD. Examinees (n = 96) with a reported score of 39 before equating (one-score point below the 1SD cut-score) have an increase in one score-point after equating the translated test. The 96 examinees were re-classified into the sample above 1SD. Consequently, 198 out of 2115 (9.4 %) French examinees are re-classified

into different sub-sample using Tucker linear equating in the confirmatory DIF condition. As a result, the net difference in the number of examinees re-classification for the sample between mean and 1SD before and after equating remains as six.

Using equipercentile equating, the results revealed differences in the number of examinees in four sub-samples in the confirmatory DIF condition (see Table 6). Before equating, 49 examinees were scored at 23 (one score below the -1SD), 102 examinees were scored at 31 (one score below the mean), and 96 examinees were scored at 39 (one score below the -1SD). After equating, these examinees gained one score-point and were placed in the next higher standard. Consequently, 247 out of 2115 (11.7 %) French examinees were re-classified into different standards after the equipercentile equating in the confirmatory DIF condition. The net difference in the numbers of examinees classification for the sample between -1SD and mean is –53 and for the sample between mean and 1SD is six.

Using IRT equating, results revealed that there were differences in the number of examinees in two sub-samples in the confirmatory DIF condition: the sample below -1SD (n = -49) and the sample between -1SD and mean (n = +49) (see Table 6). Differences are attributed to the 49 examinees with their reported scores at one score-point below the -1SD cut-score before equating who gained one score-point after equating, and were re-classified into the sample between -1SD and mean. Consequently, 49 out of 2115 (2.3 %) French examinees were re-classified into a different standard after IRT equating in the confirmatory DIF condition.

Grade 9 Social Studies

The three cut-scores before equating were determined at the score points of 31 (-1SD), 39 (mean), and 47 (+1SD) in the Grade 9 Social Studies Test. The results from either the exploratory and confirmatory DIF versus the No DIF condition comparisons revealed that all the differences are zero (see Table 6). It indicates that the classification of examinees remained unchanged in both DIF conditions.

Grade 6 Mathematics

The three cut-scores for the Grade 6 Mathematics Test before equating were determined at the score points of 29, 37, and 44. For the exploratory versus No DIF comparisons, the results reveal that there were differences in the number of examinees in two sub-samples when using the equipercentile equating method. The 118 examinees with the reported score at 43 before equating, which were one-score point below the 1SD cut-score (44), were categorized into the sample between the mean and 1SD. These examinees obtained an equated score of 43.53 in equipercentile equating and gained one score-point after the equated score was rounded up to 44 in the translated test. Therefore, 118 out of 2115 (5.6 %) French examinees were re-classified into a different sample after the equipercentile equating in the exploratory DIF condition. The same results were obtained for the confirmatory versus No DIF condition comparisons.

Grade 9 Mathematics

In the Grade 9 Mathematics Test, the three cut-scores before equating were determined at score points 25, 34, and 42. The results from the exploratory and confirmatory DIF conditions revealed that no examinees were re-classified after equating (see Table 6).

Economics

In the Economics Test, the three cut-scores before equating were determined at score points 16, 23, and 30. The results from the exploratory DIF condition comparisons revealed differences in the number of examinees in two sub-samples when using Tucker linear equating method: the sample between the mean and 1SD (n = -67), and the sample above 1SD (n = +67). The 67 examinees with a reported score of 29 before equating were originally classified in the sample between mean and 1SD. They gained one score point after equating the translated test and consequently they were re-classified into the sample above 1SD. That is, 67 out of 2000 (3.4 %) Chinese examinees were re-classified into a different sample using the Tucker linear equating in the exploratory DIF condition.

Using equipercentile equating, the results revealed differences in the number of examinees in all sub-samples in the exploratory DIF condition: the sample below -1SD (n = +96), the sample between -1SD and the mean (n = -217), the sample between the mean and 1SD (n = +54), and the sample above 1SD (n = +67) (see Table 6). Before equating, 96 examinees were scored at 16 (cut score at -1SD), 121 examinees were scored at 22 (one score below the mean), and 67 examinees were scored at 29 (one score below the -1SD). After equating, the 96 examinees who scored 16 lost one score-point and were re-classified to the sample below -1SD. However, the 121 examinees who scored 22 and 67 examinees who score 29 gained one score-point and therefore, were re-classified to the next higher standard. Consequently, 284 out of 2000 (14.2 %) Chinese examinees were re-classified into different standards after equipercentile equating in the exploratory DIF condition.

Using Levine observed score equating, the results revealed that there were differences in the number of examinees in two sub-samples: the sample between the mean and 1SD (n = +57), and the sample above 1SD (n = -57). The 57 examinees with a reported score at 30 before equating obtained an equated score of 29.45 and a rounded score 29. These 57 examinees lost one score-point after equating and they were re-classified into the sample between the mean and 1SD. Therefore, 57 of 2000 (2.9 %) Chinese examinees are re-classified into a different standard after Levine observed score equating in the exploratory DIF condition.

Using IRT observed score equating, the results revealed that there were differences in the number of examinees in three sub-samples: the sample between -1SD and mean (n = +107), the sample between mean and 1SD (n = -50), and the sample above 1SD (n = -57). Before equating, 107 examinees were scored at 23 (cut-score at the mean) and 57 examinees were scored at 30 (cut-score at 1SD). After equating, both the 107 examinees and the 57 examinees lost one score-point and were re-classified to the next lower standards respectively. Therefore, 164 out of 2000 (8.2 %) Chinese examinees were re-classified into different standards after IRT observed score equating in the exploratory DIF condition.

For the confirmatory versus No DIF conditions, the results revealed that there are differences in the number of examinees using Tucker linear equating and equipercentile equating methods. The results are similar to the results for the exploratory DIF condition, except that no re-classifications occurred using Levine and IRT observed score equating methods. Consequently, 67 (3.4 %) examinees for the Tucker linear equating and 284

(14.2 %) examinees for equipercentile equating were re-classified into different standards in the confirmatory DIF condition.

History

The three cut-scores before equating were determined at 14, 21, and 29. In both the exploratory and confirmatory DIF conditions, the results revealed that all of the differences were zero (see Table 6). The classification of examinees remains unchanged in both DIF conditions.

To summarize, the classification of examinees in the six translated tests were evaluated. The results revealed that in three of the tests, the classification of examinees in the exploratory and confirmatory DIF conditions produced important differences. These translated tests are the Grade 6 Social Studies Achievement Test, the Grade 6 Mathematics Achievement Test, and the Economics HKCEE. For the other three translated tests – the Grade 9 Social Studies Achievement Test, the Grade 9 Mathematics Achievement Test, and the History HKCEE - the classification in the exploratory and confirmatory DIF conditions remained unchanged in either DIF condition.

Section III: Summary

Six data sets were used in the present research: the Grade 6 Social Studies Achievement Test (English-French), the Grade 9 Social Studies Achievement Test (English-French), the Grade 6 Mathematics Achievement Test (English-French), the Grade 9 Mathematics Achievement Test (English-French), the Economics HKCEE (English-Chinese), and the History HKCEE (English-Chinese). Three steps were used to analyze the data: (1) identify DIF and common items in each DIF condition, (2) equate the translated achievement tests onto a common scale using the common-item

nonequivalent groups equating design, and (3) evaluate the effect of DIF on the equated translated tests. Evaluations focused on the common-item sets, the equated scores, and the classifications of examinees in meeting the standards.

To begin, three criteria were used to evaluate whether or not the common-item sets were sufficient for equating. The common-item set should: (1) have 15 to 20 common items and be at least 20% of the full-length test (test with 40 or more items) (Angoff, 1971; Kolen & Brennan, 2004; Wainer, 1999), (2) have the test specification proportions that approximate the corresponding proportions for the full-length tests (Kolen and Brennan, 2004), and (3) have a high correlation between the common-item scores and the full-length test scores (Beguin, 2002; Budescu, 1985). The results in the present research revealed that the three evaluation criteria were met by the common-item sets from the six data sets and, therefore, the common-item sets were adequate for equating (see Tables 2, 3 and 4).

Second, to evaluate the relative magnitude for the unrounded equated scores, the "Difference that matters" (DTM) score at 0.50 was used (Kolen & Brennan, 2004; Dorans, 2004; Dorans et al., 2003). When the mean differences between the unrounded equated score and the non-equated score was greater than or equal to 0.50 (i.e., the DTM score) it was considered to be an important change in the equated result. Further, to evaluate the equating results along the score scale across the three DIF conditions, the samples in each data set were categorized into four sub-samples according to three cut-scores in each translated test score distribution before equating: below 1 standard deviation (SD), between minus 1 SD and mean, between mean and plus 1 SD, and above 1 SD.

Results from the present study revealed that the equated scores in the exploratory and confirmatory DIF conditions have important changes for four of the six tests (see Table 5). These four tests included the Grade 6 Social Studies Achievement Test, the Grade 6 Mathematics Achievement Test, the Grade 9 Mathematics Achievement Test, and the Economics HKCEE. Moreover, important score changes in the equated scores occurred most frequently when Tucker linear equating and equipercentile equating were used; few changes occurred in Levine and IRT observed score equating. For the other two translated tests, the Grade 9 Social Studies Achievement Test and the History HKCEE, the equated scores in the exploratory and confirmatory DIF conditions revealed no important changes in either DIF condition.

A summary of the number of important mean score changes and the re-classification of the examinees occurred for all data sets across four equating methods is presented in Table 7. The results revealed that when exploratory DIF condition was compared with No DIF condition, the total number of mean score changes occurred five times for Tucker linear equating, five times for equipercentile equating, two times for Levine observed score equating, and two times for IRT equating. For the confirmatory DIF condition versus No DIF comparison, the number of changes occurred six times for Tucker linear equating, seven times for equipercentile equating, one time for Levine observed score equating, and two times for IRT equating. These results indicated the equating outcomes varied across different DIF conditions, different equating methods, and different tests.

Third, to evaluate the classification of examinees, comparisons were made between the numbers of examinees who wrote the translated tests and who met the

Table 7

Summary for the Total Number of Mean Score Changes and the Re-classification of Examinees across Four Equating Methods in the Six Translated Tests

Translated Test	Total Number of Mean Score Changes							
	Exploratory DIF – No DIF				Confirmatory DIF – No DIF			
	Tu	Eq	Le	IRT	Tu	Eq	Le	IRT
Grade 6 Social Studies	3*	1*	0	0	3*	4*	0	1*
Grade 9 Social Studies	0	0	0	0	0	0	0	0
Grade 6 Mathematics	0	1*	0	0	1	1*	0	1
Grade 9 Mathematics	0	0	1	0	0	0	1	0
Economics	2*	3*	1*	2*	2*	2*	0	0
History	0	0	0	0	0	0	0	0
Total	5	5	2	2	6	7	1	2

Note. Total number of mean score changes are indicated when the mean score difference, which was calculated by either the mean score from exploratory or confirmatory DIF condition, minus the mean score from the No DIF condition ≥ 0.50.

Tu= Tucker linear equating, Eq= equipercentile equating, Le= Levine observed score equating, and IRT=IRT observed score equating.

* It indicated that changes occurred in the re-classification of examinees after equating.

standards before and after equating the translated tests. The number of examinees re-classified to various standards of performance provides an important indication for evaluating the impact of equating. Based on three cut-scores established before equating, the number of examinees meeting each standard was compared before and after equating using the rounded equated scores. Comparisons were made between the DIF conditions (either exploratory or confirmatory) and the No DIF condition. The results are presented in Tables 6 and 7.

The results revealed that in three of the six tests, the classification of examinees in the exploratory and confirmatory DIF conditions changed. These translated tests were the Grade 6 Social Studies Achievement Test, the Grade 6 Mathematics Achievement Test, and the Economics HKCEE (see Table 7). The results revealed that when exploratory DIF condition was compared with No DIF condition, the re-classification of

examinees occurred in two tests for Tucker linear equating, three tests for equipercentile equating, one test for Levine and IRT observed score equating. For the confirmatory DIF condition versus No DIF comparison, the classification of examinees occurred in two tests for Tucker linear equating, three tests for equipercentile equating, none for Levine observed score equating, and one test IRT equating equating. For the other three translated tests, the Grade 9 Social Studies Achievement Test, the Grade 9 Mathematics Achievement Test, and the History HKCEE, the classification in the exploratory and confirmatory DIF conditions remained unchanged in either DIF condition. A discussion and conclusions for the current research are presented in the next chapter.

CHAPTER V DISCUSSION AND CONCLUSIONS

This chapter includes a discussion of the findings and the conclusions of the present research. It is organized into three sections. The first section is a summary of the research questions and a brief description of the methods in the study. The second section provides a discussion of the present research findings. It includes a summary and discussion of the key findings, the limitations of the present study, and the conclusions generated from the present research. The third section provides the recommendations for future practice and research.

Section I: Summary of Research Questions and Methods

Translated tests are commonly used to compare performance between different language groups. However, translation is rarely perfect. It is a challenge to compare student achievement across different language groups as performance differences may be due to test differences, group differences, or both. To disentangle the test differences from the group differences, equating can be used to place the two language tests on to a common scale. The common-item nonequivalent group equating design is commonly used to equate different language forms. The selection of common items is crucial for this equating design. Traditionally, the outcomes of statistical differential item functioning (DIF) analyses have been used to identify the common items for equating across different language tests. However, statistical DIF analyses cannot be used to explain why items are functioning differentially between the groups. It is important to identify the source of DIF for all the DIF items in order to select the appropriate common items for equating since the interpretability of the equating results using translated tests is often related to the selection of the common-item set.

Roussos and Stout (1996) proposed the multidimensionality-based DIF analysis paradigm, which uses a two-stage confirmatory approach, to detect DIF items and understand why DIF occurs. However, no study has been conducted to apply the outcomes from the multidimensionality-based DIF analysis paradigm for selecting the common items when equating translated tests and to evaluate the equated translated test scores.

The purpose of this study was to evaluate the effect of using the multidimensionality-based DIF analysis paradigm to select the common items used on the common-item nonequivalent group equating design using translated achievement tests. Three DIF conditions were created for comparison: (1) No DIF, (2) Exploratory DIF condition, and (3) Confirmatory DIF condition. Four equating methods were used: Tucker linear equating, Levine observed score equating, equipercentile equating, and item response theory observed score equating. Six data sets were analyzed: the Grade 6 Social Studies Achievement Test, Grade 9 Social Studies Achievement Test, Grade 6 Mathematics Achievement Test, and Grade 9 Mathematics Achievement Test administered in Alberta, and the Economics HKCEE, and History HKCEE administered for Secondary 5 (Grade 12) in Hong Kong.

Three research questions were addressed in the present research:

1. By comparing the DIF (exploratory and confirmatory) and the No DIF conditions, are there any differences between the equated and non-equated scores? Do these scores affect the classification of examinees in meeting the standards set for the translated tests?

2. By comparing the equating results from the exploratory and confirmatory DIF conditions, are there any differences in the common-item sets, equated scores, and the classification of examinees?
3. By comparing the equating results across different languages and subject areas, are there differences in the common-item sets, equated scores, and the classification of examinees?

To answer the research questions, three DIF conditions were compared throughout the analyses: (1) No DIF, (2) Exploratory DIF condition, and (3) Confirmatory DIF condition. For the No-DIF condition, it was assumed that either there was no DIF item or the presence of DIF items was ignored in the translated tests. No equating was performed in the No DIF condition and it was used as the control condition (before equating). The exploratory and confirmatory DIF conditions were used to detect DIF items in each data set. Then, results from the exploratory and confirmatory DIF conditions were used to develop the common-item sets for equating. Equating was performed using the four equating methods on all six data sets.

After equating the translated achievement tests on to the original test scale using the common-item nonequivalent groups equating design, results from the exploratory and confirmatory DIF conditions were compared with the No DIF condition to determine the effect of DIF on the equated translated tests. To make the comparisons consistent, evaluations were focused on the common-item sets, the equated scores of the translated tests, and the classification of examinees in meeting standards of performance.

To determine whether or not the common-item sets were sufficient for equating purpose, three criteria were used: (1) the common-item set should have 15 to 20 common

items and be at least 20% of the full-length test (test with 40 or more items) (Angoff, 1971; Kolen & Brennan, 2004; Wainer, 1999), (2) the common-item sets should have test specification proportions that approximate the corresponding proportions for the full-length tests (Kolen and Brennan, 2004), and (3) the scores from the common-item sets should have a high correlation with the full-length test scores (Beguin, 2002; Budescu, 1985). Therefore, the number of common items, the proportions (percentages) of the common items compared to the corresponding proportion for the full-length test, and the correlations of the examinees scores between the common-item sets and the full-length tests in each data set were evaluated.

To evaluate the relative magnitude for the equated scores, the "Difference that matters" (DTM) score at 0.50 was used (Kolen & Brennan, 2004; Dorans, 2004; Dorans et al., 2003). When the mean score difference between the unrounded equated and non-equated score was greater than or equal to 0.50, it was considered to be a DTM and an important change in the equating results. Moreover, to evaluate the equating results along the score scale across the three DIF conditions, the samples for each translated test were classified into four sub-samples based on three cut-scores in the translated tests before equating. These four sub-samples were below 1 standard deviation (SD), between minus 1 SD and mean, between mean and plus 1 SD, and above 1 SD. By comparing the unrounded equated mean scores of the sub-samples from the DIF condition (either exploratory or confirmatory) and the No-DIF condition, the differences in the mean equated scores for each equating method were evaluated.

To evaluate the classification of examinees using the translated tests, comparisons were made between the number of examinees who met the standards before and after

equating the translated tests. A difference in the number of examinees classified according to the standards of performance provides an important indication for evaluating the impact of equating. Based on three cut-scores established before equating, the numbers of examinees re-classified after equating the translated tests were calculated using the rounded equated scores. Comparisons were made between the DIF conditions (either exploratory or confirmatory) and the No DIF condition.

Section II: Discussion

Summary of Findings

The summary statistics for the six data sets for the original and translated tests were presented in Table 1 in the previous chapter. The numbers of examinees in all tests was at least of 2,000 which are adequate to produce stable equating outcomes (Harris, 1993; Harris & Crouse, 1993; Kolen & Brennan, 2004). Kolen and Brennan (2004) suggested that when the mean difference between the two groups to be equated was less than 0.5 standard deviation units, and the ratio of the group's standard deviations was less than 1.2, then the two groups were similar in ability. Comparisons were made for the mean scores of the examinees taking the original and translated tests and the ratios of the standard deviation. These results indicated the ability differences between the two groups of examinees were small for the six tests. Moreover, the reliabilities (internal consistency) of the original and the translated tests were similar, with Cronbach's alpha values between 0.84 and 0.92. Taken together, these results revealed that the two groups taking the original and the translated tests were of similar ability, and the tests possessed similar characteristics. Two advantages occur when the two groups are of similar ability: it enhances the accuracy for DIF detection (Hambleton, 1993) and it makes the groups

more suitable for common-item nonequivalent group equating (Kolen & Brennan, 1995; Raff & Allalouf, 2002).

The common-item sets, the equated scores, and re-classifications of examinees for each data set were evaluated. First, according to the results reported in Tables 2 to 4, all common-item sets in the six data sets fulfilled the three evaluation criteria for the common-item set. Thus, the common-item sets were adequate for equating.

Second, the results from the present research revealed that the equated mean scores in the exploratory and confirmatory DIF conditions have important score changes for four of the six tests (see Tables 5 and 7). These translated tests were the Grade 6 Social Studies Achievement Test, Grade 6 Mathematics Achievement Test, Grade 9 Mathematics Achievement Test, and the Economics HKCEE. The results revealed that when the exploratory DIF condition was compared with the No DIF condition, the total number of mean score changes occurred five times for Tucker linear equating, five times for equipercentile equating, two times for Levine observed score equating, and two times for IRT equating. For the confirmatory DIF condition versus the No DIF comparison, the number of changes occurred six times for Tucker linear equating, seven times for equipercentile equating, one time for Levine observed score equating, and two times for IRT equating. For the other two translated tests - the Grade 9 Social Studies Achievement Test and the History HKCEE - the equated scores in the exploratory and confirmatory DIF conditions revealed no important changes in either DIF condition.

Third, the results revealed that in three of the six tests, the classification of examinees in the exploratory and confirmatory DIF conditions changed (see Tables 6 and 7). These translated tests were the Grade 6 Social Studies Achievement Test, Grade 6

Mathematics Achievement Test, and the Economics HKCEE. The results revealed that when the exploratory DIF condition was compared with No DIF condition, the re-classification of examinees occurred in two tests for Tucker linear equating, three tests for equipercentile equating, and one test each for Levine and IRT observed score equating. For the confirmatory DIF condition versus No DIF comparison, the re-classification of examinees occurred in two tests for Tucker linear equating, three tests for equipercentile equating, none for Levine observed score equating, and one test for IRT equating. For the other three translated tests – the Grade 9 Social Studies Achievement Test, Grade 9 Mathematics Achievement Test, and History HKCEE - the classification in the exploratory and confirmatory DIF conditions remained unchanged in either DIF condition.

Discussion of Findings

To answer the three research questions in the present study, the findings from the common-sets, the equated scores, and the re-classification of examinees evaluations are interpreted.

Q1. By comparing the DIF (exploratory and confirmatory) and the No DIF conditions, are there any differences between the equated and non-equated scores? Do these scores affect the classification of examinees on the translated tests?

The answer is *Yes*. DIF items were identified in all six data sets in the exploratory and confirmatory DIF conditions. The results revealed that the translated tests were not equivalent to the original tests, as test differences existed. To remove the test differences, the two language tests needed to be placed on a common scale. Equating was conducted in the exploratory and confirmatory DIF conditions using four

equating methods. Then, the equating results from the exploratory and confirmatory conditions were compared to the No-DIF condition (no equating) to evaluate the effect of DIF on the equating outcomes. By evaluating the unrounded equated scores, the results across the four equating methods indicated that there were differences between the equated and non-equated mean scores, and important changes in mean scores (greater than DTM score) occurred in four data sets. Moreover, results from the rounded equated scores revealed that those scores affected the re-classification of examinees on three of the four data sets.

Important score changes in the equated mean scores and the re-classification of examinees occurred mostly when Tucker linear equating and equipercentile equating were used; few changes occurred in Levine and IRT observed score equating. In sum, by comparing the DIF and the No DIF conditions, the results across the four equating methods revealed that there were differences between the equated and non-equated scores, and the impact on the translated tests depended on the selection of the equating methods used and the DIF condition.

Often, when there were differences between the equated and non-equated mean scores, those scores also affected the re-classification of the examinees for the same test. When the mean score differences between the equated and the non-equated scores are greater than or equal to 0.5 (DTM score), it is attributed to the summation of the score differences in the sub-sample examinees, which indicates that many examinees along the score scale have important changes in scores. It is likely that the classifications of examinees in meeting the standards after equating are affected.

However, differences between the equated and the non-equated mean scores may occur without affecting the classification of examinees. It can be explained by the location of the rounded equated score at the cut-score. For example, in the Grade 9 Mathematics test, the unrounded mean score difference for the sub-sample below -1SD (score range from 0 to 24) when using Levine observed score equating was 0.53 (see Table 5), which was consider an important change. Examinees with a reported score at 24 before equating obtained an equated score of 24.47 and that was rounded to 24 after equating. Consequently, the rounded equated score at the cut-score was the same as the original test score, and thus, the re-classification of examinees remained unchanged. Kolen and Brennan (2004) suggested that in many practical situations, the scores that make the *difference that matters* depend on where the score is on the scale when the decision is made. Results from the present research demonstrated that it is important to evaluate both the unrounded and rounded equated scores.

The important score changes in the unrounded equated mean scores also occurred frequently on the extreme score groups, e.g., the subgroup sample above 1SD (See Table 5). The explanation may relate to the distribution of the standard error of equating that typically differ across different score ranges: the standard errors are high at the extreme ends and low at the median range of the score (Kolen & Brennan, 2004). However, as the sample size increases, the standard error approaches zero (Angoff, 1971; Harris, 1993; Harris & Crouse, 1993). In the present study, the sample size for each language group contained a minimum of 2,000 examinees, therefore, it was adequate to produce stable equating results at the middle of the score scale. The results were more variable at

the end of the score scale because the sample size was smaller (Harris, 1993; Harris & Crouse, 1993; Kolen & Brennan, 2004).

Results from this study revealed differences between the unrounded equated and non-equated mean scores in four of the six tests. This outcome raises an important question: Is it necessary to conduct equating when comparing performances in different language groups? Kolen and Brennan (2004) suggested that "no linking" is one alternative where the quality of the translation bears the linking burdens such as time and cost. No equating can be an alternative. However, results from the present research showed that the classification of examinees varied in three of the six tests. These tests were high-stakes examinations and the consequences of the test uses were related to meeting standards and placements decisions. The decisions of whether or not to equate may depend on the consequences of the test use.

The results from the present research also showed that two tests - the Grade 9 Social Studies and the History HKCEE tests- have no important changes in the equated scores and the classification of examinees no matter which equating methods were used. The reasons why the presence of DIF in the translated tests had no effect on the equating outcomes are not known. However, several possible explanations exist. These two tests had a smaller number of DIF items, and thus, a higher percentage of common-items (60.5 % to 97.4 %) compared to the four tests (42.0 % to 94.0 %) where important changes occurred in the equated scores and the classification of examinees (see Table 2). Thus, the effect of DIF on equating outcomes may depend on the amount of DIF present in the tests.

Q2. By comparing the equating results from the exploratory and confirmatory DIF conditions, are there any differences in the common-item sets, equated scores, and the classification of examinees?

The answer is *Yes*. First, comparisons were made between the exploratory and confirmatory DIF conditions across the common-item sets. There are differences in the common-item sets between the exploratory and confirmatory DIF conditions. All the common-item sets across the six tests met the evaluation criteria and they were sufficient for equating purpose in both DIF conditions. However, the exploratory DIF condition retained fewer items in the common-item sets (42 % to 86 %) when compared with the confirmatory condition (61.1 % to 94 %) (see Table 2). Thus, the confirmatory condition may yield better content and statistical representativeness for the common items compared to the exploratory condition (see Table 3). The corresponding correlations between the common-item set scores and the full-length test scores in the confirmatory condition were also higher than that in the exploratory condition in four tests and the correlations were the same in two tests (see Table 4). Therefore, the confirmatory DIF condition which used the multidimensionality-based DIF analysis paradigm is a suitable approach to select common items.

Second, comparisons were made between the exploratory and confirmatory DIF conditions for the equated scores. The results revealed differences in the total number of important changes in the equated mean scores in four tests. The important changes in the equated mean scores in the exploratory DIF condition tended to occur less often than that in the confirmatory DIF condition when the same equating method was compared (see Tables 5 and 7). The majority of these changes occurred in two of the four equating

methods: the Tucker linear equating and equipercentile equating. For example, the total number of score changes for the equipercentile equating occurred five times in the exploratory DIF condition and seven times in the confirmatory DIF condition.

Moreover, some of these important score changes occurred in the exploratory DIF condition but not in the confirmatory DIF condition and the vice versa when the same equating method was used for the same test. Using IRT observed score equating, important score changes occurred two times in the Economics test only for the exploratory condition and not in the confirmatory condition; whereas changes occurred two times in the Grade 6 Social Studies test and the Grade 6 Mathematics test only for the confirmatory condition and not for the exploratory condition (see Tables 5 and 7).

Third, comparisons were made between the exploratory and confirmatory DIF conditions for the re-classification of examinees. There are differences in the number of examinees re-classified between the exploratory and the confirmatory DIF conditions when the same equating method was used. For example, for the Grade 6 Social Studies test with Tucker linear equating method, 96 examinees were re-classified into a different standard in the exploratory condition whereas 198 examinees were re-classified in the confirmatory DIF condition (see Table 6).

Some of the re-classification differences of examinees occurred either in the exploratory or confirmatory DIF condition when the same test and the same equating method was used, as with the situation for the important equated score changes. An example occurred in the Economics test when using the Levine equating method where 57 examinees were re-classified in the exploratory condition, but the classification of examinees remained unchanged in the confirmatory DIF condition.

In sum, by comparing the exploratory and confirmatory DIF conditions, the results revealed differences in the common-item sets, equated scores, and the classification of examinees. Differences in the equating outcomes from the exploratory and the confirmatory DIF conditions include the total number of important changes in the equated mean scores, the total number of examinees re-classified into a different standard, and the test in which these changes occurred. Moreover, these differences are also affected by the selection of the equating methods. These differences are attributed to the different sets of common items that are identified in the exploratory and confirmatory DIF conditions. Despite of the high degree of overlap between the two common-item sets, which ranged from 62.2 % to 100 %, the differences did lead to different equating results, except for the Grade 9 Mathematics which have the same common-item sets. Among the confirmatory and exploratory DIF conditions, which equating results should be adopted?

The choice of the equating outcomes may be related to the interpretability of the equating results. Equating is a statistical process to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen & Brennan, 2004). Due to the unintended effects of translation, test differences occur between the translated tests and the original tests (Sireci, 1997, Wainer, 1999). The purpose of equating the translated tests on to the original score scale is to adjust the scores so that test differences between the translated tests and the original tests are removed. Then, performance differences between the language groups are attributed to group ability differences after equating. Thus, the equated scores from the translated tests can be interpreted as interchangeable with the original test scores. However, the equating outcomes are crucially affected by

the selection of the common-item set when the common-item nonequivalent groups equating is applied.

The exploratory and confirmatory DIF conditions used different approaches to select the common-item sets. The exploratory condition used the statistical DIF analyses to identify DIF items but it did not identify the sources of the DIF items. All DIF items, may the DIF items be attributed to translation errors or group ability differences, were removed from developing the common-item sets. These common-item sets were then used to equate the translated tests to the original test scale. The translated test scores were adjusted after equating, in which test differences and possibly or group ability differences between the tests were removed. That is, the equating outcomes in the exploratory DIF condition may have included translated test scores adjustments due to translation errors and also group ability differences that the tests are intended to measure. Without knowing the sources of DIF, it is difficult to interpret the equating results.

The confirmatory DIF condition used the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996) to identify the DIF items. The first stage involves substantive analyses to generate DIF hypotheses, and the second stage is statistical analyses to test the hypotheses. Therefore, the confirmatory DIF conditions identified the sources of translation DIF in the items, and only statistically confirmed items were identified as DIF. Thus, only items identified with translation errors were removed from developing the common-item sets. These common-item sets were used to equate the scores of the translated tests onto the original scale. The translated test scores were adjusted after equating and test differences between the language tests were removed. That is, the equated outcomes from the confirmatory DIF condition are interpreted as the

score adjustments of the translated tests due only to translation errors. Thus, performances differences between the groups are attributed to group abilities differences.

Moreover, the results in the present research revealed that the common items identified in the exploratory and confirmatory DIF conditions overlapped from 62.2 % to 100 % in six tests. Both DIF conditions identified the same set of common items (100 %) in the Grade 9 Mathematics test. Consequently, the sources of DIF for the items in this case were attributed to translation errors. In the five remaining tests, the exploratory condition tended to identify more DIF items than the confirmatory condition but the sources of the DIF were not known. In contrast, the confirmatory condition yielded information about the sources of DIF due to translation error, which enhances the interpretability of the equated outcomes. Moreover, the confirmatory condition may yield a better content and statistical representativeness for the common items compared to the exploratory condition.

Therefore, by comparing the exploratory and the confirmatory DIF conditions, the results from the common-item sets evaluation and the equating outcomes interpretability revealed that the confirmatory DIF condition has more advantages than the exploratory condition in selecting common items and interpreting the equated scores. Consequently, the multidimensionality-based DIF analysis paradigm is a suitable approach for DIF and common items detection and interpretation when equating translated tests.

Q3. By comparing the equating results across different languages and subject areas, are there differences in the common-item sets, equated scores, and the classification of examinees?

To conduct the comparisons across different languages, evaluations were made between the four French and two Chinese translated tests across the common items, equated scores, and the classification of examinees. The four French translated tests were the Grade 6 and the Grade 9 Social Studies tests, and the Grade 6 and the Grade 9 Mathematics tests. The two Chinese translated tests were the Economics and the History tests. Then, the comparisons across different subject areas are followed.

First, the common-item sets across the French and Chinese tests were compared. The answer is *No, the common-items sets characteristics across different language tests are similar.* The percentages of common items as compared to the full-length tests in the French and Chinese translated tests are similar (see Table 2). The French tests ranged from 42.0 % to 86.0 % in exploratory DIF condition and ranged from 60.0 % to 94.0 % in the confirmatory DIF condition. The Chinese tests ranged from 44.4 % to 60.6 % in exploratory DIF condition and ranged from 61.1 % to 97.4 % in the confirmatory DIF condition.

Moreover, the correlations of the common items and the full-length tests scores in the French and Chinese translated tests were also similar. The correlations in the exploratory conditions are lower than that in the confirmatory condition for the corresponding tests in two French tests –the Grade 6 and Grade 9 Social Studies tests - and two Chinese tests. The other two French tests – the Grade 6 and Grade 9 Mathematics tests - have the same correlations in the exploratory and confirmatory condition. The correlations for the French tests ranged from 0.91 to 0.99 in exploratory DIF condition and from 0.95 to 0.99 in the confirmatory DIF condition (see Table 4). The

correlations for the Chinese tests ranged from 0.86 to 0.96 in exploratory DIF condition and from 0.93 to 0.99 in the confirmatory DIF condition.

Second, the equated scores and the classification of examinees for the French and Chinese tests were compared. The answer is *No.* The results from the study revealed important changes occurred in the equated scores and classification of examinees, regardless of whether the tests were translated into French or Chinese. Important changes in the equated mean scores occurred in three French tests - the Grade 6 Social Studies test, Grade 6 Mathematics test, and Grade 9 Mathematics test - and one Chinese test – the Economics test (see Tables 5 and 7). These equated score changes also affected the re-classification of the examinees in these translated tests except the Grade 9 Mathematics test.

Third, the equated results were then compared across different subject areas. The answer is *Yes.* The results revealed that there were differences in the equated scores and the re-classification of examinees occurred in the translated tests across different subject areas. However, the differences were not consistent when tests in the same subject area were compared across different grades. For example, when using Tucker and equipercentile equating, changes in the equated mean scores and re-classification of examinees occurred in the verbal tests (e.g., the Grade 6 Social Studies) more than in the nonverbal tests (e.g., the Grade 6 Mathematics) (see Table 7). However, these changes may depend on the DIF condition and the use of the equating method. For example, the important changes occurred in the Grade 6 Social Studies test and Grade 6 Mathematics test when using the IRT equating in the confirmatory condition, and these changes did not occur in their corresponding Grade 9 tests. The reason why the Grade 9 Social

Studies test and the History HKCEE test did not show any differences or changes in the equating outcomes was not known. It may be related to the specific subject areas, the sample differences, and the amount of DIF presented in the tests, and the selection of equating methods.

In sum, by comparing across different languages, the results from the common-item sets and the equating outcomes across the French and Chinese tests were similar. The results revealed that important changes occurred in the equated mean scores and the re-classification of examinees, regardless of whether the tests were translated into French or Chinese. Previous studies suggest that the cultural distance between the language and ethnic groups may be related to the adverse effects of item bias (Rogers, 2002, cited in Puhan, 2003). Results from the present research did not find these differences. However, examinees who took the French tests are from Canada and examinees who took the Chinese tests are from Hong Kong. The effects of cultural differences for these two language groups on their test performance are not known. Moreover, the French and the Chinese tests are from different subject areas and it is difficult to conduct the comparisons directly. Results from the present research did not find the differences between different languages, the effects may vary according to different samples or subject areas.

By comparing across different subject areas, the majority of the results from this study revealed that the changes in the equated mean scores and the re-classification of examinees in the translated tests were not consistent. However, the Economics test was the only test in which the outcomes across all four equating methods displayed important changes in the equated mean scores and the re-classification of examinees in the

exploratory condition. Possible explanation may attribute to the amount of DIF as a low percentage of common items (44.4 %) that was presented in the Economics (see Table 2). However, the Grade 6 Social Studies test has the lowest percentage of common items (42.0 %) in the exploratory condition but only two equating methods displayed results with important changes. The Grade 9 Social Studies and the History tests did not display any changes across all four equating methods. Therefore, the reasons why tests in different subject areas have different equating outcomes are not known. It may be related to the specific subject areas, the specific tests, the sample differences, the amount of DIF presented in the tests, the selection of the common-item sets, and the equating method.

Furthermore, a multi-method approach was applied to determine the consistency of the equating results across the four equating methods. Results from the study revealed that Tucker linear equating and equipercentile equating tended to reveal several important and similar changes in the equated scores and the re-classification of the examinees (see Tables 5, 6 and 7). Levine and IRT observed score equating revealed a few changes and these changes were different from Tucker linear equating and equipercentile equating. Part of these results are consistent with Marais and Gierl (2002) study which reported that Tucker linear equating and equipercentile equating provided similar results when equating translated tests.

Previous studies in which multiple equating methods were used when equating same language test forms have reported inconsistency of equating results across different equating methods (Eignor, Stocking, & Cook, 1990; Lawrence & Dorans, 1990; Livingston, Dorans & Wright, 1990; Schmitt, Cook, Dorans & Eignor, 1990). Researchers reported that Tucker linear equating and equipercentile equating methods

provided similar results, whereas IRT equating differed from other equating methods such as Tucker linear equating, Levine linear equating, and equipercentile equating. There are some possible explanations. Kolen and Brennan (2004) and Skaggs (1990) explained that the common-item nonequivalent group design requires strong statistical assumptions, and thus, results are influenced by many factors. These factors include the properties of the common-item set, the assumptions of the equating methods, the ability levels of the samples, and the types of tests to be equated. The primary purpose of the present research was not design to determine the effect of these factors in relating to the inconsistency of the equating results. Further study is required.

Limitations of the Study

The present research has three limitations. First, this study was limited by the sample selection in relation to bilingualism and cultural differences. The examinee samples in the present study were living in the bilingual countries of Canada (English and French) or Hong Kong (English and Chinese). It is possible that some examinees in the sample are likely bilingual rather than monolingual. The effects of bilingualism on problem solving and test performances are not well understood. In the data used in the present research, it was not possible to identify the bilingual examinees from the monolingual examinees. However, examinees in the present study had the option of writing the tests in their preferred language. Therefore, it was expected that the examinees chose their strongest language in order to enhance their best performance. Despite the effects of bilingualism, the effects of cultural differences on the translated tests performance for these two language groups are not known. Thus, it is limited to

determine the relation for the cultural distance between the language and ethnic groups and their adverse effects of item bias.

Second, the present research focused on detecting DIF items related to translation errors. The results from the present research showed that the exploratory DIF condition tended to identify more DIF items than the confirmatory DIF condition. The sources of some DIF items are still not known. DIF may occur due to other differences such as differences between the actual cognitive processes or strategies used by the examinees in answering the question in different languages (Gierl et al., 2003). If the DIF item is attributed to the cognitive processes or strategies used by an examinee that are not intended to be measured in the test construct (e.g., testwiseness), then this DIF item would not be used as a common item. The cognitive processes of the examinees were not investigated in the present study. Moreover, no studies have been done yet to develop and to validate the sources of cognitive DIF in translated tests. Think aloud interviews and protocol analysis (Ericsson & Simon, 1993) may be needed in further study to evaluate examinees' cognitive processes.

Third, results from the present research may not be generalized to other subject areas. The present research evaluated four subject areas in French and Chinese translated tests: Social Studies (French), Mathematics (French), History (Chinese), and Economics (Chinese). Results from the present research showed that the equating outcomes in different subject areas were different. Equating outcomes may be test specific, and thus, more subject areas may be evaluated in the future study.

Conclusions

The purpose of this study was to evaluate the effect of using the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996) to select the common items used on the common-item nonequivalent group equating design using translated achievement tests. Three DIF conditions were compared throughout the analyses: (1) No DIF, (2) Exploratory DIF condition, and (3) Confirmatory DIF condition. Four equating methods were used in the exploratory and confirmatory DIF conditions and six data sets were analyzed. Evaluations were focused on the common-item sets, the equated scores, and the classification of examinees.

The results revealed that the exploratory DIF condition tended to identify more DIF items than the confirmatory DIF condition, which lead to the development of different sets of common items for the two DIF conditions in five of the six tests. All the common-item sets for the six tests fulfilled the evaluation criterion for equating. The results revealed that there were important changes in the equated mean scores in four tests, and these scores affected the classification of examinees in three tests, regardless of whether the tests were translated into French or Chinese. Differences in the equating outcomes were found between the DIF and No DIF conditions, between the exploratory and confirmatory DIF conditions, and between different subject areas. Throughout the study, the majority of these changes occurred when using Tucker linear and equipercentile equating methods; few changes occurred when using Levine and IRT observed score equating methods. Two of the six tests showed no differences in the equated scores and the classification of examinees in all DIF conditions and across all four equating methods.

Differences in the equating outcomes across the exploratory and the confirmatory DIF conditions may attributed to the use of different common-item sets, which lead to different equating outcomes and when different equating methods were used. The choice of the equating outcomes may depend on the interpretability of the results. The exploratory condition did not identify the sources of DIF and it was difficult to interpret the equating outcomes. The confirmatory condition identified the sources of translation DIF in the DIF items, and thus, it enhanced the interpretability of the results.

To conclude, the multidimensionality-based DIF analysis paradigm is a suitable approach for DIF detection and equating translated tests for two reasons. First, the paradigm tends to identify more common items than the traditional statistical DIF analyses, which enhances the content and statistical representativeness of the common-item sets for the common-item nonequivalent group equating design. Second, the paradigm helps to identify the sources of DIF which enhances the interpretation of the equated translated tests scores, in contrast to the traditional statistical DIF analyses, as they are difficult to interpret the equating outcomes. Moreover, the results from the study revealed the inconsistency in the equated outcomes across four equating methods. The reasons for the results inconsistency are not well understand. However, possible explanations that the effect of DIF in equating the translated tests may depend on the selection of the common-item sets, the amount of DIF in the translated tests, the sample characteristics, the selection of the equating methods, the magnitude and location of the equated scores, and the subject areas of the translated tests.

Section III: Recommendations

Implications for Future Practice

The results from the present research provide new information on using the multidimensionality-based DIF analysis paradigm with the common-item nonequivalent group equating design. It may have important implications in educational and psychological assessments for four reasons. First, it is important to evaluate the comparability of the original and translated tests by equating. Translated tests are commonly used in educational and psychological assessments. However, translation is rarely perfect. The results from the current research revealed that DIF items were present in all six tests. It is important to evaluate the test equivalences and transform different language tests onto a common scale before comparing the performances between different groups. When translated tests are used in the high-stakes examinations, the inferences of the test results are used interchangeable as their original tests (Cook et al., 2005; Linn, 1993), and these results may affect the decisions making in university or career placements. Thus, it is important that the translated tests are comparable to the original test.

Equating the translated test onto the original test can be used to place the different language tests onto a common scale. However, in many practical situations, the translated tests are not equated onto the original tests before comparing the performances between different language groups. Equating the translated tests may be time consuming and expensive when compared to no equating (Kolen & Brennan, 2004). However, results from the present research showed that the classification of examinees varied in

three of the six tests. The decision of whether to equate or not may depend on the consequences of the test use, sample size, cost, and administration.

Second, when equating is conducted on translated tests, the multidimensionality-based DIF analysis paradigm (Roussos & Stout, 1996) is a useful approach for detecting DIF items and provides an interpretable equating result. By comparing the exploratory and the confirmatory DIF conditions, results from this study revealed that the two conditions developed two different common-item sets, and this led to different equating results. There are two advantages for using the multidimensionality-based DIF analysis paradigm. The first advantage is that it tends to enhance the content and statistical representative of the common-item as it retains more common items than the traditional statistical DIF analyses. The second advantage is that the multidimensionality-based analysis paradigm provides an interpretable result for the equated translated tests scores by identifying the sources of translation DIF, whereas the statistical DIF analyses do not identify the source of DIF items.

The selection of the substantive analyses and statistical analyses are crucial to the success of the application of the multidimensionality-based DIF analysis paradigm. It is possible that different substantive analyses will generate different DIF hypotheses, and thus, detect a different set of DIF items (Gierl et al., 2001). The choice of substantive analyses framework may depend on the researchers' hypotheses on the sources of DIF. However, it is important to use only the validated substantive analyses framework to generate the DIF hypotheses.

Third, it is important to use a multi-method approach to cross-validate the equating outcomes (Cook et al., 2005; Kolen & Brennan, 2004; Skaggs, 1990). Four

equating methods were used in the study to determine the consistency of the equating results. The results revealed that Tucker linear equating and equipercentile equating provide similar results and revealed most changes in the equating outcomes, whereas Levine and IRT observed score equating provide different results from Tucker and equipercentile equating and revealed a few changes only. For example, results from the Grade 9 Mathematics test revealed important changes in the equated scores only when using Levine observed score equating. These results were not displayed by other equating methods. Although the results revealed that two translated tests have no important scores changes in all four equating methods, this finding does not guarantee that there will be no changes if a fifth equating method is applied. In most practical situations, test evaluators may use either IRT equating or classical equating methods. The results from the study revealed inconsistency in the equating outcomes in the translated tests. Therefore, it is important to use multi-equating methods that include IRT equating, classical equating, linear equating and non-linear equating when equating translated tests.

Fourth, by combining three different DIF conditions with the application of four equating methods and four subject areas using French and Chinese translated tests, results from the present research showed that the equating outcomes varied with the choice of the common-item sets, the equating methods, the subject areas, and samples. Therefore, it is important to consider these factors, and possibly more, when evaluating the results from the equated translated tests.

Future Studies

The success of the multidimensionality-based DIF analysis paradigm in equating translated tests may depend on the selection of the substantive and statistical analyses. The present research applied the four sources of translation DIF developed by Gierl and Khaliq (2001) that had previously been validated and confirmed by statistical analyses. The selection of other substantive analysis frameworks may generate different DIF hypotheses, and thus, detect a different set of DIF items. The selection of different statistical analyses to confirm the DIF hypotheses may also provide different results. Therefore, the use of the multidimensionality-based DIF analysis paradigm for equating translated tests may depend on both the substantive and statistical analyses.

Moreover, the present research used substantive analyses results from Gierl and Khaliq (2001) and Gierl et al. (2000) which focused on identifying the sources of translation DIF between different language groups. Future studies may be directed at the evaluation of examinees' cognitive processes in relating to DIF in different language tests. Despite the presence of translation errors, DIF may attribute to other reasons related to differences in different language groups. For example, if the DIF item is attributed to the cognitive processes or strategies used by the examinee that are not intended to be measured in the test construct (e.g., testwiseness), then this DIF item would not be used as a common item. With the use of the substantive analyses to identify sources of cognitive DIF, it is possible to generate a different set of DIF hypotheses. This in turn would lead to a different set of common-item and different equating outcomes. Therefore, it is necessary to generate DIF hypotheses relate to cognitive processes to explain why DIF occurs (Gierl et al., 2001; Gierl et al., 2003). Further studies are

required to evaluate the use of the multidimensionality-based DIF analysis paradigm when different substantive and statistical analyses are selected.

Moreover, the present research was designed to evaluate the effect of different common-item sets on the equating outcomes. Future studies may evaluate the factors that affecting the equating outcomes in the multi-method approach. The results from the current research revealed the similarity between Tucker linear and equipercentile equating results, whereas Levine and IRT observed score equating have different equating results from Tucker and equipercentile equating results. Therefore, the majority of the equating results across different equating methods were inconsistent and the reasons are unknown. Possible explanations include the differences in the assumptions of the equating methods, the sample characteristics of different language groups, the amount of DIF in the translated tests, and other factors that affect the equating results. Real data has been used in the present research. Thus, simulation studies can also be performed to evaluate these factors in a more systematic way.

Finally, this research is limited by the use of six tests in French and Chinese to determine the differences in languages and the equated outcomes. Results from the present research revealed that important equated scores changes and re-classification of examinees occurred in some of the French and Chinese translated tests. However, with the use of the samples from different ages (e.g., Grade 6, Grade 9) and different subject areas (e.g. Mathematics, History) in the current study, it was difficult to compare the results across different languages. To conduct evaluation in equating across different language forms, data from educational and psychological tests with multiple language forms administered in the same subject area, same country, or same cultural group would

be necessary. Future studies using data from educational and psychological tests with multiple language forms are recommended (e.g., the Israeli Psychometric Entrance Test with six languages forms) to evaluate the effect of different language in affecting the equated scores.

References

Alberta Education. (1989). *Program of Studies: Social Studies.* Edmonton, AB: Curriculum Standards Branch, Alberta Education.

Alberta Education. (1996). *Program of Studies: Mathematics.* Edmonton, AB: Curriculum Standards Branch, Alberta Education.

Alberta Education. (2003). *Provincial Achievement Tests: Supporting excellence in student learning in Alberta.* Edmonton, AB: Alberta Education.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36,* 185-198.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Angoff, W. A. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education. (Reprinted as W. A. Angoff, Scales, norms, and equivalent scores. Princeton: Education Testing Service, 1984).

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academic and the Scholastic Aptitude Test.* College Board Report No.88-2. Princeton, NJ: Educational Testing Service. ERIC document: ED 304 457.

Beguin, A. A. (2002). *Robustness of IRT test equating to violations of the representativeness of the common items in a non-equivalent groups design.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23,* 67-95.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test equating* (pp.9-49). New York: Academic.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22,* 13-20.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park: Sage.

Choi, C. C. (1999). Public examinations in Hong Kong. *Assessment in Education, 6*, 405-417.

Choi, S. W., & McCall, M. (2002). Linking bilingual mathematics assessments: A monolingual IRT approach. In G. Tindal, & T. M. Haladyma (Eds.), *Large-scale assessments for all students: Validity, technical adequacy, and implementation.* (pp. 317-338). Mahwah: Lawrence Erlbaum.

Churchill, S. (1998). *Official languages in Canada: Changing the language landscape. New Canadian Perspectives.* ERIC document: ED 429 440.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 3,* 225-244.

Cook, L., & Schmitt-Cascallar, A. (2005). Establishing score comparabiity for tests given in different languages. In R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 131 - 170). Mahwah, NJ: Lawrence Erlbaum.

Cook, L., Schmitt-Cascallar, A., & Brown, C. (2005). Adapting achievement and aptitude tests: A review of methodological issues. In R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 171-194). Mahwah, NJ: Lawrence Erlbaum.

Dorans, N. J. (2004). Using subpopulation invariance to assess score equity. *Journal of Educational Measurement, 41,* 43-68.

Dorans, N.J., Holland, P.W., Thayer, D.T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three advanced placement program exams. In N. J. Dorans (Ed.), *Population invariance of score linking: Theory and applications to advanced placement program examinations* (pp. 79-118), Research Report 03-27. Princeton, NJ: Educational Testing Service.

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

Engelhard, G., Davis, M., & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education, 12,* 199-210.

Engelhard, G., Hansche, L., & Rutledge, E. (1990). Accuracy of bias review judges in identifying differential item functioning on Teacher Certification Tests. *Applied Measurement in Education, 3,* 347-360.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education, 3,* 37-52.

Ercikan, K. (1999, April). *Translation DIF on TIMSS.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Ercikan, K., Gierl, M.J., McCreith, T., Puhan, G., & Koh, K. (2002). *Comparability of English and French versions of SAIP for reading, mathematics and science items.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Toronto.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Feuer, M. J. & Fulton, K. (1994). Educational testing abroad and lessons of the United States. *Educational Measurement: Issues and Practice, 13(2),* 31-39.

Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education, 25,* 280-296.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. Educational Measurement: Issues and Practice, 23 (1), 3-14.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A. & Khaliq, S. N. (2001). *Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, USA.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement, 40,* 281-306.

Gierl, M. J., & Bolt, D. M. (2003). *Implications of the multidimensionality-based DIF analysis framework for selecting a matching and studied subtest.* Paper presented in the annual meeting of National Council on Measurement in Education, Chicago.

Gierl, M. J., Cheng, L., Rogers, W. T., Gotzmann, A., & Vandenberghe, C. (2000). *Translation differential item functioning on the Hong Kong Certificate of Education Examinations in six content areas* (CRAME Research report No. RR-00-01). Edmonton: Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translating achievement tests. *Journal of Educational Measurement, 38,* 164-187.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R. K. (1993). Translation achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-244.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17,* 164-172.

Hambleton, R. K., & Jones, R. W. (1994). Comparison of empirical and judgmental methods for detecting differential item functioning. *Educational Research Quarterly, 18,* 21-36.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological assessment, 11,* 147-157.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45,* 153-171.

Hambleton, R. K., & Patsula, L. (1999). *Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices.* Http: //www.testpublishers.org/journal01.htm.

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

Han, T. (1993). *Comparison of IRT observed-score equating with both IRT true-score and classical equipercentile equating.* Unpublished doctoral dissertation, Southern Illinois University, Carbondale, IL.

Han, T., Kolen, M., Pohlmann, J. (1997). A comparison among IRT true-and observed-score equating and traditional equipercentile equating. *Applied Measurement in Education, 10,* 105-121.

Hanson, B. A., & Feinstein, Z, S. (1997). *Application of a polynomial log linear model to assessing differential item functioning for common items in the common-item equating design.* (ACT Research Report No. 97-1). Iowa, IA.

Hanson, B. A., & Zeng, L. (1995) *PIE program.* IA: University of Iowa.

Harris, D. J. (1991). *Equating with non-representative common item sets and nonequivalent groups.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Harris, D. J. (1993). *Practical issues in equating.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195-240.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braum (Eds.). *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hong Kong Examinations Authority. (1999). *Hong Kong Certificate of Education Examination 1999.* Hong Kong: Hong Kong Examinations Authority.

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

International Association for the Evaluation of Educational Achievement. (1994). *TIMSS main study manuals: population 1 and 2.* Hamburg: Author.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197-206.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18,* 1-11.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25-44.

Kolen, M. J. (2003). *CIPE program.* Iowa: Iowa Testing Programs, University of Iowa.

Kolen, M. J. (2004). Linking assessment: Concept and history. *Applied Psychological Measurement, 28,* 219-226.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 3,* 263-277.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices (2nd Ed.).* New York: Springer.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education, 3,* 19-36.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (Research Bulletin 55-23). Princeton, NJ: Educational Testing Service.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3,* 73-95.

Lonner, W. J. (1990). An overview of cross-cultural testing an assessment. In R. W. Brislin. (ed.). *Applied cross-cultural psychology* (pp. 56-76). Newbury Park: SAGE.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Marais, A., & Gierl, M. (2002). *Examining the impact of DIF items on the equating function.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Toronto.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Educational , Macmillan.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods and prospects.* Princeton, NJ: Educational Testing Services.

Mislevy, R. J., & Bock, R. D. (1997) *BILOG 3.11 Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific software.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland and D. B. Rubin (Eds.), *Test equating.* New York: Academic Press, 71-135.

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

Poortinga, Y. H. (1983). Psychometric approaches to inter-group comparison: The problem of equivalence. In S. H. Irvine & J. W. Berry (Eds.), *Human assessment and cross-cultural factors* (pp. 237-258). New York: Plenum.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basis issues. *International Journal of Psychology, 24,* 737-756.

Puhan, G. (2003). *Evaluating the effectiveness of two-stage testing for English and French examinees on the SAIP Science 1996 and 1999 tests.* Unpublished dissertation. Edmonton: University of Alberta.

Raju, N. S., Bode, R. K., Larsen, V. S., & Steinhaus, S. (1986). *Anchor-test size and horizontal equating with the Rasch and three-parameter models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Rapp, J., & Allalouf, A. (2002). *Evaluating the effect of ability differences between groups and the use of a non-representative anchor on equating in cross-lingual circumstances.* (NITE Research Report). Jerusalem: National Institute for Testing & Evaluation.

Rapp, J., & Allalouf, A. (2003). Evaluating cross-lingual equating. *International Journal of Testing, 3,* 101-117.

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. (2003). Differential validity and utility of successive and simultaneous approaches to the development of equivalent achievement tests in French and English. *The Alberta Journal of Educational Research, XLIX (3),* 290-304.

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

Roussos, L. A. & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Education Measurement, 33,* 215-230.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3,* 53-71.

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

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

Skaggs, G. (1990). To match or not to match samples on ability for equating: a discussion of five articles. *Applied Measurement in Education, 3,* 105-113.

Skaggs, G., & Lissitz, R. W. (1986). *The effect of examinee ability on test equating invariance.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16(1),* 12-19.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Van de Vijver, F., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (vol. 1, pp. 257-300). Boston: Allyn & Bacon.

Van de Vijver, F., & Looner, W. J. (1995). A bibliometric analysis of the Journal of Cross-cultural Psychology. *Journal of Cross-Cultural Psychology, 26,* 591-602.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13,* 29-37.

Van de Vijver, F., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263-279.

Wainer, H. (1999). Comparing the incomparable: An essay on the importance of big assumptions and scant evidence. *Educational Measurement: Issues and Practice, 18,* 10-16.

Wingersky, M. S., Cook, L. L., & Eignor, D.R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (Research Report 87-24). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347-364.

Yang, W. (1997). *The effects of content mix and equating method on the accuracy of test equating using anchor-item design.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Yang, W., & Houang, R. T. (1996). *The effect of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT -based equating using an anchor-item design.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Zeng, L., & Hanson, B. A. (1995). *ST program.* IA: University of Iowa.

Zhang, Y., Matthews-Lopez, J., & Dorans, N. J. (2003). *Using DIF dissection to assess effects of item deletion due to DIF on the performance of SAT I: Reasoning test sub-populations.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp.337-347). Hillsdale, NJ: Erlbaum.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20,* 136-147.

APPENDIX A

Summary for three DIF conditions in identifying DIF and non-DIF items