THE EFFECT OF DIFFERENT ANCHOR TESTS ON THE ACCURACY OF TEST EQUATING FOR TEST ADAPTATION

A Dissertation Presented to

The Faculty of the College of Education of

Ohio University

In Partial Fulfillment

of the Requirement for the Degree

Doctor of Philosophy

Hua Gao

June 2004

This dissertation entitled

THE EFFECT OF DIFFERENT ANCHOR TESTS ON THE ACCURACY OF TEST EQUATING FOR TEST ADAPTATION

BY

Hua Gao

has been approved for

the Department of Educational Studies

and the College of Education by

George Johanson

Professor of Research and Evaluation

James Heap

Dean the College of Education

GAO, HUA. Ph.D. June 2004. Educational Research and Evaluation

The Effect of Different Anchor Selection Approaches on the Accuracy of Test Equating for Test Adaptation (326pp.)

Director of Dissertation: George Johanson

The focus of this study was to evaluate the effect of different approaches of anchor test construction on the accuracy of equating for test adaptation. The term "equating" in cross-lingual studies refers to a statistical procedure that adjusts test scores from the source language (SL) version of the test and the target language (TL) version of the test using a set of common translated items of the same examination so that scores can be interpreted interchangeably. In each test, the verbal section and the non-verbal section of the test were investigated. The Levine Linear equating method and Mean-Sigma equating method were utilized with an anchor item design and an equivalent group design, respectively. The double linking method and the standard errors of equating method were used to evaluate the accuracy of the equating for different anchor tests. The average difference between the two anchor tests for the verbal and non-verbal sections of the test over three target language groups reflected the degree of overall instability that existed in the cross-lingual equating process. These differences were associated with real and systematic variance that underlies the cross-lingual equating process. Scoring outcomes of an actual certification examination with a sample of nearly 9,000 examinees taking both SL and TL versions of the test data set were utilized for this research study.

Findings indicated that the differences between the double linking chains for each anchor test were greater for the verbal section than the non-verbal section of the test. The results of the double linking method supported the notion that different choices for

anchor items can result in different equatings and using items with the more stable parameters was a better choice than using items with less DIF. The results of MSEE did not show large differences between the parameter and the DIF methods of anchor item selection. However, the MSEE differences were in the same direction as the double linking method differences. That is, the parameter method was superior to the DIF method using both criteria.

Approved:

George Johanson

Professor of Research and Evaluation

ACKNOWLEDGMENTS

Completing my doctorate and writing this dissertation have been quite an ordeal, and I never thought I could do it. However, I have many people to thank for their help. I express appreciation to Sheila and Mary who provided the data of this study for taking the time and the effort to help me. They were kind, courteous, and very interesting to listen to. Good luck to all of you in your future pursuits.

I would like to thank my committee members for their support and guidance. I felt each one of you really cared about what I was doing and was trying to make it the best research possible. To Professor George Johanson, Professor Gordon Brooks, Professor Rajindar Koshal and Professor Gary Moden: thank you! I would especially like to thank Professor Johanson for being my dissertation committee chairman. He was always willing to lend a listening ear and philosophize with me about anything. His keen insight, considerable guidance, positive attitude, and continuous support were all essential for my personal and professional growth. I would also like to thank the others for supporting me both temporally and mentally. I consider all of you as my great friends.

I would also like to thank my family for supporting me through this time, and praying for me: my son, for understanding and believing in me; my husband, whom I love so much, is my best friend and who encourages and supports me at all times. He has paid the price for my Ph.D. as much as I have. Thank you all for giving me the hope and the encouragement.

My very special thanks and appreciation go to my parents for giving me the motivation and desire to complete this dissertation and for all the sacrifices that they had since I came to the United States. I grew closer to both of them during this dissertation,

felt guided and inspired in all that I did, and give them credit for all that I have accomplished in life. I also felt as if they were always preparing the way for me, always making it easier to accomplish the tasks that looked unbelievably formidable. I know mom that you prayed a lot for me

Table of Contents

Page

List of Tables

Table Page

List of Figures

CHAPTER ONE

Introduction

Background of the Study

International collaborations arising from the growing international marketplace provide exciting opportunities for researchers who are interested in cross-cultural and international research. Adapted tests are being increasingly used to assess the knowledge and skills of individuals from other cultures and who speak different languages (Sireci & Berberoglu, 2001). Consequently, interest in adapting educational and psychological exams from one language (source language) to another language (target language) has increased in recent years. Increased interest in test adaptation is a natural result of the spread in comparative studies across national, ethnic and cultural groups and the desire to compare the achievement or aptitude of examinees in different countries and cultures. For example, in Europe there are presently 38 countries, 727 million persons, and at least 30 languages. Efforts to adapt educational and psychological exams to ease the transition of persons from one country to another are underway. In the United States, there are large test adaptation projects underway with the National Assessment of Educational Progress, the Scholastic Assessment Test and many others. Some of the most popular American intelligence and personality tests have been translated into more than 50 languages (Muniz, Hambleton, & Xing 2001).

When we talk about test adaptation, we must first distinguish between test adaptation and test translation. Test adaptation is more descriptive of the process that usually takes place with directions, formats, and contexts. However, test translation is

only a small part of the process of test adaptation (Hambleton, 1994). The phrase test adaptation is considered preferable by cross-cultural researchers because it does not imply only a literal word-to-word translation. And the test adaptation process is typically flexible and allows for complex word and situational substitutions so that the intended meaning is retained across languages even though the translation is not completely literal (Geisinger, 1994). Therefore, in this dissertation research, the phrase test adaptation will be utilized instead of test translation.

Test adaptation is necessary as the need for multi-language versions of achievement, aptitude and personality tests increases. Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) have been developed to help test developers administer and use educational and psychological tests. Among these standards, three seem to be relevant to test adaptation. Standard 6.2 states the need to revalidate a test subsequent to major revisions such as when a test is adapted for use in a second language. Standard 13.2 addresses the need to assess reliability and validity of adapted tests for their intended uses. Finally, Standard 13.4 states the need to establish the comparability of tests. These three standards provide a framework for sources of error or invalidity that might arise in the test adaptation process.

However, when a test is adapted from a source language to a target language the result is generally not a psychometrically equivalent test (Allalouf, 1999). In many cases, psychometricians who deal with testing scores across different languages and who try to achieve score comparability between the adapted version of a test and its source version face serious difficulties. These difficulties are related both to differences between the languages of the test and the cultural differences between the examinee groups. In order

to use scores from two different forms or different versions of a test interchangeably they must be based on a common scale, and more specifically, a corresponding relationship between the scores on two given versions of the test must be established. Consequently, a test equating must be employed.

The importance of equating began to be recognized by a broader spectrum of testing researchers in the early 1980s (Woldbeck, 1998). Recently, a great deal of progress in addressing the importance of test equating has been made. For example, the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council for Measurement in Education (NCME) (1999) Standards for Educational and Psychological Testing (ETS) devoted a substantial portion of a chapter to equating, whereas the previous edition did not list equating in the index (Kolen & Brennan, 1995). In addition, several studies have illustrated innovative methodologies for statistically evaluating the comparability of translated items to their original versions (e.g., Angoff & Cook, 1988; Budgell, Raju, & Quartetti, 1995; Ellis & Kimmel, 1992; Hulin, Drasgow, & Parsons, 1983).

If equating has been successful, it is possible to discuss individual examinees' growth over several administrations of a test, to identify changes in population performance over a given period of time, and to compare students who take tests at various times during a year (Angoff, 1971). Or in the case of test adaptation, successful equating will illuminate the true change between the source language version of the test and the target language version of the test. Therefore, test equating is critical in making important decisions. Regardless of the type of decision that is to be made, it should be based on the most accurate information possible: the more accurate the information, the

better the decision will be (Kolen & Brennan, 1995). As a result, accurate test equating is necessary. The focus of this research study is to evaluate the impact of different anchor tests on the accuracy of test equating.

Test Equating

Test equating is a statistical procedure establishing the relationships between scores from two or more tests and placing two or more tests on a common scale. It is often used in situations where multiple forms or multiple versions of a test exist and examinees taking different forms or different versions of the test are compared to one another. In addition to statistical procedures, successful equating involves many aspects of testing including procedures to develop tests, to administer and score tests, and to interpret scores earned on tests.

Other terms used for equating are: linking, scaling, calibration, projection, statistical modification, and social modification. However, these terms cannot be used interchangeably because they are actually different (Linn, 1994). Many researchers believe that a procedure may be called equating only if it is used strictly to equate two testing forms or two versions of a test with the same content. For this reason, equating is utilized in this dissertation study strictly because it is assumed that the two adapted tests are from the same test specification, of equal length, and measure the same construct.

Although the terms used to describe these test equating procedures are different, they are generally classified into two categories: horizontal and vertical. Horizontal equating is appropriate when multiple forms or multiple versions of a test are required to maintain test security. The forms or versions of a test may be not identical, but are expected to be parallel in content and difficulty (Kolen, 1988). Equating procedures do

not function as well when there are large differences in form-to-form or version-to-version difficulty, reliability, or test content. The ability distribution of examinees is expected to be approximately the same in a horizontal equating. When there are large differences among ability distributions, traditional equating methods (e.g., linear and equipercentile equating) may not be appropriate.

Vertical methods are used to equate scores on two tests intentionally designed to be different in difficulty but still measure the same general knowledge or domain or skills. Unlike horizontal equating, the ability distributions of examinees at the various levels will be different (Barnard, 1996). The problem with vertical equating is that it is considerably more complex than horizontal equating. Some measurement experts point out that vertical equating is an appropriate term since equating adjusts for difficulty differences rather than differences in content. However, others do not believe that vertical equating should be included in test equating because the test content at various levels is often different (Kolen, 1981).

Equating Assumptions

Test equating should meet four criteria before being successfully employed. These four criteria are: same ability, equity, population invariance and symmetry (Angoff, 1982; Kolen & Brennan, 1995; Lord, 1980). The same ability criterion means that the tests to be equated must measure the same ability. If the tests are different in content, they should not be equated. The equity criterion implies that the conditional frequency distribution of scores on Test A after equating is the same as distribution of scores on Test B. That is, scores on Test A and Test B should be interchangeable after equating. The population invariance criterion means that the test should be independent

of the sample of examinees employed in the equating process, and a conversion derived from the equating should apply to all similar situations (Kolen, 1988). Last, the symmetry criterion means that the transformation result should be the same regardless of which test is used as converting reference or base, or the interpretation of the test scores should be the same based on either equating from Test A to Test B or that from Test B to Test A. In equating practice, every effort should be made to assure that the above criteria are satisfied to the greatest extent possible (Kolen & Brennan, 1995).

Equating Methods

Test-equating methods can be classified as traditional equating or item response theory equating. Traditional equating methods are based on classical test theory (CTT). In CTT method, score correspondence of tests is established by setting characteristics of the score distribution equal for a specified group of examinees (Kolen, 1998). Three often used equating methods are (a) mean equating, (b) linear equating, and (c) equipercentile equating (Barnard, 1996).

In mean equating, the means of two forms or two versions of the test are set equal to one another for a particular group of examinees or two different language groups. That is, the Form B scores are converted so that their mean will equal the mean scores on Form A; and the source language version of the test is converted so that its mean will equal the scores on the target language version of the test. This type of equating assumes the differences in difficulty between the forms are constant throughout the entire score range.

The second type of traditional equating is known as linear equating. Linear equating is a special case of equipercentile equating. In this equating, the mean and

standard deviation on the two forms for a particular group of examinees are set to be equal, thus:

$$\frac{x_1 - \mu_1}{\sigma_1} = \frac{x^*_2 - \mu_2}{\sigma_2} \quad (1)$$

where x_1 is the raw scores, and x^*_2 is the equated or adjusted scores, μ_1 and μ_2 are means and σ_1 and σ_2 are standard deviations of Test A and Test B respectively. Actually this is a z-score transformation. Therefore, linear equating is also considered as establishing equivalent z-scores for two different tests. This type of equating allows the relative difficulty of the forms to vary along the score scale. For instance, Form A might be relatively more difficult than Form B for low achieving students than for high achieving students (Kolen, 1990).

In equipercentile equating, score distributions are set to be equal so that the same percentile ranks from different tests are considered to be the same level of performance. That is, the Form B distribution is set equal to the Form A distribution for a particular group of examinees by scoring the two tests as percentages. Form B scores that are converted using equipercentile equating have approximately the same mean, standard deviation and distributional shape as do scores on Form A. Scores on Form A and Form B with the same percentile rank for a particular group of examinees are considered to indicate the same level of performance. This provides for even greater similarity between distributions of equated scores than does linear equating. A plot can be constructed between percentile ranks and raw scores for each of the two tests.

The following guidelines should be followed in choosing from among the different traditional methods. Linear equating requires more restrictive assumptions than

equipercentile equating. If tests to be equated have equal standard deviations, then mean equating and linear equating will produce the same results. If the distributions have the same shape, the linear and equipercentile methods produce the same results. Equipercentile equating normally requires larger sample sizes than the other two methods and is more complicated in computation. In this study, the linear equating method will be utilized. See Chapter Three for detailed description of this equating method.

Test Equating Evaluation

The specification of criteria for evaluating a test equating is a critical issue. Over a 50-year span, equating criteria have been developed, reviewed, and criticized as to both usefulness and validity. Without appropriate measures of accuracy, a thorough evaluation of the equating results is not possible. However, there is no single definitive criterion and one criterion may not be appropriate for all equating contexts (Harris & Crouse, 1993). The following summary of different equating criteria is from Harris and Crouse:

1. Weak equity or tau equivalence is considered a special case of Lord's (1980) equity definition. It only requires that means of the conditional distribution on each test after equating be equal. This special criterion includes equivalent expected scores and conditional variance of the equating function. The advantage of this criterion over the other equating criteria is that it is directly aligned with a special case of Lord's equity definition of equating. Therefore, whenever Lord's definition is adopted, it is suggested that the weak equity criterion be used. However, the disadvantage of weak equity is that it is relatively difficult to compute and explain.
2. Summary indices are often used to compare two sets of equating conversions. The root mean square error (RMSE) is frequently used. The advantage of using indices is

that they are easy to interpret. A disadvantage is that the index may not specify the loss function or choice of standard.

3. Standard error of equating is an analytical method to estimate amount of equating error from sampling, that is, one aspect of the accuracy of equating. This method is easy to apply and interpret; however, it ignores systematic errors. The difficulty in using this criterion is that although smaller errors are preferable to larger errors, whether the magnitude of differences between standard errors is important or whether the size of the errors of equating themselves is "large", is unanswered.
4. Estimated scaling constants can be compared to actual constants with generated data. Generated data means that data are generated or simulated. The advantage of this method is that the true equating relation is known and can be used to evaluate the results. This method is most useful when the generated data closely resemble the real data of interest. The disadvantage of this method is the potential bias and the question of how well generated data mimics real data remains unanswered.
5. Estimated scaling constants can be compared to actual constants if a test is equated to itself. Equating a test to itself is also known as circular equating. A test is equated to itself either directly or through a chain of intervening forms. Traditionally, the circular equating criterion was intended to assess systematic error. This method has the advantage of knowing the true equating. However, the drawback is that no equating always works well.
6. A large sample criterion is used as an estimate of the population conversion to evaluate the equating results from smaller groups. This criterion is easy to interpret; however, a large sample is not always available.

7. Consistency criterion means that equating results are compared across methods. Usually all that can be concluded from such a comparison is whether the methods can provide similar or dissimilar results and then one method will be substituted for another for practical reasons. This method does not address accuracy.
8. A Stability criterion compares new procedures to conventional equating methods to assess similarity but not necessarily accuracy. Cross-validation is a common example. This method is easy to apply; however, it does not address the accuracy.

Two additional equating evaluation methods are recommended by Allalouf and Rapp (2000). The first method is called the double-linking method. In this method, a new test form is independently equated to two old forms. The two conversion functions are averaged to produce a single conversion. If the two conversion functions differ more than would be expected by chance, a systematic error would be expected in at least one of the equating processes. This method was specifically developed for evaluating cross-lingual equating. The second method is called the three channel method. These three channels are: (a) using only the non-differential functioning translated items as internal anchor; (b) using non-verbal (largely quantitative) translated items as external anchor; and (c) using an internal, within-language equating channel, in which every new translated form is equated to an already equated translated form. The advantage of this method is using both internal anchor and external anchor as criteria in one study.

Based on the recommendation of the Kolen and Brennan (1995), and Allalouf and Rapp's (2000) studies for evaluating cross-lingual equating accuracy, the double linking method is the only available equating evaluation tool for cross-lingual studies. Therefore, in this research study the double linking equating criterion will be chosen to check the

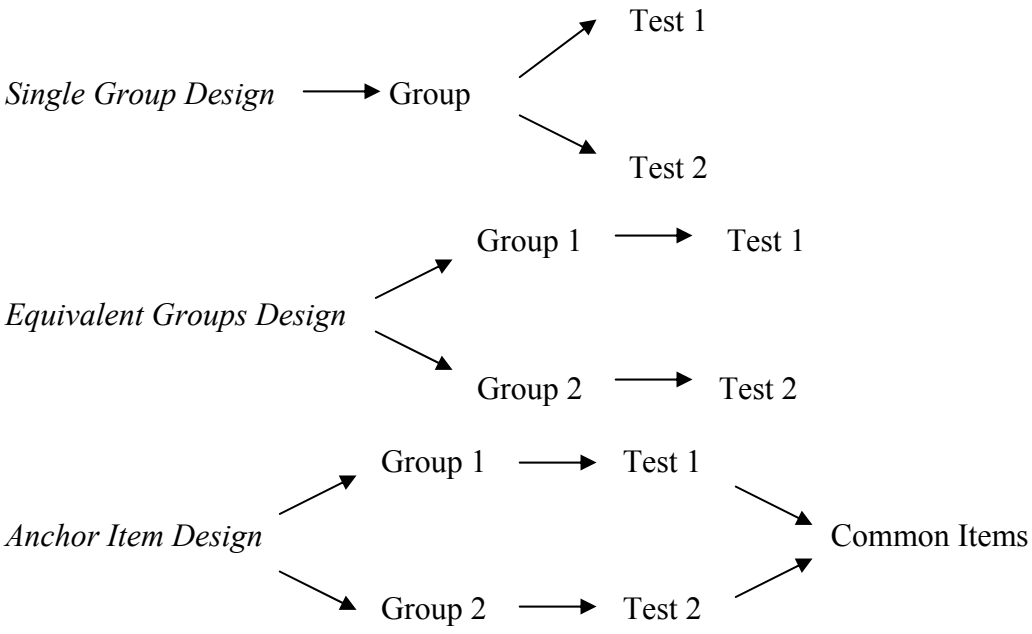
accuracy of different anchor tests. In addition, the mean standard errors of equating criterion will be examined as well. See Chapter Three for detailed description.

Data Collection Designs for Equating

Test equating like other aspects of test development starts with data collection. The three commonly used data collection designs are (1) single group design, (2) equivalent groups design, and (3) anchor item design, which are illustrated in Figure 1 (Kolen & Brennan, 1995).

Figure 1

Illustration of Data Collection Designs of Test Equating



In the single group design, two or more testing forms are administered to the same group of examinees. The advantage of this design is that measurement errors are relatively small since there is only one group of examinees. Differences among tests are

not confounded by differences among groups. The major factors to be concerned within this design are fatigue and practice effects especially when tests are mentally and physically demanding. To avoid fatigue and practice effects, either a spiraling process should be applied or the order of testing forms should be counterbalanced.

The equivalent group design method involves two tests administered to two equivalent groups of examinees. The groups are randomly formed and that is why this design is also called the random groups design. An advantage in using this method is that fatigue and practice effects are eliminated and testing time is minimized. However, a negative factor is the unknown degree of bias introduced because groups often are not identical in terms of their ability distribution. To control for this possible bias, larger groups are generally required for this design (Kolen, 1998; Kolen & Brennan, 1995).

Finally, in the anchor item design, tests are administrated to two different groups of examinees. This design is also called common-item nonequivalent groups design. A set of common or anchor items is included in both tests or both forms, so that the differences between the two can be adjusted based on common item statistics (Zhu, 1998). Two variations of the design are employed depending on whether the common items are administered internally or externally. This design is also useful in measuring growth when two groups are known to be non-equivalent, and is necessary when it is impossible to administer more than one test due to test security or other practical concerns like test adaptation. It is often used when developing an item bank in which testing items are cumulated into a common scale. However, the use of an anchor item design requires strong statistical assumptions for effects of group and test differences; therefore, there should be enough common items with representative content to be

measured. The number of common items to use should be considered on both content and statistical grounds. The major disadvantage of the common item nonequivalent groups design is the stringent statistical analysis underlying the technique.

Data collection design choice depends on practical concerns. Both single-group design and equivalent-groups designs require administration of two tests to the same or equivalent groups with little or no time intervening between test administrations. However, this may be difficult to implement in practice. The anchor item design is less restrictive and may be preferable in terms of practicality (Zhu, 1998). In this study, both anchor item design and equivalent groups design will be chosen.

Selecting Anchor Items for Anchor Tests

When using anchor item non-equivalent groups design, anchor item sets should be chosen very carefully. Equating can be successful only if the anchor items are well selected. Even when an equating study is well designed and statistical assumptions are met, an acceptable equating can be ineffective because anchor items differ from one form to another. Therefore, the procedure for selecting anchor items deserves considerable emphasis because problems with anchor items have serious consequences. If anchor items are not properly selected, the data gathered in an equating study can lead to erroneous conclusions about the comparability of the test forms (Kolen & Brennan, 1995).

In test adaptation, the anchor items selection procedure often requires considerably more effort than that same language anchor items selection procedure because the anchor items in cross-lingual equating are chosen from adapted items. These adapted items are treated as if they were identical, as if they measure the same construct,

and as if they have the same psychometric characteristics (Rapp & Allalouf, 2002). However, it is very difficult to assure that all adapted items used as anchor items are equivalent across languages. In addition, the equivalence of anchor items across languages is the most basic requirement for any equating system to be effective. Therefore, selecting proper anchor items in cross-lingual study is crucial for successful equating.

The different approaches of selecting anchor items (hereafter different approaches of selecting anchor items is referred to as anchor tests) will also have impact on the accuracy of test equating for test adaptation. However, previous studies did not place emphasis on the importance of different anchor tests. Therefore, the focus of this dissertation research study is to find which anchor test results in the most accurate equating.

The following criteria are utilized for choosing proper anchor items: (1) content representations; (2) 20 items or 20% of the total items, whichever is larger; (3) best translation; (4) identifying item difficulty and item discrimination; and (5) finding DIF items using delta plot analysis. Combinations of (1) (2) (3); (1) (2) (4); and (1) (2) (5) of the above anchor tests will be analyzed. More detailed description regarding different anchor tests will be discussed in Chapter Three.

Statement of the Problem

The problem to be addressed focuses on the differential impact of anchor item selections on equating accuracy between English language version and other target language versions of the test. Until recently, the effect of different anchor tests on test equating accuracy has been overlooked. Different test equating criteria for test adaptation

have been used in many research reports; however, the impact of the different anchor tests on the accuracy of equating remains unaddressed. Questions naturally arise as to the extent to which different anchor tests produce more accurate equating results. What problems are likely to result from treating different anchor tests as being equivalent on equating accuracy if they are not?

Research Questions

The following research questions will be addressed in this research study. These focal questions are intended to provide a means of investigating the general problems previously described.

1. What is the best anchor test for equating two forms of the Korean language version of the test and the English language version of the test?
2. What is the best anchor test for equating two forms of the Chinese language version of the test and the English language version of the test?
3. What is the best anchor test for equating two forms of the Spanish language version of the test and the English language version of the test?

Significance of the Study

The basic purpose of accurate test equating is to establish an effective equivalence between test scores (Harris & Crouse, 1993). If accurate equating has been successful, it is possible to discuss the true change over several versions of a test and to compare students who took tests at various locations. The increased attention to accurate test equating has been furthered by an expansion in the number of testing programs. Also, test developers have referenced the role of accurate equating in arriving at reported scores to address issues raised by testing critics while the accountability movements in education

and resultant issues of fairness in testing have become much more visible (Kolen, 1995). The content, format and related issues are vitally important in linking and developing cross-lingual tests. The lack of strong comparability in these areas prevents the development of reliable and valid linkages. Therefore, the goals of accurate equating need to be clearly classified. The design for data collection, the equating linkage plan, the statistical methods used, and procedures for choosing among results should be appropriate for achieving the goals in the particular practical context in which equating is conducted.

The significance of this study can be summarized as follows. First, this study can improve test score integrity to assure fairness of a test or eliminate practice effect. Second, this study addresses the issue of test score exchangeability. When cross-lingual tests are being used to measure the same variable in practice, test scores often are not comparable because they are set on different scales. Therefore, in order to compare examinees or criteria across tests, it is necessary to first convert test scores through test equating onto the same scale. Third, this study addresses test continuity. That is, it allows for cross-lingual tests being used at different levels to measure growth or change in an ability or trait.

Delimitations and Limitations of the Study

This study is delimited in that it used only one data set with three target language groups to evaluate the test equating for the test adaptation. The design of this study hinges on the test data available and the tests are limited to test adaptation only. The sample size for target language versions of the test is small. As a result, the classical test theory (CTT) method equating is the only option for equating different versions of the

test in this study. The number of the anchor items, content of anchor items and the different statistical analyses will effect the interpretations of the results. However, the present study is limited in the following four aspects:

1. The available Certified Quality Engineer (CQE) data that will be analyzed were not necessarily adapted according to the best practices described in Chapter Two. The secondary nature of the data analyzed in this study limited its design and the generalizability of the study outcomes.
2. The criterion that used for evaluating accuracy of test equating had some inherent limitations. The consequence of using arbitrary criteria for evaluating equating accuracy is self-evident. The major drawback is that such criteria do not address equating accuracy directly. The consistency between the chains for double linking method and standard errors of equating (SEE) are measured. In addition, double linking method has not been used widely, SEE method is very subjective. Therefore, the evaluation outcomes based on the arbitrary criteria are interpreted with cautions.
3. The CTT-based equating methods used in this study assume linearity for the test forms being equating. Therefore, the equating results should not be generalized to testing context where non-linearity prevails. In addition, generalizability of this study's findings should be limited to the context where the Levine linear method and Mean and Sigma method apply.

Definitions of Terms

Anchor-Item Design

Anchor-Item Design is also known as common-item non equivalent group design. In this design the tests to be equated are given to two different groups of examinees and each test has a set of common items that may be internal or external to the tests. This design is feasible and frequently used. If the anchor items are chosen properly, it avoids the problems in the other designs (Hambleton & Swaminathan, 1991).

Anchor Test

Different approaches of selecting anchor items are called anchor test. The following criteria will be utilized for choosing proper anchor items in this study: (1) content representatives; (2) 20 items or 20% of the total items, whichever is larger; (3) best translation; (4) identifying item difficulty and item discrimination; and (5) using delta plot analysis. The combinations of (1) (2) (4) and of (1) (2) (5) of the above will be analyzed. The combinations of (1) (2) (4) is called anchor test one and combination of (1) (2) (5) is called anchor test two.

CQE Certified Professional Examination

CQE exam stands for Certified Quality Engineer exam. It consists of 160 multiple-choice questions that are carefully designed, and reviewed for correctness. All these items are computer-scored and analyzed to properly determine the degree of comprehension of the prescribed body of knowledge. There are in total six content areas covered in this exam. The Exam is given in June and December twice a year in the different languages.

Culture

Hofstede (1997) defined culture as "the collective programming of the mind which distinguishes the members of one group or category of people from another" (p. 260) and "derives from one's social environment" (p. 5). In other words, it can refer not only to race, but groups within race, such as the corporate culture, the culture of small-town America or even the culture of a book-club that meets weekly. Marquardt and Engel (1993) gave names to these different levels of culture, including corporate culture, ethnic culture, regional culture, national culture, and global culture. They also stated that culture is determined by religion, education, economics, politics, family, class structure, language, history, and geography.

Globalization

The American Heritage Dictionary of the English Language defines globalization as "making something global or worldwide in scope or application." In business, this term typically refers to the economic and social interaction and integration between cultures, but relates as well to "political, social, cultural and environmental spheres" (Walters, 1997, p. 4). This process of globalization can influence and affect many parts of an organization, including the employees, customers, products, and the instructional materials.

Internationalization

The Localisation Industry Standards Association (LISA, http://www.lisa.org/), an international voluntary association developing guidelines in this area, has defined internationalization as "the process of designing and implementing a product which is as culturally and technically 'neutral' as possible, and which can therefore easily be

localized for a specific culture or cultures" (LISA, 2001). As internationalization is a forerunner to localization, it "reduces the time and resources required for the localization process" (LISA). However, internationalization can take place without localization.

Test Adaptation and Test Translation

Test adaptation includes such activities as (a) deciding whether or not a test can measure the same construct in a different language and culture, (b) selecting translators, (c) deciding on appropriate accommodations to be made in preparing a test for use in a second language, and (d) adapting the test and checking its equivalence in the adapted form. Test translation, on the other hand, is only one of the steps in the process of test adaptation and, even at this step, adaptation is often a more suitable term than translation to describe the actual process that takes place (Hambleton, 1992 p. 3).

Test Equating

Test equating is a statistical procedure to establish the relationships between scores from two or more tests. This is a procedure to place two or more tests on a common scale. It is often used in situations where multiple forms or multiple versions of a test exist and examinees taking different forms or different versions of the test are compared to one another.

Organization of the Dissertation

There are five parts included in this dissertation. In Chapter one, a background study was provided including related issues of equating, test adaptation, the problem of the study, research questions, significance of the study, limitations of the study and some definitions of terms. Chapter two provided an extensive overview of previous study

related to test equating and test adaptation. Research design, data collection and statistical method were given detailed description in Chapter 3. The results of the analysis and corresponded to the research questions can be found in Chapter four. In last Chapter five, summary, conclusions and suggestions were provided in details.

CHAPTER TWO

Literature Review

Recently, the issues and interest involved in test adaptation and test equating have received considerable attention. This chapter presents an overview of previous research from different authors focusing on test adaptation, test equating and the methods appropriate to assess the accuracy of such test adaptation.

Test Adaptation

Previous Studies in Test Adaptation

Test adaptation stems largely from an increasing number of students worldwide who are not proficient in English and the desire to compare the educational achievement of students in different countries. There are numerous examples of the use of tests to compare individuals across languages. The International Association for the Evaluation of Educational Achievement (IAEEA) (1994) and Miura, Okamoto, Kim, Steere, and Fayol (1993) compared the educational achievement of students in different countries who received instruction in different languages. Ellis (1995), Hulin (1987), Hulin and Candell (1986), and Sireci and Berberoglu (2000) evaluated the cross-cultural generalizability of attitudes or psychological constructs. Angoff and Cook (1988) evaluated the academic proficiency of non-English speaking students in the United States with respect to their English-speaking peers. However, most of these test adaptation studies have focused on tests that were adapted into Spanish from an original English language version of the test.

Test Adaptation and Test Translation

During the past several decades, the unique challenges of cross-cultural studies have attracted considerable attention. Cross-cultural assessment has become a sensitive issue due to specific concerns regarding the use of standardized tests across cultures and languages. Within this context, Butcher and Garcia (1978) identified test translation or test adaptation as two main problems associated with cross-cultural testing. One problem arising from different languages and different cultures is comparing the accuracy of test translation or test adaptation between English language version of the test and other language version of the test. They found that the translated or adapted version of the test is not identical with respect to its English language version of the test. Consequently, they concluded that translation or adaptation of a test into another language is an important task.

Hambleton (1992) distinguished the difference between test translation and test adaptation:

> The term test adaptation is preferred to the more popular and frequently used term test translation in our work because the term test adaptation is broader and more reflective of what should happen in practice when preparing a test that is constructed in one language and culture for use in a second language and culture. Test adaptation includes such activities as (1) deciding whether or not a test can measure the same construct in different language and culture, (2) selecting translators, (3) deciding on appropriate accommodations to be made in preparing a test for use in a second language, (4) adapting the test and checking its equivalence in the adapted form. Test translation, on the other hand, is only one of

> the steps in the process of test adaptation and even at this step, adaptation is often a more suitable term than translation to describe the actual process that takes place (p. 3).

Further, Hambleton (1993) outlined test adaptation procedures for international achievement instruments. It was the basis for the test adaptation guidelines. Hambleton acknowledged that any translation procedure is not merely of test translation, but test adaptation:

> Some researchers prefer the term test adaptation to test translation because the former term seems to more accurately reflect the process that often takes place: Producing an equivalent test in a second language or culture often involves not only a translation that preserves the original test meaning, but also additional changes such as those affecting item format and testing procedures. Such changes may be necessary to insure the equivalence of the versions of the test in multiple languages or cultures (pp. 3-4).

When tests are adapted from one language to another, they may not retain their psychometric properties. For example, adapting a test from one language to another, typically from English, may mean that items are organized by order of English difficulty, rather than reflecting the developmental order of the target language.

Restrepo and Silverman (2001) found several item difficulty discrepancies between the original English and the translated Spanish version when tested with predominately Spanish-speaking preschoolers. For example, items that related to prepositions were relatively easy for English speakers but were more difficult for Spanish speakers. On the other hand, the function items requiring students to point out objects

based on a description of their use (something like "Show me what people use for cooking" or "What do you sweep with?") were easier for the Spanish speakers than the English speakers.

Hambleton (1994) provided a nice example of how simple adaptation can create a problem: “Where is a bird with webbed feet most likely to live? (a) in the mountains (b) in the woods (c) in the sea (d) in the dessert.” When adapted into Swedish, the question becomes much easier as the swimming feet will be used for Swedish word webbed feet. Therefore, adaptation needs to consider the whole cultural context within which a test is to be used.

Gierl, Rogers, and Klinger (1999) provided another example. In an English and French comparison, English-speaking examinees were presented the item in Figure 2. The phrase “historical record” was included in the correct option D for the English form whereas the phrase “source of information” was used for the French form. Because the caption for this item was a picture of an “ancient Greek vase,” the word “historical” provided English examinees with a clue about the correct option. The outcomes from these items will yield misleading test score interpretations if they are attributed to achievement differences between language groups instead of translation errors. Consequently, bias is always a concern when a test is adapted from one language or culture to another.

Figure 2

The English form of a Grade 6 Social Studies Achievement Test Item.

2. Ancient pottery, such as the vase, is important today mainly because it

    a. shows a primitive aspect of Greek cultural

    b. is fragile and must be kept in the museum

    c. is considered priceless as art collectors

    d. becomes a type of historical record

Figueroa (1989) noted that words may generally represent the same concept but have variations and different levels of difficulty across languages. An illustration of this is found in a study of vocabulary test adaptations (Tamayo, 1987). When test items were adapted from English to Spanish, they differed in frequency of occurrence in each language. Because the Spanish adaptations were of lower frequency within Spanish, test scores obtained from Spanish speakers were lower compared to scores obtained from the original English version. However, when the vocabulary items were matched for their frequency of occurrence in the original and target language and matched for meaning, test scores obtained from Spanish and English speakers were equivalent.

Similarly, across different languages the same general category may have different prototypical members, and different words may be associated with each language for the same situation. These contextual variations make adapted vocabulary tests particularly vulnerable to imbalance. When Pena, Bedore, and Zlatic-Giunta (2002) asked bilingual four- to six-year-olds to give examples of animals, the children's three

most frequent English responses were "elephant," "lion," and "dog," while in Spanish they used "caballo" (horse), "elefante" (elephant), and "tigre" (tiger) in these orders.

In addition to vocabulary differences, grammatical structure also affects the validity of test adaptation practices. For example, nouns are marked by gender in Spanish, but not in English. An English test adapted to Spanish will miss aspects of Spanish, such as gender marking, that are not present in the English language. Furthermore, Spanish subject information is frequently carried in the verb resulting in more complex verbs and less salient pronouns as compared to English. In English language assessment, pronoun omission is a hallmark of language impairment yet this would not be true for Spanish. Thus, adapted language tests may target inappropriate features for the target language, resulting in inaccurate assessment of language ability.

Hambleton (1992) emphasized the need for care in adaptation and for ensuring the equivalence:

> Unless the adaptation work is done well, and evidence is compiled to establish, in some sense, the equivalence of the two versions of the test, questions about the validity of the adapted tests will arise. Also, the validity of comparisons among countries where different versions of the test have been administered will be in doubt until questions about the equivalence of the versions are resolved (p. 3).

Lonner and Berry (1986) summarized four types of equivalence in test adaptation: functional equivalence, conceptual equivalence, metric equivalence, and linguistic equivalence. Functional equivalence refers to the role or function that behavior plays in different cultures. One cannot assume that behaviors play the same role or function across cultures; therefore, assumptions made about the function of behavior in a cultural group

must be verified. Conceptual equivalence refers to the similarity in meaning attached to behavior or concepts. Certain behaviors and concepts may have different meanings across cultures. Metric equivalence refers to the psychometric properties and indicates that the scales measure the same constructs in different cultures. Finally, linguistic equivalence refers to the actual translation process.

Fouad (1993) and Geisinger (1994) indicated that before selecting an assessment instrument for use in test adaptation, researchers are trained to verify that the test is appropriate for use with their population. This includes investigation of validity, reliability, and appropriate norm groups to which the population is to be compared. Validity and reliability take on additional dimensions in cross-cultural testing as does the question of the appropriate norm group. The instrument must be validly adapted, the test items must have conceptual and linguistic equivalence, and the test and the test items must be bias free.

Hambleton (1993, 1994) identified two basic methods for test adaptation: forward translation and back-translation. In forward translation, the original test in the source language is translated into the target language and then bilinguals are asked to compare the original version with the adapted version. In back-translation, the test is translated into the target language and then it is re-translated back to the source language. It is possible to repeat this process several times. Once the process is complete, the final back-translated version is compared to the original version. Each adaptation process has its strengths and limitations.

Hambleton (1993), Hambleton and Kanjee (1995) and Hambleton and Patsula (1999) recommended adapting an existing instrument instead of developing a new one.

The advantages of adapting an existing test are: (a) need, (b) cost, (c) security and (d) fairness.

First, for cross-national, cross-language and cross-ethnic comparative studies, test adaptation is necessary. Recent development of International Guidelines on Test Use is a good example. That is, a detailed set of guidelines for adapting psychological and educational tests in various different language and culture contexts has been presented.

Second, adaptations can conserve more time and expenses than creating a new test for second language group. Normally, it will take years to develop and validate a new test and cost a lot of money as well. By adapting a test, the existing database will provide a framework to design and interpret the validity of the studies.

Third, by adapting an instrument, the researcher is able to compare the already-existing data with newly acquired data, thus allowing for cross-cultural studies both on the national and international level. Therefore, researchers often have a sense of security when adapting a test instead of initiating a new test. Last, test adaptation can lead to increased fairness in assessment by allowing individuals to be assessed in the language of their choice.

Hambleton (2000) summarized sources of errors for test adaptation process: (a) cultural/ language differences, (b) technical methods, and (c) interpretation of the results. Failure to attend to the sources of error in each of these categories can result in an adapted test which is not equivalent in the two languages and cultural groups.

Cultural and language differences can affect test scores for test adaptation. Construct equivalence, test administration, test format and speed of responses should be taken into consideration when evaluating the results of the two versions of the tests.

Other response styles, such as acquiescence, tendency to guess and social desirability are major concerns as well. Test format is an important factor for this category. Differential familiarity with particular item formats presents an important source of invalidity of test results in cross-cultural studies. For example, students from United States are all very familiar with the selected response questions such as multiple-choice questions. However, nationalities that follow the British system of education place great emphasis on essays and short answer questions. Therefore, students from these countries are positioned at a possible disadvantage as compared to their American counterparts.

For technical designs and methods, there are five major sources of errors that can influence the validity of adapted tests: (a) the test itself, (b) selection and training of translators, (c) the process of translation, (d) judgmental designs for adapting tests, and (e) empirical analyses for establishing equivalence.

The last category is the factor affecting interpretation of results. In large scale cross-cultural studies, the purpose of the test is to provide a basis for making comparisons between various cultural and language groups in order to understand the differences and similarities that exist. Therefore, when interpreting scores relevant factors external to the tests or assessment measures should also be considered to minimize errors.

Lonner and Berry (1986) argued that the disadvantages of test adaptation include the risk of imposing conclusions based on concepts that exist in one culture but that may not exist in the other. That is, there are no guarantees that the concept in the source culture exists in the target culture. Another disadvantage of adapting existing tests for use in another culture is that if certain constructs measured in the original version are not found in the target population, or if the construct is manifested in a different manner, the

resulting scores can prove to be misleading. However, they concluded that despite the difficulties associated with using adapted instruments, this practice is important because it allows for greater generalizability and allows for investigation of differences among a growing diverse population.

Another issue that must be considered in cross-cultural assessment is test bias. Fouad (1994) asserted that the test user must ascertain that the test and the test items do not systematically discriminate against one cultural group or another. Test bias may occur when the contents of the test are more familiar to one group than to another or when the tests have differential predictive validity across groups. Culture plays a significant role in cross-cultural assessment. Whenever tests developed in one culture are used with another culture there is the potential for misinterpretation and stagnation unless cultural issues are considered. Therefore, issues of test adaptation, test equivalence and test bias must be considered in order to fully utilize the benefit of cross-cultural assessment.

van de Vijver and Poortinga (1991) summarized five possible problems in cross-cultural testing. These five problems are: (a) problems related to the testers, (b) problems related to the examinees, (c) problems related to the interaction between tester and examinee, (d) problems related to the response procedure, and (e) problems related to the stimulus materials.

Testers and examinees could be the obstacles to measurement of the trait being measured. Although the effects of testers have been generally small, they may have been a threat to the validity of the measurement. The choice of examinees can affect the results. Differences in the different culture groups may be responsible for observed differences in performance rather than differences in the trait being measured.

Interaction between the tester and examinees may also be a source of difficulty. Establishing clear communication between testers and examinees as the expectation of the test is important to proper test use.

Clear response procedure is very important. Different level of familiarities with the response medium could affect measurement of the trait of interest. The method of presentation of stimuli can also be a difficulty. That is, a problem of differential familiarity with the materials used to respond to the exam.

Akagi (1991) addressed more problems encountered in adapted tests: (a) validity; (b) familiarity with the material used; and (c) ceiling effects. Validity is the first concern, that is, when adapting a test, it should stem from the context in which the items exist. The concern about the familiarity with the material used is that the adaptations must be made to conform to the standards used in the country of interest. Different level of familiarity with the system may make the item easier or harder, and thus distort the comparisons between languages. Ceiling effects may affect comparisons between languages if one language group shows the effect and the other does not. This may result in misunderstanding the effects of instruction.

Important Guidelines for Test Adaptation

Guidelines for test adaptation are very important for cross-cultural assessment and many researchers agreed that there is a need for guidelines for test adaptation. Bullinger, Anderson, Cella, and Aaronson (1993) proposed both a minimum and an optimum set of criteria for conducting test adaptation studies. At a minimum level, forward and backward translation studies, reliability and validity studies in each language and cultural groups on samples of at least 100, and clear documentation of the test adaptation process

and findings would be needed. At an optimum level, more translators would be used and more reviews of the translations would be conducted, and expended efforts to establish empirically the equivalence of the test in multiple languages and cultures would be carried out.

The International Test Commission (ITC) is an association of national psychological associations, test commissions, test publishers and other organizations committed to promote effective testing and assessment policies and to the proper development, evaluation and uses of educational and psychological instruments (ITC, 1995). ITC consists of a 13-person committee of psychologists representing a number of international organizations to prepare a set of guidelines for adapting educational and psychological tests. Among its various activities, the ITC is responsible for a number of international projects. Over the past few years, the ITC has adopted a policy of focusing attention on those areas where international coordination of effort is most important. As a consequence of this, two major projects have been initiated. One has been concerned with guidelines on adapting tests, the other more recently on developing guidelines for test use. Bartram (1995) summarized a number of reasons why guidelines on test adaptations are needed at an international level:

1. Difference in Statutory Control. Countries differ greatly in the degree of statutory control that can be exercised over the use of testing and its consequences for those tested. Some national professional societies have statutory registration, whereas others do not; some have mechanisms for the control of standards of test use by non psychologists, whereas others do not. The existence of a set of internationally accepted guidelines would provide national psychological associations and other

relevant professional bodies and organizations with a degree of support in their endeavors to develop standards in those countries where they are currently either lacking in some respect or nonexistent.

2. Pattern of Access. Patterns of access in terms of the rights to purchase or use test materials vary greatly from country to country. In some countries, access is restricted to psychologists, in others to users registered with formally approved national test distributors, in yet others test users may be free to obtain materials without restriction from suppliers in their country or directly from suppliers abroad.
3. Background of Test Users. A recent international survey (Bartram & Coyne, 1998) showed that for both educational and work-related testing, non psychologist users far outnumber psychologists. Only in the area of clinical testing, which in volume terms is relatively small, do psychologists tend to account for the majority of test users.
4. International Copyright. A number of well-known instruments have appeared on the Internet in violation of copyright without acknowledgment of the test authors or publishers, and without regard to issues of test security.
5. Mobility of Labor. Within the occupational testing arena, the greater international mobility of labor has increased the demand for tests to be used on job applicants from a number of different countries often with the tests being administered in one country on behalf of a potential employer in another.
6. Internet Applications. Development work is being carried out in the United States and in Europe on the use of the Internet for distance-assessment or remote-assessment in both occupational and educational settings. This raises a whole host of issues relating

to standards of administration and control over the testing process, including test security.

The ITC has worked for 3 years to produce near final drafts of 22 guidelines organized into 4 categories: (a) context; (b) instrument development and adaptation; (c) administration; and (d) documentation and score interpretations. Each guideline by itself is described by a rationale for inclusion, a set of steps for achieving the guideline, a list of common errors, and references for follow-up study. Sireci (1997) provided the following critical guidelines for test adaptation.

1. Get to Know The Culture As Well As The Language. Cultural difference should first take consideration in test adaptation. Familiar features of tests in one culture may be completely unfamiliar in another culture. Therefore, the construct equivalence (Sireci, Bastari, & Allalouf, 1998) of the knowledge, skills, and abilities tested must be considered, as well as the cross-lingual generalizability of the practice analyses and test specifications. Becoming familiar with the cultures to be tested will help in deciding whether it is sensible, legitimate, and feasible to adapt an existing test.

2. Select Translators Carefully. The quality of a test adaptation depends on the quality of the translators. Hambleton and Patsula (1999) summarized at least four criteria to be considered in selecting translators: (a) proficiency in both languages, (b) familiarity with both cultures, (c) proficiency in the subject matter tested, and (d) item writing expertise. When choosing translators, all four criteria should be considered because item writing expertise is trainable, however, the other qualities are more difficult to find or teach.

3. Involve As Many People in The Adaptation Process As Possible. The rule in test adaptation is the more people involved in the adaptation process, the better the adaptation will be. Several adaptation designs are available, such as forward translation, backward translation, parallel development, and combinations of these designs (Brislin, 1986; Hambleton, 1994). Critiques of these designs consistently suggest that independent teams of translators be used whenever possible, so that they can check one another's work. A related important issue to consider is the diversity of dialects within a language.

4. Pilot-Test the Adapted Examination. A pilot test is essential in test adaptation process. A pilot test can help evaluate construct equivalence and item functioning across languages. In addition to statistical analyses, interviews of examinees who sit for the pilot test should prove illuminating regarding the quality of the adaptation and the comparability of test scores across languages.

5. Conduct Statistical Analyses of Test Quality and Comparability. Statistical analysis plays an important role in test adaptation. The equivalence of the constructs measured, the functioning of the items, and the validity and comparability of the passing standards across languages are all issues that can be evaluated statistically. For example, reliability and validity statistics, factor or multidimensional scaling analyses, and differential item-functioning analyses all can be computed.

6. Document The Adaptation Process. Test adaptation like most test development activities is not static one-time events. The entire process from the decision to adapt to selection of the translators, through conduct of the validation studies, should be

thoroughly documented. This documentation will be useful to examinees, licensing authorities, and other invested parties; it will also be useful to test developers when they need to replicate the process.

Test Equating

When we talk about test equating, no matter what equating procedure is chosen, first we have to understand the conditions of equivalence. Lord (1980), Angoff (1984), and Dorans (1990) summarized the following: (a) same construct, (b) equity, (c) symmetry, (d) population invariance, and (e) unidimensionality.

The requirement of same construct of two tests can be achieved by carefully selecting items that measure the same construct during the test construction process. Since the equating is a process of transforming scores for the purpose of comparison, it makes no sense for the forms of a test or two versions of the test to measure different constructs.

Equity requires that individuals have the identical proficiency no matter what forms or versions of the tests are taken. That is, every ability level of the conditional frequency distribution on one form of the test is the same as that of another form. The equating transformation is symmetric, that is, the equating of A to B is in inverse of the equating of B to A.

The population invariance refers to no matter which groups of examinees are used; the equating results should not change with the characteristics of the particular groups except for the underlying construct that the test is measuring. It can be assessed by

the examining the relationship of equivalence across sub-groups. The condition of population invariance is one of the goals of test equating.

Unidimentionality is a requirement for IRT equating. Therefore, IRT equating is more restrictive than the other equating methods.

Different Ways of Defining Equating

Test equating has been defined in many ways. Angoff's (1971) proposed the equipercentile equating definition:

> Two scores, one on form X and the other on from Y (where X and Y measure the same function with the same degree of reliability) may be considered equivalent if their corresponding percentile ranks in any given group is equal (p. 563).

Lord's equity (1980) requires that the conditional distributions of scores on each test after equating must be equal. Divgi (1981) presented two approaches to equating based on the concept of equity and the "given group" must consist of persons with exactly the same ability. Weak equity or tau equivalence (termed weak equity by Divgi, 1981, and tau equivalence by Yen, 1983) are considered special cases of Lord's equity definition, and only requires that the means of the conditional distributions of scores on each test after equating being equal.

Morris (1982) summarized the difference between strong and weak equating by stating:

> Two tests are strongly equated if every individual in the test population has the same probability distribution for the score on both tests. Two tests are weakly equated if each individual in the test population has the same expected score on both tests (p. 171).

In the cross-lingual equating case, when the general assumptions do not hold perfectly, it had led some researchers to label the relationship between two language forms as linking rather than equating. Brennan (2001) stated a typical justification for using the term linking:

> I use the word "equating" to refer to a statistical relationship between scores on forms of a test constructed according to the same content and statistical specifications and administered under the same conditions. By contrast, when any of these conditions are not fulfilled, I use the term "linking" (p.10).

Aspects That Influencing Satisfactory Equating

Brennan and Kolen (1987a) provided a set of guidelines for satisfactory equating. First, for test structure, the test content and statistical specifications for tests being equated ought to be defined precisely and be stable overtime. The test should be reasonably long with at least 35 items and the scoring keys should be consistent. Item statistics should be obtained from pretest or previous use of the test. Second, a list of the ideal situations for equating are: (a) two sets of common items embedded in the full length test were desired; (b) the anchors should be at least half of the total test in length and reflect the total test in content and specification and statistical characteristics; (c) at least one link form was administrated no earlier than one year in the past, and at least one link form was administered in the same month as the form to be equated; and (d) each common item should be in the same position between the two forms. Last, the characteristics of the examinee group should be stable over time. That is, the curriculum, training materials and field should be stable. Also, the size of the groups should be relatively large (i.e., more than 400).

There are many literature reviews about how to select or tailor an equating method to practical needs. For adapted tests, it is expected to be highly accurate when selecting an equating method that functions better for that particular test. Crocker and Algina (1986) summarized the aspects to consider in selection of equating methods: (a) are the underlying assumptions tenable? (b) is the procedure practical? and (c) how good is the equating result?

First, the premise of a model application is that all the underlying assumptions hold. Both equipercentile equating and linear equating assume that the tests being equated measure the same trait with equal reliability. In addition, linear equating assumes that the tests being equated have identical shape for the score distributions differ only in the mean and/or standard deviations. IRT equating requires unidimentionality and item and ability invariance assumptions. If the assumptions do not hold, these equating methods may lead to erroneous results.

Second, random assignments may save time and money for equating, but it is not always practical or feasible because tests are usually administered to convenient intact groups of examinees. Thus anchor design will be a solution. However, if either linear or equipercentile methods is used, the results will not be accurate because the assumptions can not be held without random assignments. Methods based on latent trait theory are more adequate although this method is more laborious and costly.

Third, equity accuracy depends on the conditions of equivalency, that is, same construct, equity, and symmetry and group invariance. Perfect equivalency is very difficult to determine since the true score cannot be known and can only be estimated from the observed scores. Therefore, there is no absolute criterion for equating accuracy.

The degree of accuracy is often estimated by comparing the equating result against arbitrarily sound criteria.

Different Designs for Cross-Lingual Studies

Hambleton (1993, 1994), Sireci (1997), Hambleton and Patsula (1998) and Cook (2000) reviewed the various design methods for equating tests across languages. They concluded that there are three designs for cross-lingual equating: (a) the bilingual group design; (b) the matched monolingual groups design; and (c) the separate monolingual groups design.

The bilingual group design is to assure that a group of bilingual examinees equally proficient in both languages with respect to the construct being measured are tested in the two languages versions of the test. If there is a difference in achievement between the two language versions, it is attributed to differences in the difficulty of the two versions. Although promising, a problem is that in practice it is very difficult to find examinees who are equally proficient in both languages.

The second design is that a group of examinees from each language is selected so that they are matched on particular criteria, such as socio-economic status and the level of education. They compare the achievement of these groups. A major problem with this design is the need to choose relevant and available criteria for matching.

The third separate monolingual groups design is the most popular design in cross-lingual equating. It is a variation of common-item non equivalent groups design, which is used for 'regular' same language equating (Angoff, 1984). In this design, source language and target language versions of a test are administered separately to source- and target- language examinee groups respectively. A set of items common to the two tests is

used to link the scores. These items are treated as if they were identical and measure the same construct, and as if they have the same psychometrical characteristics. Since this method does not require examinees with special characteristics (bilingual, 'matching') that might be difficult or impossible to find, this method seems to be relatively easy to apply. However, due to uncertainty that all the translated items used as anchor items are equivalent across languages, which is the basic requirement for equating to hold, a separate monolingual group's procedure suffers from a theoretical flaw. In addition, it is difficult and practically impossible to ensure that different languages test versions measure exactly the same construct. Therefore, a high risk of equating error will occur.

Anchor Item Design

Equating results depend on the accuracy of the anchor tests in cross-lingual studies. Consequently, it is crucial to adequately select anchor items. The most important characteristics of the anchor item selection for test adaptation are content representation, adequate anchor items, literal translation, and items showing no DIF.

Content Representation

Budescu (1985) pointed out that whether the anchor items are representative to the overall items of the tests being equated in terms of content and statistical properties is very important when groups are vary in ability. The magnitude of the correlation between the anchor test and the unique components of each test form was the single most important determinant of the efficiency of the equating process. Brennan & Kolen (1987b) further indicated that any substantial content change entailed a re-scaling and re-norming of the test with a new 'origin' form to which subsequent forms were equated.

Number of Anchor Items

It is impossible to offer universal guidelines for selecting the length of anchor items (Kolen & Brennan, 1995). However, for its specific purposes, each test needs to take into account the time, cost and context constraints as well as the particular index of efficiency when determine the length of the anchor. Angoff (1984) summarized a rule of thumb for the appropriate number of anchor items is at least 20 items or 20% of the total number of items in a test, whichever is larger.

McBride and Weiss (1974) claimed that 40 to 60 anchor items may be needed to calibrate an item pool using classical test model. Based on theoretical values of standard errors of item estimates, Wright (1997) considered an example of 400 persons and anchor items of 10 to 20 as sufficient for most equating situations. Wright contended that ten anchor items may be adequate if the items are good.

McKinley and Reckase (1981) investigated effects of sample size and anchor test length on precision of the item parameter estimates. There were three levels in test length: 5, 10 and 25 items. Correlation between linked estimates and estimates obtained from the original total sample was used as an evaluation criterion. Obtained correlation values under all conditions were close to unity. Despite trivial differences among the correlations, results generally indicated the longer the anchor item and the larger the sample size, the better the precision. Only in one condition was the five-item anchor better than the 15 anchor. This investigation concluded that a five-item anchor might be adequate, but a 15 anchor was suggested.

Raju, Edwards, and Obsberg (1983) and Lord (1980) suggested that as few as five or six carefully chosen items could perform as satisfactory anchors in IRT equating when

the item parameters of both tests were estimated in a single analysis using IRT concurrent methods. However, Hills, Subhiyah, and Hirsch (1988) studied the effect of anchor test length and found that five randomly chosen anchor items of a mathematics test were not sufficient to produce satisfactory equating results. An anchor of ten items was found satisfactorily sufficient when IRT method was adopted.

None DIF Items

Differential item functioning (DIF) analyses are often used during the test adaptation process to identify items that function differently between language groups (e.g., Allalouf, Hambleton, & Sireci, 1999; Budgell, Raju, & Quartetti, 1995; Hambleton, 1994; van de Vijver & Leung, 1997). DIF is present when examinees from different language groups have a different probability or likelihood of answering an item correctly after conditioning on overall ability.

When items show DIF, these items should be removed from anchor tests because these items lower the reliability and validity of the adapted tests. Further, the DIF items should remove from the item bank so that they will not be used in the future tests. However, removing these items from an item bank involves a financial aspect since new adapted items are expensive to produce.

Allalouf (2003a) identified the methods in detection DIF in test adaptation. He stated that in test adaptation and cross-lingual assessment, DIF detection methods assist in making crucial decisions before and after adapting a test. Before adapting a test refers to a process of determining the translatability of tests and items, and after adapting a test is a process of scoring, equating and maintaining a cross-lingual item bank.

They are so many findings for DIF in test adaptation studies**.** Generally, adapted items do vary in the amount of DIF found. Angoff and Cook (1988) analyzed the equivalence between the SAT and its Spanish-language counterpart, the Prueba de Aptitud Academica (PAA). They found that verbal items that contain more text have higher DIF than items containing less text, where every word is critical and every adaptation problem has an effect on item performance. For example, reading comprehension items have higher DIF than verbal analogies. On the other hand, no DIF is expected in non-verbal items such as math or figural items, as noted by Gafni and Melamed (1994).

Some studies list the possible causes of DIF between test forms in different languages. One study for example, Allalouf, Hambleton and Sireci (1999) studied the adaptation of the verbal reasoning domain of the PET (Psychometric Entrance Test, which is used in selecting candidates for universities in Israel) from Hebrew to Russian. They found that DIF is likely to occur if there are differences between source and target language in: (a) word difficulty, (b) item format, (c) cultural relevance, and (d) content.

In another study, Gierl, Rogers, and Klinger (1999) identified four similar sources of DIF in Canadian Achievement Test administered in English and French. The sources they found were: (a) omission or addition of words or phrases that affect meaning, (b/c) differences in words or expressions inherent/not inherent to the language or culture, and (d) format differences between the test forms in different languages. They created an eleven member panel that, by using these sources, had significant success in predicting the language group that would perform better on item bundles.

Allalouf, Hambleton, and Sireci (1999) found the causes of DIF in verbal reasoning that were associated with specific item types. These causes are: (a) item is adapted from source to target language, (b) translation is not correct, (c) the format does not remain exactly the same, (d) the words do not have the same level of difficulty, and (e) there are some differences in culture relevance. This study not only tells us the importance of identifying of DIF items in cross-culture testing, but also demonstrates the cause of DIF that are so crucial in development of translated tests and enhancing score validity. However, this study only had two languages involved, the generability of other languages is a big challenge.

Allalouf (2003b) examined item revision as a tool for improving test adaptations. A panel of eight translators and researchers are formed to revise the items shown DIF in author's previous study. The author found revising items (a) can retain translated items and maintain item bank size, (b) provides a better understanding of the sources of DIF, and (c) determines which revision is more effective. This study created an empirically based guideline for future studies. That is, the cause of DIF could have been eliminated earlier during the translation process of an item. When DIF is found, implementing a revision design similar the study can eliminate or reduce DIF and improve the validity of adapted tests.

Evaluating Test Equating Accuracy

The purpose of equating is to obtain comparable scores that well estimate the underlying true scores. How good are true scores estimates and to what extent are the equated scores are comparable? In this section, a review of previous studies of double

liking method and standard errors of equating method on equating accuracy will be presented.

Double Linking Method

Kolen and Brennan (1995) proposed a double linking procedure for common-item non equivalent groups design. This procedure that is often used to solve the problems associated with developing linkage plans, that is, to use two old forms to equate new forms. It provides a built-in stability check on the equating process leading to greater equating stability. With two links, a second link still is available to be used for equating even if the strong statistical assumptions required under this design are violated for one of the links. In addition, in anchor item design if a significant number of common items on one link is found to have problems, or if security problems are discovered with one of the versions of the test, then a second link still exists that can be used to conduct the equating. Therefore, double linking method is strongly recommended when feasible. However, double linking requires greater effort in test development and in equating than does equating using a single link.

Rapp and Allalouf (2002) used double linking method for evaluating and validating cross-lingual equating for test adaptation. In this method, a new test form is independently equated to two old forms. Then the two conversion functions are averaged to produce a single conversion. If the two conversion functions differ more than would be expected by chance, it would suggest that a systematic error occurred in at least one of the equating processes. This method provides a built-in check on equating and leading to greater equating stability for cross-lingual studies; however, it also introduces more complications into the equating process.

Standard Errors of Equating (SEE)

There are two general sources of error in estimating relationship when equating is conducted: random error and systematic error (Kolen & Brennan, 1995). Standard errors of equating provide estimates of the amount of error due to sampling examinees. As the size of the sample approaches infinite, the standard errors of equating approach to zero (Harris & Crouse, 1993). Crouse (1991) compared the accuracy of equating conducted using various methodologies for three data collection designs. They used bootstrapping to obtain estimates of error. Their single-group counterbalanced design was chosen as their criterion. However, the fact they employed real data in the study prevented them from knowing the true equating conversions.

Loret (1975) described the method used to empirically estimate the standard errors of linear and equipercentile equating for the anchor test study. The equatings were replicated eight times on half samples of schools. Error was defined as the square root of the average squared deviations of the equivalents determined by each of the eight replications. These errors provided a basis for evaluating the equating results for the seven standardized reading tests studied.

Zeng (1993) estimated the standard errors of linear equating for the single group design with and without the normality assumption. A computer simulation was generated to obtain bootstrap standard errors, and a real data example was used to evaluate the behavior of the estimated standard errors.

Summary

Literature reviews of test adaptation confirm that adapting an existing instrument instead of developing a new one has many advantages (Hambleton, 1993; Hambleton &

Kanjee, 1995). That is the reason why test adaptation is very important in cross-lingual assessment. Hambleton (1992) distinguished the difference between test adaptation and test translation. However, among all these reviews, the disadvantages of test adaptation have been widely discussed (Lonner & Berry, 1986; van de Vijver & Pootinga, 1991). Cultural and language differences are the major concern. They argued that it is very difficult to impose conclusions based on the concepts in one cultural that may not exist in the other culture (Fouad, 1994). Some studies found that the test and test items do not systematically discriminate against one culture groups or the other. Therefore, validity and reliability studies in test adaptation should take more care than the other studies (Froud, 1993; Geisinger 1994).

ITC (International Test Commission) developed the *Guidelines for Adapting Educational and Psychological Tests* and later summarized by Hambleton (1994, 2001). These guidelines provide guidance regarding the adaptation process and encourage test developers to conduct statistical analysis to check cross-lingual equivalence. Some critical guidelines for test adaptation have been summarized (Sireci 1999).

Test equating has been defined in several ways. Lord's equity requires that the conditional distributions of scores on each test after equating must be equal (Lord, 1980). Weak equity or tau equivalence requires that the means of the conditional distributions of scores on each test after equating being equal (Divgi, 1981; Yen, 1983). Equipercentile equating only requires that the corresponding percentile ranks in any given groups are equal (Angoff, 1971).

When selecting an equating method, it is expected to be highly accurate in that equating method function better for that particular test. That is, test equating assumptions,

data collection methods, and ways to evaluate test equating should take into consideration before choosing an equating method (Crocker & Algina, 1986). There are three linking designs for cross-lingual equating and the monolingual groups design is the most popular one (Cook 2000; Hambleton, 1993, 1994; Sireci, 1997; Hambleton & Patsula, 1998; Wainer, 1999).

Studies regarding anchor item design showed that whether the anchor items are representative the overall test is very important (Busescu, 1985; Kolen & Brennan, 1987). Many studies summarized different rules of thumb in choosing anchor items under different conditions (Angoff, 1984; Lord, 1984; Wright, 1977; Mckinley & Reckase, 1981; Mebride & Weiss, 1974; Hills, Subhiyah, & Hirsch, 1988; Raju, Edwards, & Obsberg, 1983). Differential item functioning (DIF) study and test equating cannot be treated as two separate issues (Angoff & Cook, 1988). Items with DIF not only increase the errors of test equating, but could also be biased towards some examinees. Therefore, studies of DIF in anchor item design focus on detecting the differential statistical properties in order to delete DIF items before performing test equating. On the other hand, studies of test equating usually assume that all items are free from DIF influences. There were many studies summarized the causes of DIF in test adaptation (Gierl & Khaliq, 2001; Allalouf, Hambleton, & Sireci, 1999). One study found that revision DIF items instead of deleting them can improve the validity of the adapted tests (Allalouf, 2000).

There are many studies on how to evaluate the accuracy of equating. Double linking method provides a built-in stability check on the equating process leading to greater equating stability (Kolen & Brennan, 1995). Further, a study applied this double

linking method for evaluating and validating cross-lingual equating in test adaptation study (Rapp & Allaloff, 2000). Standard error of equating (SEE) is an important method in evaluating the accuracy of equatings as well. Several studies addressed standard errors of linear equating using different designs (Loret, 1972; Zeng, 1991). All these studies preferred smaller errors to larger errors. However, whether the magnitude of the differences between standard errors is important or whether the size of the errors themselves is large, is a subjective determination (Harris & Crouse, 1993).

CHAPTER THREE

Research Methodology

Chapter Three has two major parts. Part one presents a detailed description of the research design including research instrument, research subjects, items and contents, and data collection procedure. Part two provides a detailed description of the basic data analysis employed in this study including ways of choosing anchor items, equating methods, and anchor tests evaluation methods.

Research Design

Instrument

The test used in this study is a two-form and three target language version of a certification test. Based on the available test data, there are only three target language groups that will be investigated in this study: Chinese, Korean and Spanish. The test data are the scores on the two forms of each target language. The source language (SL) and target language (TL) versions of the tests are selected from a total of 8322 examinees taking the source language version of the test and another 620 examinees taking the target language versions of the test. Both SL and TL versions of the test consist of 160 items in each test.

The tests are administered using computers. The two groups taking different test forms in each language are randomly formed. The test forms are comprised of four-alternative multiple choice questions. The items are administered in random order, and all the item responses are scored as correct or incorrect (coded as 1 or 0, respectively). There are different anchor items in different anchor tests and all the anchor items are identically

embedded in each TL and SL in terms of location. The anchor items are chosen from the total items and are a part of the total test. The stem, alternatives, and stimulus materials for the anchor items are identical for the two versions of the test in three different TLs.

The test forms generally meet the equating requirements that were previously mentioned in the review of equating guidelines. All tests have sufficient numbers of items and all the items are reasonably long. The test items are administered and secured under standardized conditions. Some items have been administered in previous years under the same standardized testing situations and found to be satisfactory. In addition, the scoring keys are clear and consistent for the two forms of the test in different TLs.

Subjects

A total of nearly 9,000 (8322 + 620) examinees took the test in both SL version and TL version over a period of several years. The data obtained for this study contained no identifiers of individuals who took the tests. Examinees taking the test in SL all took the test in the United States; examinees taking the test in TLs all took the test in the TL countries. Table 1 through Table 3 presents the number of examinees from TLs and SL; Table 4 and Table 5 show a breakdown of TLs and SL, respectively. In Table 4 and Table 5, SL and TL are matched by year.

Table 1

Number of Examinees for Korean Language and English Language

TL One	Frequency (Form1)	Percentage (Form1)	Frequency (Form2)	Percentage (Form 2)
English Language	875	92	1422	92
Korean Language	71	8	123	8
Total	946	100	1545	100

Table 2

Number of Examinees for Spanish Language and English Language

TL Two	Frequency (Form1)	Percentage (Form1)	Frequency (Form2)	Percentage (Form 2)
English Language	1441	93	1455	96
Spanish Language	116	7	64	4
Total	1557	100	1519	100

Table 3

Number of Examinees for Chinese Language and English Language

TL Three	Frequency (Form1)	Percentage (Form1)	Frequency (Form2)	Percentage (Form 2)
English Language	1463	93	1677	93
Chinese Language	116	7	130	7
Total	1579	100	1807	100

Table 4

Comparison of Number of Examinees for Different Target Languages

TL	Frequency	Percentage
Chinese (Year 1999-2000)	246	40
Spanish (Year 2000-2001)	180	29
Korean (Year 2002-2003)	194	31
Total	620	100

Table 5

Comparison of Number of Examinees for Source Language in Different Years

SL	Frequency	Percentage
English (Year 1999-2000)	3130	38
English (Year 2000-2001)	2895	35
English (Year 2002-2003)	2297	27
Total	8322	100

To become certified, the examinees are strongly encouraged to participate in training programs before they start to take the formal certification tests. Since the training program provides examinees valuable insights regarding the formal certification tests, it is assumed that all the examinees from all countries had knowledge before formal tests were administrated.

Item Format and Test Specification

The test consists of items in a multiple choice (MC) format and divided into two major sections: non-verbal and verbal. In each version of the test there are four response options and all items are scored dichotomously. Each item can be located in the content specification that it belongs. In section one there are five verbal content specifications; there is only one non-verbal content specification covered in section two. Table 6 shows the test specifications.

Table 6

Number of Items by Content Specification

	Content Specification	Items
Verbal	Management and Leadership in Quality Engineering	19
	Quality System Development, Implementation, and Verification	19
	Planning, Controlling, and Assuring Product and Process Quality	33
	Reliability and Risk Management	11
	Problem Solving and Quality Improvement	25
Non - Verbal	Quantitative Methods	53
	Total	160

Data Source

The test data were provided by American Society of Quality to the researcher upon request for the purpose of this research study. The tests were administered in three TLs from year 1999 to year 2003.

Data Analysis Procedure

In evaluating the psychometric properties of tests that are adapted into multiple languages three types of empirical analyses are typically utilized (Sireci, Harter, Yang, & Bhola, 2000). First, descriptive analyses are conducted to provide preliminary information on the impact of the exam (e.g., differences in average exam performance across language groups), the reliability of the scores from each version of the test, and the functioning of each item in each language (within-language item analysis). Second, dimensionality analyses are carried out to assess the equivalence of the dimensional structure across the source language and target language versions of the test. Third, differential item functioning (DIF) and test equating will be analyzed to assess the differences in item difficulty across language groups as well as to place two versions of the test into the common scale to identify the potential translation problems or other sources of item bias. In this study, the following data analysis procedures will be employed.

Phase 1: A Preliminary Study of the Data

A preliminary inspection of test data is the procedure used to find whether the two versions of the test in each language are parallel within languages and across languages. By simply comparing the mean and standard deviation between two versions and two forms of the test we would expect them to have the same mean and standard deviation (within sampling error). A two-way ANOVA using form (Form A vs. Form B) as one factor and language (source language vs. target language) as the other factor will be employed to examine whether the means and standard deviations differ. If mean and standard deviation are the same between two versions of the test, it is likely that the two

groups are similar in their abilities; otherwise the two groups may differ in their abilities. Also, if the mean and standard deviation are the same for the two forms of the test, we can conclude that the two test forms are identical. A F-test will be employed as well to examine whether there is a significant difference exists in the variability of the groups and variability of the tests.

Kolen and Brennan (1995) suggest that with the common-item non equivalent groups design, mean differences between two groups of approximately 0.1 or less standard deviation unit on the common items seem to cause few problems for any of the equating methods. Mean group differences of around 0.3 or more standard deviation unit can result in substantial differences among methods, and differences larger than 0.5 standard deviation units can be especially troublesome. Additionally, ratios of group standard deviations on the common items of less than 0.8 or greater than 1.2 tend to be associated with substantial differences among methods. Differences in group standard deviations have the potential to lead to differences among methods that are at least as great as those caused by differences in means.

Phase 2: Investigating Reliability and Validity of the Items

Reliability and validity are two very important issues in test adaptation. If a test gives different results at different versions of a test, the results may indicate that the test is not valid. In addition, it is impossible for an adapted test give the same results over time but not measure what it supposes to measure.

Reliability

The reliability of measurement refers to "the degree to which test scores are free from errors of measurement" (AERA, APA, & NCME, 1999, p. 19). The two most

frequently reported indices of reliability are the standard error of measurement and the reliability coefficient. The standard error of measurement (SEM) is a measure of the extent to which an individual's scores vary over numerous parallel tests. It is the standard deviation of an individual's scores if he or she took numerous parallel tests. The standard error of measurement (SEM) is estimated using the following formula:

$$\text{SEM} = \text{SD}\ \sqrt{1 - r_{xx}} \qquad (2)$$

where SD is the standard deviation of observed scores for a single test and r_{xx} is the reliability coefficient for the test.

Among the several approaches to estimate the reliability of a test, Cronbach's alpha is probably the most frequently used. It is a measure of internal consistency (i.e., how homogeneous test items are) appropriate for a test containing only multiple choice (MC) items.

In this study, Cronbach's alpha will be examined between two forms and two versions of the test in each TL. Moderate to high reliability indices are desirable. Also, each pair of forms and each pair of tests should have similar degrees of internal consistency.

Validity

The validity issue is the most important psychometric property of any measurement and this is true for cross-lingual assessment as well. Oosterhof (2001) defined "validity pertains to the degree to which a test measures what is supposed to measure. Validity is the most central and essential quality in the development, interpretation, and use of educational measures" (p. 45).

van der Vijver and Tanzer (1997) provided guidance to cross-cultural researchers for evaluating translated instruments for validity issues. In providing their taxonomy of test validity in cross-cultural assessment, they discussed three levels of equivalence. The first level of equivalence is construct equivalence, which signifies that the same construct is measured by instruments in all cultural groups. The second level of equivalence is measurement unit equivalence, which occurs when the assessments are measuring the same construct using a common metric, but the origin of the metric differs, such as in the case of the Farenheight and Celsius temperature scales. The third level of equivalence is scalar equivalence, which occurs when all assessments are measuring the same construct using the "same measurement unit and same origin" (p. 266).

Construct equivalence is most often established through rational analysis and familiarity with the cultural groups being assessed. The primary issue to be resolved is whether the construct to be measured exists in all cultures and can be measured in an equivalent manner. Measurement unit equivalence and scalar equivalence are more difficult to establish. Therefore, many test specialists and cross-cultural researchers have stressed the need to ensure construct equivalence in different language versions of an assessment (e.g., Geisinger, 1994; Hambleton, 1993, 1994; Sireci, 1997, in press; van der Vijver & Poortinga, 1997). For example, the Guidelines for Adapting Educational and Psychological Tests developed by the International Test Commission stipulate that

> instrument developers/publishers should apply appropriate statistical techniques to establish the equivalence of the different versions of the instrument (Hambleton, 1994, p. 232).

This requirement relates to construct equivalence. That is, if a test lacks construct comparability, it can lead to test bias, which implies that inferences derived from test scores are not equivalent across groups. In this study, the data set will be analyzed to investigate construct equivalence using SPSS principle components analysis (PCA) and compare the results with parallel analysis (PA).

Phase 3: Choosing Anchor Test Items

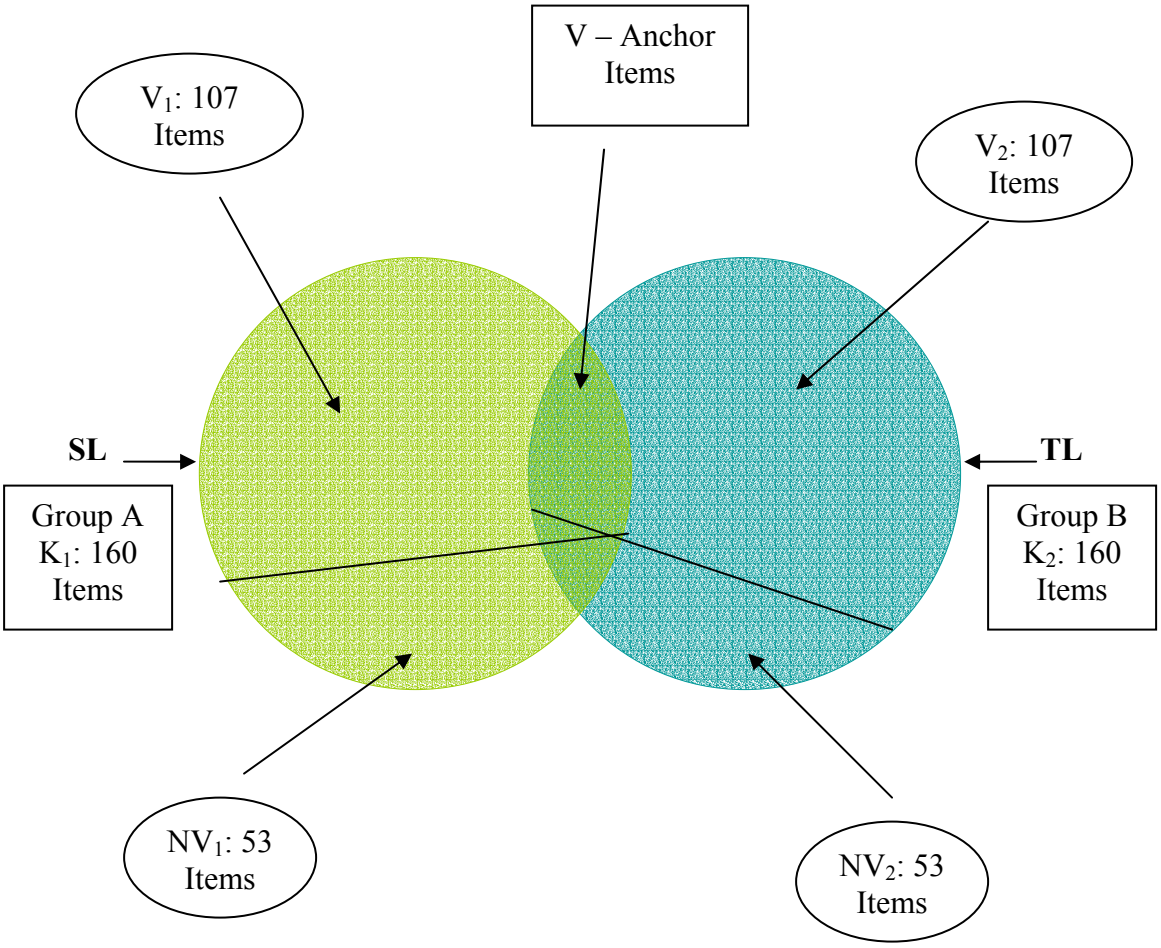
Anchor items are the test items that represent for various subject area, instructional level, instructional objective measured, and various pertinent item characteristics (e.g., item difficulty and discriminating power) (Gronlund, 1998). There are two kinds of anchor items: internal anchor items and external anchor items. Items that contribute to the examinees total scores are referred to as internal anchor items. Internal anchor items are often interspread throughout a test. External items are the items that do not contribute to the examinees total scores and are frequently administered as a separate and timed block of items (Kolen & Brennan, 1995). Internal anchor items will be used in this study.

The fundamentals of anchor item design in this study can be perceived as similar to Angoff's (1971) Equating Design IV or Equating Design VI. The essentials of these designs are as follows: source language test X is administered to Group A; target language test Y is administered to Group B. Different test versions X and Y have a set of items in common (i.e., the anchor items). The anchor items are administered to both group A and B and are used to adjust differences that exist between the two versions of the test. This is illustrated in Figure 3. The rationale underlying almost all the equating methods under this design is:

1. The sample statistics of test X are projected to the group that takes only test Y through the relationship between test X and the common set, V. This is then repeated for test Y.
2. A synthetic group (sometimes called the synthetic population) is formed as a weighted combination of the groups that completed tests X and Y. The sample statistics for tests X and Y are projected to the synthetic group.

Figure 3

Simple Pattern for Common-Item Non Equivalent Group Design



The selection of anchor items in test adaptation can depend on both practical considerations and statistical considerations. The following criteria will be utilized for choosing proper anchor items in this study: (1) content representations; (2) 20 items or 20% of the total items, whichever is larger; (3) best translation; (4) identifying item difficulty and item discrimination; and (5) using delta plot analysis. Practical consideration means those criteria (1) or (2) or (3) will be examined, while statistical consideration is related to criteria (4) or (5). However, statistical analysis does not make anchor items identical in context, it only maximizes the similarity of the common items (Allalouf & Rapp, 2002). Therefore, the anchor set that is eventually used is a combination of both practical consideration and statistical consideration, that is, the items must be content representative and limited to certain numbers of the items. In addition, statistical analysis or best adaptation practices need to be examined as well.

In this study, the combinations of (1) (2) (4) and of (1) (2) (5) of the above will be analyzed. Combination of (1) (2) (3) can not be examined in this study due to the difficulty of getting actual test items to identify the test adaptation procedure.

Content Representation

Content representation means that anchor tests should be built to have the same specifications proportionally as the test itself. Klein and Jarjoura (1985) defined content representation as a match between anchor test and total test of the percentage of items in each of the several content areas. They concluded that content representativeness of anchor items was critical to equating accuracy. In addition, in constructing anchor items, the number of anchor items should be long enough to adequately represent test format (Kolen & Brennan, 1995). It may seem safest to use a long, content-representative anchor

having item statistics that reflect the item statistics of the total test. In a practical setting, however, this may not be possible. If items are frequently updated or changed, the questions required for long content-representative anchors may no longer exist in the item pool.

Cook and Peterson (1987) reported that inadequate content representation of the common-item set creates especially serious problems when the examinee groups that take the alternate forms differ considerably in achievement. In addition, serious problems can result if the contexts in which the common items appear differ from the old form to the new form, or from one version of the test to the other of version of the test. One way to avoid having the common items function differently in the two groups is to administer common items in approximately the same position between the two forms or the two versions of the test.

In this study, content representation focuses on tightly defined content areas within the test, all of which fall within a somewhat restricted domain. In other words, the content areas correspond to the table specification used to assemble the test. Additionally, these common items were in about the same positions.

The Number of Anchor Items

The number of the anchor items used should be considered on both content and statistical grounds. Budescu (1985) and Wingersky et al. (1987) concluded that too few items could lead to many equating problems, while large numbers of anchor items would lead to less random equating error.

Some studies support the very small anchor item design. For example, Harris (1993) conducted a simulation study using a small pool of items and recommended that a

small number of anchor items could lead to the same results as a large number of anchor items could. However, because educational tests tend to be rather heterogeneous, a large number of anchor items are likely required for adequate equating in practice.

Kolen and Brennan (1995) suggested a rule of thumb that the number of anchor items should be at least 20% of the length of a total test containing 40 or more items, unless the test is very long, in which case 30 anchor items might be enough. Another rule of thumb for the minimum length of the anchor items is 20-25% of the number of items on either of the tests (Woldbeck, 1998). Angoff (1971) proposed 20% of the total length or 20 anchor items, whichever is greater. In this study, Angoff's suggestion will be utilized as a minimum number of anchor items.

Best Translation

Test adaptation and test translation are two very important tasks in cross-lingual study. Test translation is part of test adaptation and is a very important procedure of test adaptation as well. Here we focus only on test translation in test adaptation procedure.

Translation is a kind of activity which inevitably involves at least two languages and two cultural traditions (Toury, 1978). Newmark (1988) defined culture as the way of life and its manifestations that are peculiar to a community that uses a particular language as its means of expression. Vermeer (1989) stated that language is part of a culture, therefore, language and culture can be seen as being closely related and both aspects must be considered for translation.

An instrument sometimes can be translated on a question-by-question basis, however, at other times, it must be translated only in concept (Gersinger, 1994). There are four types of test translations (Casagrande, 1954): (a) Pragmatic translation: the sole

interest lies in communicating accurately in the target language what was contained in the source language; (b) Aesthetic-poetic translation: the purpose of which is the evocation of moods, feelings and affect in the target language that are identical to those evoked in the source language; (c) Ethnographic translation: is aimed at maintaining the meaning and the cultural content of the source language in the target language; (d) Linguistic translation: is concerned with equivalence of meanings of both morphemes and grammatical forms of the two languages.

Best translation in psychological instruments must be concerned with evaluating translations of ability tests, measures of attitudes, interests that are designed to assess individual differences (Hulin, Drasgow, & Parsons, 1983). They claimed that translations carried out in this area would most likely be classified as ethnographic translations although it does not fit with this category perfectly. Translators producing these translations must be familiar with both the source and target cultures as well as with the source and target languages. They must know how words and phrases are interpreted in a culture and use them appropriately in the translated version.

Item Difficulty and Item Discrimination

Item difficulty is simply defined as the percentage of students taking the test who answered the item correctly. The larger the percentage responding correctly the easier the item. The higher the difficulty index, the easier the item is understood to be (Wood, 1960). To compute the item difficulty, divide the number of people answering the item correctly by the total number of people answering item. The proportion for the item is usually denoted as p and is called item difficulty (Crocker & Algina, 1986). For example, an item answered correctly by 85% of the examinees would have an item difficulty, or p-

value, of .85, whereas an item answered correctly by 50% of the examinees would have a lower item difficulty, or p value, of .50. Item difficulty ranges for 0 to 1. Zero item difficulty means that no one answered the item correctly, whereas item difficulty of 1 means that all examinees answered the item correctly.

Item difficulty is basically a behavioral measure. Rather than defining difficulty in terms of some intrinsic characteristic of the item, difficulty is defined in terms of the relative frequency with which those taking the test choose the correct response (Thorndike et al, 1991). One cannot determine which item is more difficult simply by reading the questions. One can recognize the name in the second question more readily than that in the first. But saying that the first question is more difficult than the second, simply because the name in the second question is easily recognized, would be to compute the difficulty of the item using an intrinsic characteristic.

Another implication of item difficulty is that difficulty is a characteristic of both the item and the sample taking the test. For example, an English language test item that is very difficult for an elementary student could be very easy for a high school student. Item difficulty also provides a common measure of the difficulty of test items that measure completely different domains. It is very difficult to determine whether answering a history question involves knowledge that is more obscure, complex, or specialized than that needed to answer a math problem. When item difficulties are used to define difficulty, it is very simple to determine whether an item on a history test is more difficult than a specific item on a math test taken by the same group of students.

Item difficulty has a profound effect on both the variability of test scores and the precision with which test scores discriminate among different groups of examinees

(Thorndike et al, 1991). When all of the test items are extremely difficult, the great majority of the test scores will be very low. When all items are extremely easy, most test scores will be extremely high. In either case, test scores will show very little variability. Thus, extreme item difficulties directly restrict the variability of test scores.

Item discrimination refers to its ability to distinguish between more and less knowledgeable students (Oosterhof, 2001). That is, if the test and a single item measure the same thing, one would expect people who do well on the test to answer that item correctly, and those who do poorly to answer the item incorrectly. A good item discriminates between those who do well on the test and those who do poorly. The higher the discrimination index, the better the item because such a value indicates that the item discriminates in favor of the upper or more knowledgeable group, which should get more items correct. Two indices can be computed to determine the discriminating power of an item, the item discrimination index and discrimination coefficients.

In computing the discrimination index, first, score each student's test and rank order the test scores. Next, the 27% of the students with the highest scores and the 27% with the lowest scores are separated for the analysis. Wiersma and Jurs (1990) stated that "27% is used because it has shown that this value will maximize differences in normal distributions while providing enough cases for analysis" (p. 145). There need to be as many students as possible in each group to promote stability, at the same time it is desirable to have the two groups be as different as possible to make the discriminations clearer. Although Nunnally (1972) suggested using 25%, according to Kelly (1981) the use of 27% maximizes these two characteristics.

The discrimination index is the number of people in the upper group who answered the item correctly minus the number of people in the lower or less knowledge group who answered the item correctly, divided by the number of people in the larger of the two groups. Wood (1960) stated that

> when more students in the lower group than in the upper group select the right answer to an item, the item actually has negative validity. Assuming that the criterion itself has validity, the item is not only useless but is actually serving to decrease the validity of the test. (p. 87)

A negative discrimination index is likely to occur when an item covers complex material and is written in such a way that it is possible to select the correct response without any real understanding of what is being assessed. A less knowledgeable student may make a guess, select that response, and come up with the correct answer. More knowledgeable students may be suspicious of a question that looks too easy, may take the harder path to solving the problem, read too much into the question, and may end up being less successful than those who guess. As a rule of thumb, in terms of discrimination index, .40 and greater are very good items, .30 to .39 are reasonably good but possibly subject to improvement, .20 to .29 are marginal items and need some revision, below .19 are considered poor items and need major revision or should be eliminated (Ebel & Frisbie, 1986).

Two additional indicators of the item's discrimination effectiveness are point biserial correlation and the biserial correlation coefficient. The choice of correlation depends upon what kind of question we want to answer. The advantage of using discrimination coefficients over the discrimination index (D) is that every person taking

the test is used to compute the discrimination coefficients and only 54% (27% upper + 27% lower) are used to compute the discrimination index.

Biserial correlation coefficients (r_{bis}) are computed to determine whether the attribute or attributes measured by the criterion are also measured by the item and the extent to which the item measures them. The r_{bis} gives an estimate of the well-known Pearson product-moment correlation between the criterion score and the hypothesized item continuum when the item has been dichotomized into right and wrong (Henrysson, 1971). Ebel and Frisbie (1986) state that the r_{bis} simply describes the relationship between scores on a test item (e.g., "0" or "1") and scores (e.g., "0", "1",..."50") on the total test for all examinees.

The point-biserial (r_{pbis}) correlation is used to find out if the right people are getting the items right, and how much predictive power the item has and how it would contribute to predictions. Henrysson (1971) suggests that the r_{pbis} tells more about the predictive validity of the total test than does the biserial r, in that r_{pbis} tends to favor items of average difficulty. It is further suggested that the r_{pbis} is a combined measure of item-criterion relationship and of difficulty level. Therefore in this study, point-biserial correlation will be used. It is calculated as the following:

$$r_{pbis} = \frac{\mu_+ - \mu_x}{\sigma_x}\sqrt{p/q} \qquad (3)$$

where μ_+ means of criterion score for examinees who get the item correct, μ_x is the mean score of the test for the entire group, σ_x is the standard deviation of the test for the entire group, *p* is the proportion of examinees who get the item correct, and *q* equals to 1-*p*. The point-biserial correlation is similar to Pearson correlation between an item score

and the total score (Crocker & Aligina, 1986). In this study, items that will be chosen as anchor items are the items that are of moderate difficulty and discriminate well.

Delta Plot Method

The Delta Plot method can be utilized to screen for DIF items. The non-DIF items that are closest to principal axis will be used as anchor items (Angoff, 1982). The principal axis is an orthogonal least square line that best fits the data symmetrically. It minimizes the sum of the squared deviations between the two variables so that the role of both variables is the same. In the delta plot method, the principal axis is also called equal difficulty line (Angoff, 1982). This method receives wide use by many testing agencies because it provides a useful impression of the functioning test items across two groups, especially when sample sizes are small and an analysis that can be completed quickly is needed (Robin, Sireci & Hambleton, 2003). The key to using this method successfully is to view the findings as exploratory. However, a well-known weakness of this method is that Type I and Type II detection errors are likely to increase when item discriminations are not homogeneous (Angoff, 1982; Camilli & Shepard, 1994; Dorans & Holland, 1993).

The history and suitability of the Delta Plot procedure is described by Angoff (1982). This method is also good for studying cultural differences (Beller, 1996). In its simplest form, items are deemed non-DIF when item difficulties from one group are perfectly correlated with difficulties in another group, thus creating a straight line in the scatter plot. This is illustrated by placing item difficulties for one group on the y axis and items difficulties for another group on the x axis (Crocker & Algina, 1986). The researchers look at the plots of item difficulties obtained in the two groups and identify

the anchor items, the items that are very close to the principal axis of the data. Because the item difficulties are ordinal measurements, it is customary to assume that the item difficulties were obtained by examinees from normal ability distributions, and report the item difficulties as normal deviates on a scale with mean and standard deviation equal to 13 and 4 respectively (referred to as "ETS delta values" after the organization that pioneered their use in test development work). Therefore, for a item difficulty of .50 the corresponding delta value would be 13. If the item difficulty were .84, the delta value would be 9.0, and for a item difficulty of .16, the delta value would be 17.0.

When using the delta plot method, there are three values that are necessary to identify possible anchor items: item difficulty, z-scores, and delta measures. First, item difficulty scores for the two different groups are computed on the items chosen. Second, z-scores are found by using a z-scores table and finding the cut off score for the item difficulty of each item. Third, the cut off z-scores for the p-values are then converted to a normal deviate with an arbitrary mean and standard deviation using mean as 13 and standard deviation as 4. The formula for calculating deltas, the transformed normal deviates is the following:

$$Delta = 4z + 13 \quad (4)$$

The pair of normal deviates, one pair for each item, are plotted on a bivariate graph to demonstrate possible items that are close to the principle axis are identified as anchor items (Fisk, 1991). When the groups are of the same type and of the same level of ability, the plot of these points will ordinarily appear in the form of an ellipse extending from lower left to upper right. This often represents a correlation of 0.98 or even higher, indicating that the rank orders of difficulty of the items is essentially the same in the two

groups. However, when the groups are drawn from different populations, the points will be dispersed in the off-diagonal direction and the correlation represented by the points will be lower. Delta plot analyses (Angoff, 1982, 1993) are the easiest to implement and can be done directly in a spreadsheet program such as Microsoft Excel or in a statistical package such as SPSS.

Phase 4: Levine Linear Equating Method for Anchor Item Design

The most basic of the equating methods is linear equating. Linear equating assumes that the two tests to be equated differ only with respect to means and standard deviations. The distributions of the raw scores for the two tests are assumed to be equal except for mean and standard deviation. Crocker and Algina (1986) define equivalent scores as those that "can be identified as determining the pair of scores, one on form X and one on form Y, that have the same z-scores" (p. 458). The conversion from one test to another is accomplished using additive and multiplicative constants in the forms of the following equation (Angoff, 1971) for a synthetic group R:

$$Y = AX + B \tag{5}$$

R is a weighted combination of the two groups X and Y. This equation is used for all of the designs, the only difference being the calculations of A and B.

In converting the above equation to a linear equating, a transformation is found such that scores on X and Y on R are said to be equated if they correspond to the same number of standard deviation units above and below the mean in R. See below for detailed description of synthetic group. The linear equating function is (Dorans & Lawrence, 1990):

$$L_p(y) = m_x + s_x / s_y (y - m_y) \tag{6}$$

where $L_p(y)$ is the linear equating function for Y to X, and m_x, m_y, s_x and s_y are means and standard deviations, respectively, of the score distributions of X and Y on R.

Among linear equating methods, two of the more popular methods are Tucker's equating method and Levine's equally reliable method. Woodruff (1989) looked at both Tucker and Levine methods and concluded that the Levine method was more sensitive to group differences than Tucker method. Woodruff also noted that the Levine method should be an appealing method because it permits large group differences on the anchor test.

Kolen and Brennan (1995) compared the relationship between Tucker and Levine equating methods and concluded that the Levine methods are more appropriate than the Tucker methods when groups are dissimilar. They suggested that one of the Levine methods should be chosen when it is known or strongly suspected that populations differ. However, if the groups are too dissimilar, then any equating is suspect.

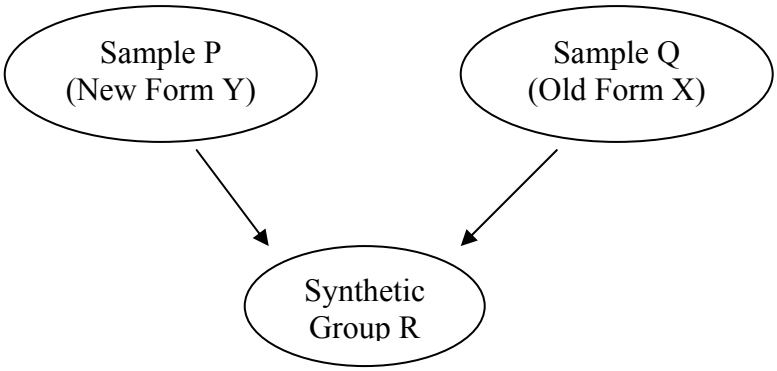
The Levine equally reliable linear equating method will be employed in this study to compare the different anchor tests. This method was originally developed by Levine in 1955, although he did not explicitly consider the concept of synthetic population. Kolen and Brennan (1987) formulated Levine equally reliable linear equating method by emphasizing the notion of a synthetic population, a combination of the proportionally weighted populations of examinees taking different test forms. The synthetic group is conceived of as containing two strata. Examinees from Form One are considered to be a random sample from stratum 1 and examinees from Form Two are considered to be a random sample from stratum 2. Weights w_1 and w_2 are used to weight the strata defining the synthetic group. This process is illustrated in Figure 3.

There are three ways to choose weights. One is to choose weights proportional to the sample size of examinees from each groups, that is, $w_1 = n_1 / (n_1 + n_2)$ and $w_2 = 1 - w_1$. Second, the weights are chosen to be equal, where $w_1 = w_2 = .5$. Third, synthetic group is defined as the new group, therefore $w_1 = 1$ and $w_2 = 0$ may be chosen.

From a practical perspective, the synthetic group that leads to the most direct score interpretation is preferable (Kolen & Brennan, 1987). When a new form is administered and scored, the focus of score interpretation is on the group that just took the new form. Since equating based on $w_1 = 1$ and $w_2 = 0$ allows a direct comparison and interpretation of how the new group performed on the new form to how the new group performed had it been administered the old form. Therefore, in this study, $w_1 = 1$ and $w_2 = 0$ of choosing weights will be utilized.

Figure 4

Synthetic Group Using Levine Linear Equity Method



$w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$

Given the sample P takes new-form Y and the set of anchor items V, sample Q takes old-form X and the set of anchor items V and sample R is a composite of P and Q, the linear equating method makes strong statistical assumptions as follows:

1. The true scores on Y and V are perfectly related, and the ratio of the standard deviation of true scores on Y to the standard deviation of the true scores on V is the same in P and R and the same as the true score on X and V.
2. The intercept of the regression line relating true scores on Y to true scores on V is the same in P and R and the same as the true score on X and V.
3. The standard error of measurement for Y and for V is the same for groups P and R and the same as the true score X and V.

Under these assumptions, the Levine equally reliable method is parameterized by:

$$A_L = Z_L / W_L \tag{7}$$

$$B_L = U_L - O_L \tag{8}$$

$$Z_L = \sqrt{\left[S^2{}_{xQ} + (S^2_{xQ} - S^2_{x^*Q})(S^2_{vR} - S^2_{vQ}) / (S^2_{vQ} - S^2_{v^*Q}) \right]} \tag{9}$$

$$W_L = \sqrt{\left[S^2{}_{vP} + (S^2_{yP} - S^2_{y^*P})(S^2_{vR} - S^2_{vP}) / (S^2_{vP} - S^2_{v^*P}) \right]} \tag{10}$$

$$U_L = M_{xQ} + (M_{vP} - M_{vQ}) \left[(S^2_{xQ} - S^2_{x^*Q}) / (S^2_{vQ} - S^2_{v^*Q}) \right] \tag{11}$$

$$O_L = A_L M_{yp} \tag{12}$$

where A_L and B_L are the parameters in equation 5. M and S refer to means and standard deviations respectively. Z_L and W_L are the estimate of variances of tests X and Y, respectively, on sample R, U_L is the estimate of the mean of X and R, and O_L is the

scale-adjusted estimate of the mean of Y on R in the standard deviation metric of X. Also, x^* y^* and v^* refer to the errors of measurement on the old form, the new form and equating test, respectively.

A common misconception holds that the Levine equally reliable equating method is a true-score equating method. In fact, it is not. It estimates observed-score means and standard deviations using assumptions about true-score regressions and standard errors of measurement and is an observed-score equating method based on assumptions about true scores.

The Levine equally reliable equating method will be facilitated by Common Item Program for Equating (CIPE) (Kolen, 1995) and confirmed by LEQUATE program (Waldron, 1988). The CIPE program is based on the frequency estimation methodology described by Kolen and Brennan (1995). The LEQUATE program displays the estimated means and standard deviations of Forms A and B for the synthetic population, as well as the slope and intercept of the equating line described by Kolen and Brennan (1987).

Phase 5: Mean and Sigma Equating Method for Equivalent Group Design

Mean and sigma equating method is the most commonly used method in equating (Kolen & Brennan, 1995). In this method, the linear conversion is defined by setting standardized deviation scores (z-scores) on the two forms to be equal such that

$$z_x = z_y^* \tag{13}$$

where z_x and z_y^* are the z-scores for Form X and Form Y respectively. Equating 13 can also be written as

$$\frac{x - \mu(X)}{\sigma(X)} = \frac{y^* - \mu(Y^*)}{\sigma(Y^*)} \tag{14}$$

where x is the raw scores for Form A and y is the equated or adjusted scores for Form B; $\mu(X)$ and $\sigma(X)$ are mean and standard deviation for Form X; $\mu(Y)$ and $\sigma(Y)$ are mean and standard deviation for Form Y. Solving for y in equating 14,

$$l_y(x) = y = \sigma(Y)\left[\frac{x-\mu(X)}{\sigma(X)}\right] + \mu(Y) \tag{15}$$

where $l_y(x)$ is the linear conversion equation for converting observed scores on Form X to the scale of Form Y. By arranging terms, an alternate expression for $l_y(x)$ is,

$$l_y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \tag{16}$$

where $\frac{\sigma(Y)}{\sigma(X)}$ is the slope (a) and $\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)$ is the intercept (b). Therefore, equation 16 also can be written as

$$y = ax + b\,. \tag{17}$$

Once a and b are determined, scores for Form X will be put on the same scale as scores for Form Y.

Phase 6: Anchor Tests Evaluation

The purpose of different equating is to obtain comparable scores that accurately estimate the underlying true scores. Therefore, a relevant question is: How good are the true score estimates and to what extent are the equated scores comparable in different anchor tests? This section focuses on a brief description of the double linking equating evaluation method and standard error of equating (SEE) evaluation method, and how these two evaluations on the accuracy of equating in different anchor test can be done.

Double Linking Method

Double linking is useful in equating into a common-item nonequivalent groups design. It provides a built-in check on the equating process and leads to greater equating stability (Kolen & Brennan, 1995). With two links, a second link is available to be used for equating if the strong statistical assumptions required under the common-item nonequivalent design are violated for one of the links. Also, if a significant number of common items on one link are found to have problems, then a second link exists that can be used to conduct the equating. Therefore, it is strongly recommended that double linking be used when feasible (Kolen & Brennan, 1995).

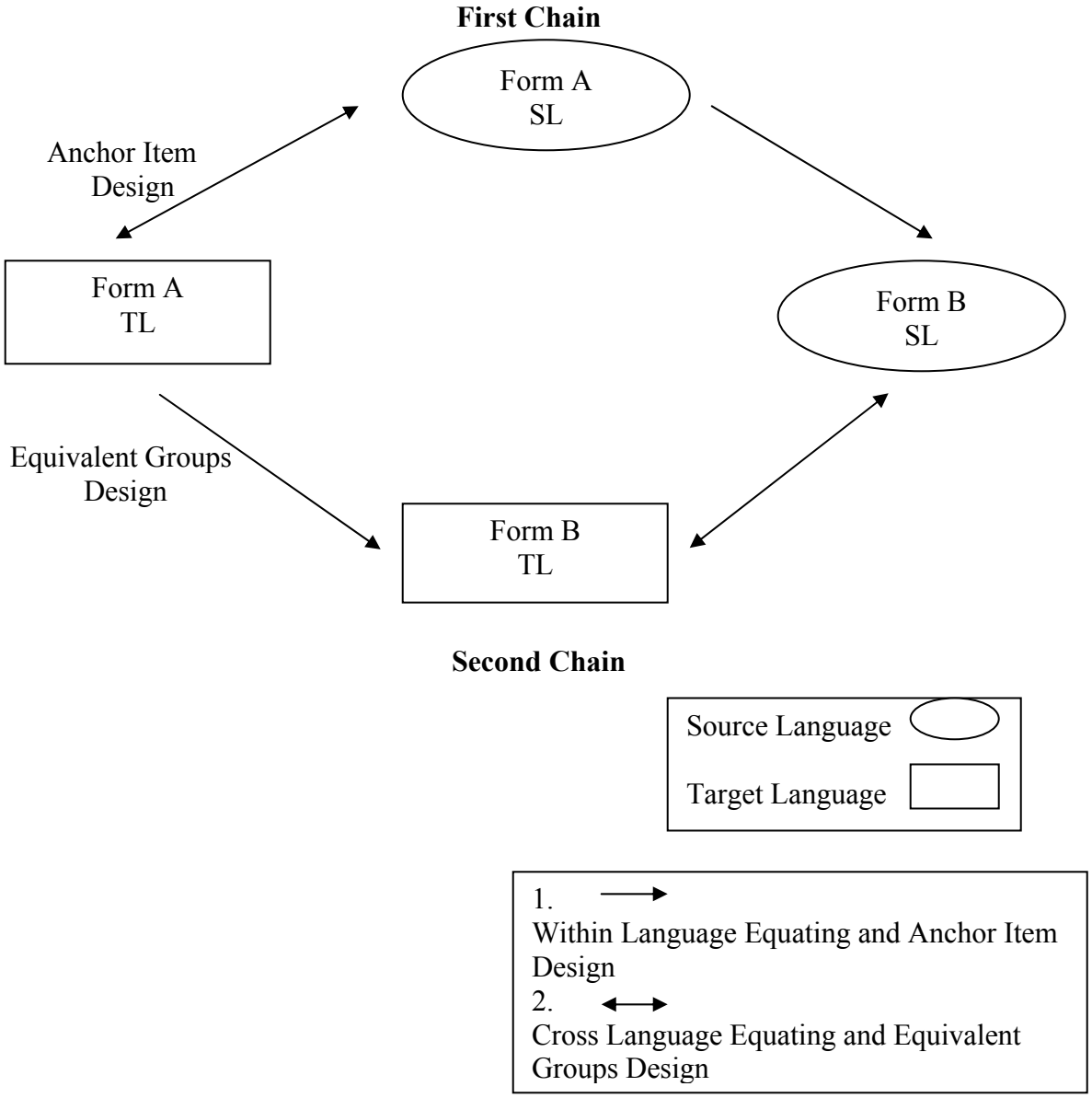
Rapp and Allalouf (2002) applied this double linking to cross-lingual equating cases. The minimal requirement of this method is two forms in each language version of the test. In this method, new translated forms must be assembled with two sets of translated items, each taken from a different source language form, and each represent the test in terms of content and statistical parameters. Then each of the two parallel sections of the translated exam form is equated to its respective section in the source language form. In the first chain, anchor items equate via section one in source language, while in the second chain, anchor items equate via section two in target language. Here, both chains consisted of one cross-lingual equating link and one within language link, however, in reverse order. The special contribution of this method is introducing the 'within-language' links between two sections in both source language and target language.

Applying this method using different data, a similar method can be summarized to evaluate equating in different anchor tests. For each anchor test, both "within language"

and “across language” links will be examined, however, in this case, the “within language” links will be investigated between forms. That is, items of TL in section one of Form A will be equated to items of TL in section one of Form B; items of TL in section two of Form A will be equated to items of TL in section two of Form B as well. Here section one and section two refer to verbal and non-verbal respectively. The “within language” will be executed using the “equivalent group design” and the “across language” will be examined using the “anchor item design”. We assume that within language equating link between two forms to be fairly stable. Therefore, the difference found between the equating results in the two chains for different anchor tests should be the differences between the cross-lingual equating links. Figure 5 is a diagram to show this double linking plan. In the first chain, the equating will be employed via SL in Form A, while in the second chain it is equated via TL in Form B. Therefore, both chains include one within language equating and one across language equating.

Figure 5

The "Double Linking" Plan



Standard Errors of Equating (SEE) Method

There are two kinds of standard errors in equating: random error and systematic error. Random equating error is present when the scores of the examinees that are

considered to be samples from a population or populations of examinees are used to estimate equating relationships. Systematic errors can occur in the following ways: (1) equating methods used introducing the bias in estimating the equating relationship; (2) statistical assumptions are violated in utilizing different equating methods; (3) improper implementation of data collection design in equating; and (4) the groups of examinees used to conduct equating differ substantially (Kolen & Brennan, 1995).

In this study, only random equating error will be examined. The amount of the random equating error associated with different anchor tests will be quantified by the standard error of equating. The pattern and behavior of standard errors of linear equating methods for the single-group, random-group, and common-item nonequivalent group designs have been researched widely and the results are well-known. The primary purpose of this criterion as utilized in this study is to provide some initial information about these standard errors for different anchor tests.

The mean standard error of equating (MSEE; Kolen & Brennan, 1995) is reported as a summary index of equating accuracy. This index can be used to compare the overall accuracy of different anchor tests. MSEE is defined as follows:

$$\sqrt{\sum_{i=1} f(x_i)\left[se^2(x_i)\right]} \qquad (18)$$

In Equation 18, the error variance at each score point i [$se^2(xi)$] is weighted by the relative Frequency $f(x_i)$ at the score point for the original sample examinees who took source language and then summed over score points. Weighting by the density is done so that the error variance for each examinee in the population is weighed equally. For

chained equating, the MSEE will be the sum of MSEE of the two component equatings (Braun & Holland, 1982).

CHAPTER FOUR

Results

This chapter reports the results for all relevant statistical analysis for the source language (SL) and for each target language (TL) group and responds to the research questions that were discussed in Chapter One. Data from nearly 9000 examinees were analyzed using SAS, SPSS, CIPE, and LEQUATE software to conduct the statistical analysis. The results were divided into the following sections for each language group: a preliminary study of the data for two forms, results of Principle Component Analysis, results of item difficulty indices and item discrimination indices, results of Delta Plot, results of equating using the Levine Equating method and the Mean - Sigma Equating method, and finally the results of Double Linking method and Mean Standard Errors of Equating to evaluate the equating accuracy for different anchor tests.

Target Language One - Korean Language

A Preliminary Study of the Data

The initial step in this phase of the analysis was a simple comparison of the means and variances of the scores distributions of the test across forms and languages. The summary statistics for these scores are presented in Table 7. In Form A, the mean of source language (M = 110.67) was more than 1 point higher than target language (M = 109.37), and the standard deviation of SL (SD = 17.64) was more than 1 point higher than TL (SD = 15.95) as well across languages. In Form B, the differences of mean and standard deviation were larger than in Form A. The mean of TL (M = 112.20) was six points higher than SL (M = 106.87), and the standard deviation of SL (SD = 18.60) was

three points higher than TL (*SD* = 15.22). Within English language, mean of Form A (*M* = 110.67) scored three points higher than Form B (*M* = 106.87), and standard deviation of Form A (*SD* = 17.64) was scored close to one point lower than Form B (*SD* = 18.60). In TL, Form B (*M* = 112.20) surpassed Form A (*M* = 109.37) three points of their means, and their standard deviations were about the same (*SD* = 15.95, SD = 15.22, respectively).

Table 7

Korean Language: The Statistics for Examinees Total Right Scores by Language and Test Forms

Statistics	English Language		Korean Language	
	Form A	Form B	Form A	Form B
Mean	110.67	106.87	109.37	112.20
Standard Deviation	17.64	18.60	15.95	15.22
Count	1422	875	123	71

A two-way ANOVA was also conducted at this preliminary stage with one factor being the language and the other being test form. The results are presented in Table 8. In this analysis, ANOVA was used to test the null hypotheses that there was no difference in the scores between languages or between forms. The interaction between the two factors (i.e., language and form) was also of interest.

As can be seen from Table 8, both factors of language groups, F (1, 2487) = 2.126, p = 0.145 and forms, F (1, 2487) = 0.125, p = 0.724 were not statistically significant, however, the interaction, F (1, 2487) = 5.726, p =0 .017 was statistically significant. Since each factor had only two levels, we can infer that the mean values for each level of the two factors were not significantly different from each other. The mean scores for SL in Form A were higher than that of SL in Form B. The order was reversed for the TL tests. The mean scores for TL in Form B were considerably higher than that of TL in Form A. The relationship between the means is presented graphically in Figure 6.

The effect size (eta squared) was also calculated to measure the magnitude of the difference between the language and the form. The eta squared formula is often used as an estimate of the effect size or the strength of association between the variables (Tabachnick & Fidell, 1996). According to this formula, eta square = 0.01 is a small effect size; eta square = 0.06 is a medium effect size, and eta square = 0.14 is a large effect size. Results showed that eta square was 0.002 for the interaction. This was a very small effect size, which indicated that the significant difference might be relatively unimportant.

Table 8

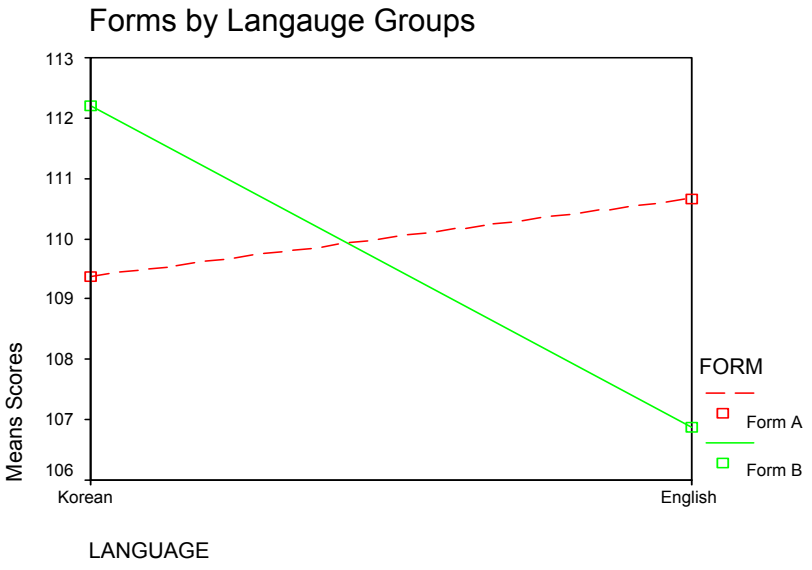
Korean Language: ANOVA Results for Total Number Right Score

Source of Variation	Sum of Squares	DF	Mean Square	F	p	Effect Size
Language	676.897	1	676.897	2.126	0.145	0.001
Form	39.708	1	39.708	0.125	0.724	0.000
Language * Form	1823.527	1	1823.527	5.726	0.017*	0.002
Error	791981.988	2487	318.449			
Corrected Total	800417.387	2490				

* significant at the 0.05 level.

Figure 6

Korean Language: Means Scores by Forms and Language Groups



Investigating Reliability and Validity

Results of Reliability Estimation of the Test Items

Internal consistency was examined for the languages and forms using a total of 160 items. The results are presented in Table 9. In Form A, the alphas were 0.9217 and 0.8914 for SL and TL, respectively for the verbal section of the test, and the alphas were 0.9077 and 0.9111 for SL and TL, respectively for the non-verbal section of the test. In Form B, the alphas were 0.9191 and 0.8977 for SL and TL, respectively for the verbal section of the test, and the alphas were 0.9123 and 0.9029 for SL and TL, respectively for the non-verbal section of the test. All these coefficient alphas for the languages and forms are very reliable, which means that the test items are quite homogeneous.

Table 9

Korean Language: Cronbach's Alpha by Language and Form

	English Language		Korean Language	
	Verbal	Non-Verbal	Verbal	Non-Verbal
Form A	0.9217	0.9077	0.8914	0.9111
Form B	0.9191	0.9123	0.8977	0.9029

Results of Validity Estimation of the Test Items

The construct validity was investigated using principle component analysis (PCA) and parallel analysis (PA) to determine the number of factors to extract. PA is a method

that generates random data from the same mean and the same standard deviation of item responses. Eigenvalues for both actual data using PCA method and generated data were computed and compared. If the eigenvalues of the real data exceed the eigenvalues of the random data then the factor would be retained (Thompson & Daniel, 1996).

For Form A, PCA was performed on all 158 items and extracted two components with eigenvalues of 13.56 and 4.70, respectively, and these two components accounted for 11.56% of the total variance. The eigenvalue of the third component that PCA produced was 2.43. The maximum eigenvalue that the PA produced was 4.81, and the next two eigenvalues were 4.72 and 2.37; therefore, two factors were retained. The scree plot can be found in Figure 21, Appendix B. The rotated component matrix using the Varimax method showed that the majority of items loaded on the first component and the second component, with 85.44% (135/158) of the items having a correlation coefficient larger than 0.10 and 26.58% (42/158) a correlation coefficient larger than 0.30. These two components represented the verbal section and non-verbal section of the test, respectively.

For Form B, PCA was also performed on all 156 items and extracted two components with eigenvalues of 13.23 and 5.23, respectively, and these two components accounted for 11.83% of the total variance. The eigenvalue of the third component that PCA produced was 2.01. The maximum eigenvalue that the PA produced was 5.89, and the next two eigenvalues were 5.14 and 1.96; therefore, two factors were retained. The scree plot can be found in Figure 22, Appendix B. The rotated component matrix using the Varimax method showed that the majority of items loaded on the first component and the second component, with 84.62% (132/156) of item having a correlation coefficient

larger than 0.10 and 29.49% (46/156) a correlation coefficient larger than 0.30, and these two components represented the verbal section and non-verbal section of the test, respectively.

Choosing Anchor Items

Anchor Test One - Results of Combination of (1) (2) and (4)

Anchor items should be a miniature version of the total test (Pansy & Kromrey, 1993). In this section, item difficulty indices and item discrimination indices were analyzed in selection of proper anchor items. Table 32 in Appendix C presents the item difficulty indices and item discrimination indices. First of all, the items that were discriminated well and had mild and similar difficulty indices were chosen as anchor items. Then, all of these anchor items must be a miniature of the total test. Since 20% of the total items were bigger than 20 items, 20% of the total items criterion was chosen as a minimum number of anchor items. In addition, anchor items should be administered in the same order for both language groups. For Form A, there were in total 36 items (23 items for V_1 and 13 items for NV_1) chosen; for Form B, a total of 36 items (23 items for V_2 and 13 items for NV_2) were selected. Here V_1 and V_2 stood for the verbal section of the test for Form A and Form B, respectively; NV_1 and NV_2 were the non-verbal section of the test for Form A and Form B, respectively. A summary of the number of anchor items for anchor test one is presented in Table 10. See Table 35 in Appendix E for a summary of how items are allocated in each content specification for anchor test one.

Anchor Test Two - Results of Combination of (1) (2) and (5)

Delta plot is a method based on item difficulty values that converted to a normal deviate with an arbitrary mean and standard deviation. For this method, items that were

closest to the equal difficulty line and also within the limit sets for the difficulty indices were chosen as anchor items. In addition, these anchor items must be a miniature of the total test and about 20% of the total items. Figure 27 though Figure 40 in Appendix D are the diagrams of the delta plot for all 160 items for Form A and Form B. Since it was difficult to identify each item when they were clustered together, the actual anchor items were visually selected from the delta plots of the items categorized by each content specification of the test. There were in total 36 items (23 items for V_1 and 13 items for NV_1) chosen as anchor items for Form A, and 36 items (23 items for V_2 and 13 items for NV_2) selected for Form B. A summary of the number of anchor items for anchor test two is presented in Table 10. See Table 35 in Appendix E for a summary of how items are allocated in each content specification for anchor test two.

Table 10

Korean Language: A Summary of the Number of Anchor Items for Two Anchor Tests

Form / Anchor Test	Anchor One	Anchor Two
Form A	V_1: 23	V_1: 23
	NV_2: 13	NV_2: 13
Form B	V_1: 23	V_1: 23
	NV_2: 13	NV_2: 13

Results of Levine Linear Equating

There were in total four Levine linear equating results for each anchor test: two for the verbal section and two for the non-verbal section of the test. In either the verbal or

the non-verbal section for each anchor test, an equating for the first link and an equating for the second link were conducted. Both of these two anchor tests used a common-item nonequivalent design. All detailed Levine Linear equating results are presented in Table 40 and Table 41, Appendix G for the verbal section and the non-verbal section of the test, respectively.

Table 11 provides a summary of all the statistics for the slopes and the intercepts pertaining to each anchor test. For the verbal section of anchor test one, the slope and intercept were 1.34 and -24.16, respectively for the first link; and 0.97 and 5.38 for the second link. For the non-verbal section of anchor test one, the slope and intercept were 1.13 and -5.78, respectively; and 0.96 and 2.11 for the second link. The slopes of 0.92 and 1.05 and intercepts of 5.69 and -3.87 for the first link and the second link, respectively were reported for the verbal section of anchor test two. In the non-verbal section of anchor test two, slope of 1.02 and intercept of -1.61 were for the first link; and slope of 0.96 and intercept of 1.62 were for the second link.

Table 11

Korean Language: A Summary of the Slopes and the Intercepts for Levine Equating of Anchor Item Design

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
Slope	1.34	0.97	1.13	0.96	0.92	1.05	1.02	0.96
Intercept	-24.16	5.38	-5.78	2.11	5.69	-3.87	-1.61	1.62

Results of Mean-Sigma Equating Method

The four Mean-Sigma equating results for each anchor test were reported: two for the verbal section and two for the non-verbal section. In either the verbal or the non-verbal section for each anchor test, there were two equatings for the first link and for the second link. Note that the equivalent groups design link for the first chain in the same for both anchor tests construction methods. The same is true for the second chain. Detailed Mean-Sigma equating results can be found in Appendix G.

For the verbal section of both anchor tests, the slope and intercept for mean-sigma method were the same: 0.98 and 5.51, respectively for the first link; and 1.20 and -14.03 for the second link. For the non-verbal section of two anchor tests, the slope and the intercept were 0.90 and 2.72, respectively for the first link; and 0.95 and -2.17 for the second link. Table 12 provides a summary of all these statistics.

Table 12

Korean Language: A Summary of the Slopes and the Intercepts for Mean-Sigma Equating of Equivalent Groups Design

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
Slope	0.98	1.20	0.90	0.95	0.98	1.20	0.90	0.95
Intercept	5.51	-14.03	2.72	-2.17	5.51	-14.03	2.72	-2.17

Results of Double Linking Equating Evaluation Method

Double linking provides a built-in check on the equating process and leads to greater equating stability (Kolen & Brennan, 1995). In this method, both "within language" and "across language" links were examined for each anchor test. The "within language" was executed using the equivalent group design and the "across language" was examined using the anchor item design. Table 13 provides a summary of all the statistics for the slopes and the intercepts pertaining to each anchor test. In order to calculate the slopes and intercepts of the first chain or the second chain for each anchor test, we need to compose the two first links or the two second links from Table 11 and Table 12.

For the verbal section of anchor test one, the slope and intercept were 1.31 and -18.17, respectively for the first chain; and 1.16 and -8.23 for the second chain. For the non-verbal section of anchor test one, the slope and intercept were 1.01 and -2.48, respectively; and 0.91 and 0.02 for the second chain. The slopes of 0.90 and 1.26 and

intercepts of 11.09 and -18.60 for the first chain and the second chain, respectively were reported for the verbal section of anchor test two. In the non-verbal section of anchor test two, slope of 0.92 and intercept of 1.27 were for the first chain; and slope of 0.91 and intercept of -0.46 were for the second chain.

Table 13

Korean Language: A Summary of the Slopes and the Intercepts for Double Linking Equating of Two Anchor Tests

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Chain	Second Chain	First Chain	Second Chain	First Chain	Second Chain	First Chain	Second Chain
Slope	1.31	1.16	1.01	0.91	0.90	1.26	0.92	0.91
Intercept	-18.17	-8.23	-2.48	0.02	11.09	-18.60	1.27	-0.46

Figure 7 illustrates the difference between the two different functions of two anchor tests for the verbal section of the test. For the verbal section of anchor test two, the difference between the two chains was from -10 to 30 raw score points; for the verbal section of anchor test one, the difference between the two chains was from -10 to 10 raw score points. The absolute mean difference for anchor test one and anchor test two in the verbal section were about 4.28 and 12.52 raw score points, respectively. As can be seen from Figure 7, anchor test one was closer to the zero perfect line than anchor test two. See Appendix H and Appendix I for detailed double linking statistics. Figure 69 through

Figure 72 in Appendix J present more detailed diagrams of the double linking results of the equating function differences between the two chains for the verbal section of two anchor tests.

Figure 7

Korean Language: Different Functions of Two Anchor Tests for Verbal Sections

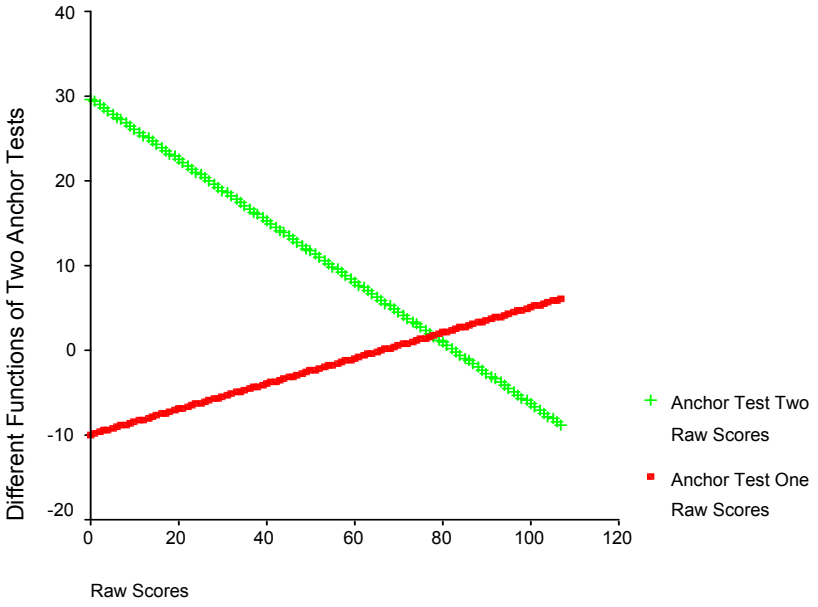
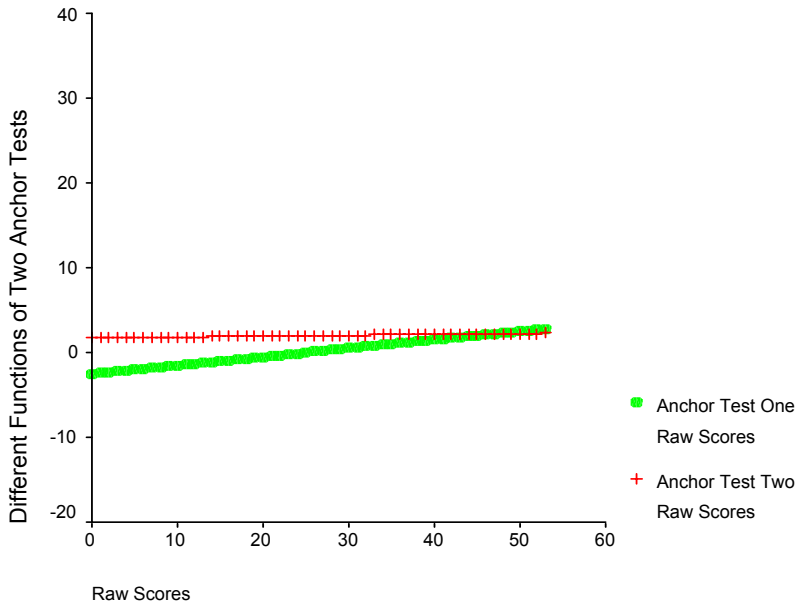


Figure 8 presents the difference between the two different functions of two anchor tests for the non-verbal section of the test. For the non-verbal section of anchor test one, the difference between two chains was from -2.5 to 2.75 raw score points; for the non-verbal section of anchor test two, the difference between two chains was from 1.75 to 2.15 raw score points. The absolute mean difference for anchor test one and anchor test two in the non-verbal section were about 1.35 and 2.00 raw score points, respectively. As can be seen from Figure 8, anchor test one was closer to the zero perfect line than anchor

test two. See Appendix H and Appendix I for more detailed double linking statistics. Also, Figure 73 through Figure 76 in Appendix J present more detailed diagrams of the double linking results of the equating function difference between the two chains for the non-verbal section of two anchor tests.

Figure 8

Korean Language: Different Functions of Two Anchor Tests for Non-Verbal Sections



Results of Mean Standard Error of Equating (MSEE) Evaluation Method

The values of the mean standard errors of equating are shown in Table 14. For the first chain in the verbal section, anchor test two (5.29) had higher SEE than anchor test one (3.49), but in reverse order for the second chain (6.48, 7.28, respectively). If we averaged these two links, anchor test two (5.89) scored slightly higher SEE than anchor test one (5.39). However, the MSEE (0.1178) for the verbal section of anchor test one

was similar to the MSEE (0.1080) for anchor test two. The diagram of SEE for the verbal section is presented in Figure 9.

For the first chain in the non-verbal section, anchor test one (1.96) scored slightly higher SEE than the anchor test two (1.93), but in the reverse order for the second chain (2.72, 3.06 respectively) as well. If we average these two chains, anchor test two (2.50) had higher SEE than anchor test one (2.34). However, the MSEE (0.0501) for the non-verbal section of anchor test two was also similar to the MSEE (0.0469) for anchor test one. The diagram of SEE for the non-verbal is presented in Figure 10. Detailed SEE results can be found in Table 54 and Table 55, Appendix K.

Table 14

Korean Language: MSEE between Two Anchor Tests

	Anchor Test One		Anchor Test Two	
	Verbal	Non-verbal	Verbal	Non-verbal
First Chain	3.49	1.96	5.29	1.93
Second Chain	7.28	2.72	6.48	3.06
Average	5.39	2.34	5.88	2.50
MSEE	0.1080	0.0469	0.1178	0.0501

Figure 9

Korean Language: SEE of the Verbal Section for Two Anchor Tests

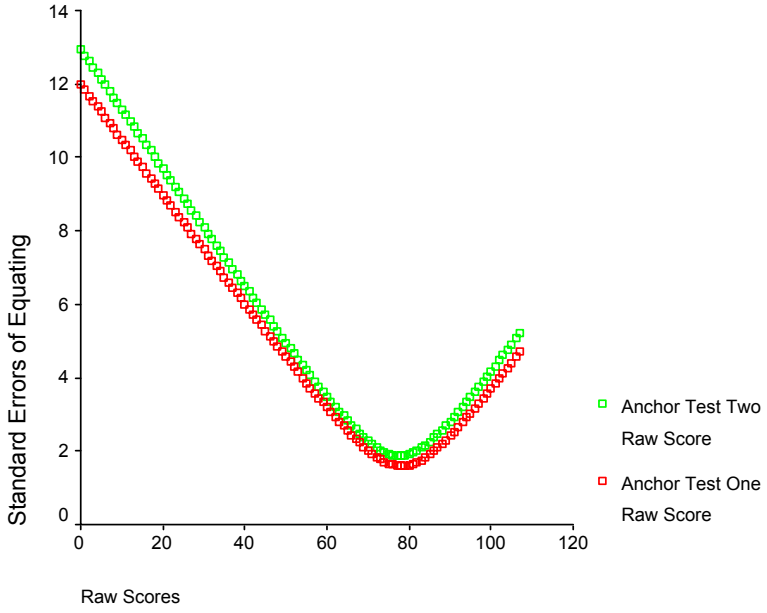
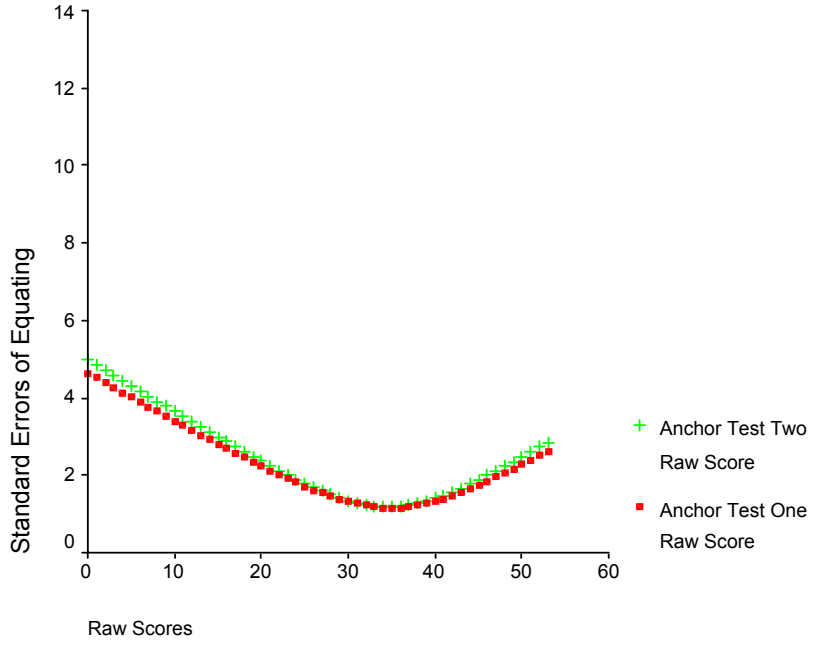


Figure 10

Korean Language: SEE of the Non-Verbal Section for Two Anchor Tests



Target Language Two - Spanish Language

A Preliminary Study of the Data

The initial step in this phase of the analysis was a simple comparison of the means and variances of the scores distributions of the test across forms and languages. The summary statistics for these scores are presented in Table 15. In Form A, the mean of SL (M = 108.87) was more than 6 points higher than TL (M = 102.50), and the standard deviation of SL (SD = 18.67) was close to TL (SD = 18.88) across languages. In Form B, the differences of mean and standard deviation were larger than in Form A. The mean of SL (M = 108.96) was more than 8 points higher than TL (M = 100.89), however, the standard deviation of TL (SD = 19.71) was more than 1 point higher than SL (SD = 18.48). Within Spanish language, mean of Form A (M = 102.50) scored about two points higher than Form B (M = 100.89), and standard deviation of Form A (SD = 18.88) was scored close to one point lower than Form B (SD = 19.71). In the English language, Form A (M = 108.87) was very close to Form B (M = 108.96) of their means, and their standard deviations were about the same as well (SD = 18.67, SD = 18.48, respectively).

Table 15

Spanish Language: The Statistics for Examinees Total Right Scores by Language and Test Forms

Statistics	English Language		Spanish Language	
	Form A	Form B	Form A	Form B
Mean	108.87	108.96	102.50	100.89
Standard Deviation	18.67	18.48	18.88	19.71
Count	1454	1441	62	114

A two-way ANOVA was also conducted at this preliminary stage with one factor being the language and the other being test form. The results are presented in Table 16. In this analysis, ANOVA was used to test the null hypotheses that there was no difference in the scores between languages or between forms. The interaction between the two factors (i.e., language groups and forms) was also of interest.

As can be seen from Table 16, both factors of forms, F (1, 3067) = 0.253, p = 0.615 and interaction, F (1, 3071) = 0.320, p = 0.572 were not statistically significant, however, the language, F (1, 3067) = 22.872, p = 0.000 was statistically significant. Since each factor had only two levels, we can infer that the mean values for forms and interactions of each level were not significantly different from each other. The mean scores of SL were higher than that of TL in both Form A and Form B. The relationship between the means is presented graphically in Figure 11. The effect size (eta squared) was also calculated to measure the magnitude of the difference between the language and

the form. Results showed that eta square was 0.007 for the language. This was a very small effect size, which indicated that the significant difference might be unimportant.

Table 16

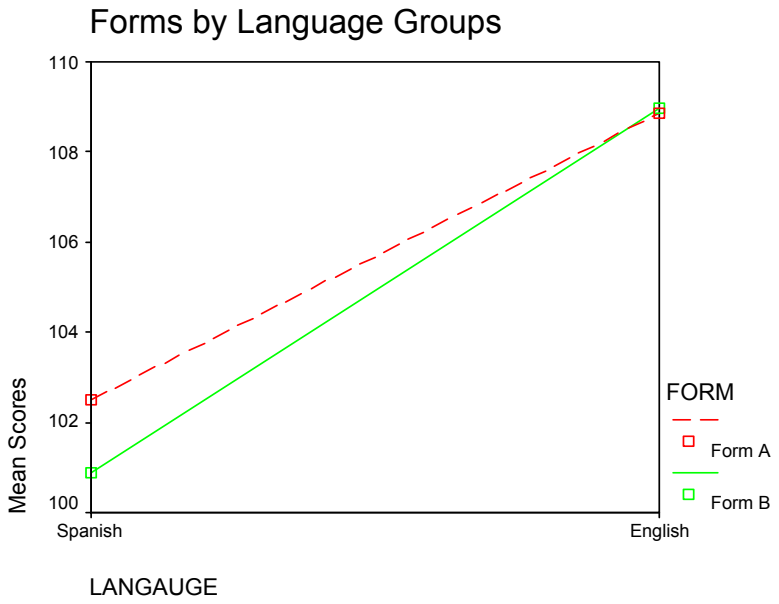
Spanish Language: ANOVA Results for Total Number Right Score

Source of Variation	Sum of Squares	DF	Mean Square	F	p	Effect Size
Language	7935.419	1	7935.419	22.872	0.000*	0.007
Form	87.845	1	87.845	0.253	0.615	0.000
Language * Form	111.073	1	111.073	0.320	0.572	0.000
Error	1064106.957	3067	346.954			
Corrected Total	1073449.162	3070				

* significant at the 0.05 level.

Figure 11

Spanish Language: Means Scores by Forms and Languages



Investigating Reliability and Validity

Results of Reliability Estimation of the Test Items

Internal consistency was examined for the languages and forms using a total of 160 items. The results are presented in Table 17. In Form A, the alphas were 0.9244 and 0.9201 for SL and TL, respectively for the verbal section of the test, and the alphas were 0.9141 and 0.9019 for SL and TL, respectively for the non-verbal section of the test. In Form B, the alphas were 0.9229 and 0.9260 for SL and TL, respectively for the verbal section of the test, and the alphas were 0.9237and 0.9162 for SL and TL, respectively for the non-verbal section of the test. All these coefficient alphas for the languages and forms are very reliable, which means that the test items are quite homogeneous.

Table 17

Spanish Language: Cronbach's Alpha by Language and Form

	English Language		Spanish Language	
	Verbal	Non-Verbal	Verbal	Non-Verbal
Form A	0.9244	0.9141	0.9201	0.9019
Form B	0.9229	0.9237	0.9260	0.9162

Results of Validity Estimation of the Test Items

Construct validity was investigated using principle component analysis (PCA) and parallel analysis (PA) to determine the number of factors to extract. PA is a method that generates random data from the same mean and the same standard deviation of the item responses. Eigenvalues for both actual data using PCA method and generated data were computed and compared. If the eigenvalues of the real data exceed the eigenvalues of the random data then the factor would be retained (Thompson & Daniel, 1996).

For Form A, PCA was performed on all 156 items and extracted two components with eigenvalues of 14.14 and 3.92, respectively, and these two components accounted for 11.58% of the total variance. The eigenvalue of the third component that PCA produced was 2.27. The maximum eigenvalue that the PA produced was 4.02, and the next two eigenvalues were 3.89 and 2.11; therefore, two factors were retained. The Scree Plot can be found in Figure 23, Appendix B. The rotated component matrix using the Varimax method showed that the majority of items loaded on the first component and the second component, with 81.41% (127/156) of the items having a correlation coefficient

larger than 0.10 and 31.41% (49/156) a correlation coefficient larger than 0.30. These two components represented the verbal section and non-verbal section of the test, respectively.

For Form B, PCA was performed on all 156 items and extracted two components with eigenvalues of 14.04 and 3.77, respectively, and these two components accounted for 11.42% of the total variance. The eigenvalue of the third component that PCA produced was 1.93. The maximum eigenvalue that the PA produced was 4.17, and the next two eigenvalues were 3.72 and 1.89; therefore, two factors were retained. The Scree Plot can be found in Figure 24, Appendix B. The rotated component matrix using the Varimax method showed that the majority of items loaded on the first component and the second component, with 82.05% (128/156) of the items having a correlation coefficient larger than 0.10 and 30.77% (48/156) a correlation coefficient larger than 0.30, and these two components represented the verbal section and non-verbal section of the test, respectively.

Choosing Anchor Items

Anchor Test One - Results of Combination of (1) (2) and (4)

Anchor items should be a miniature version of the total test (Pansy & Kromrey, 1993). In this section, item difficulty indices and item discrimination indices were analyzed in selection of proper anchor items. Table 33 in Appendix C reports the item difficulty indices and item discrimination indices. First of all, the items that discriminated well and had mild and similar difficulty indices were chosen as anchor items. Then, all of these anchor items should be a miniature of the total test and these items represented 20% of the total items. In addition, anchor items should be administered in the same order for

both language groups. For Form A, there were in total 37 items (24 items for V_1 and 13 items for NV_1) chosen; for Form B, a total of 37 items (24 items for V_2 and 13 items for NV_2) were selected. Here V_1 and V_2 stood for the verbal section of the test for Form A and Form B, respectively; NV_1 and NV_2 were for the non-verbal section of the test for Form A and Form B, respectively. A summary of the number of anchor items for anchor test one is presented below in Table 18. See Table 36 in Appendix E for a summary of how items are allocated in each content specification for anchor test one.

Anchor Test Two - Results of Combination of (1) (2) and (5)

Delta plot is a method based on item difficulty values converted to a normal deviate with an arbitrary mean and standard deviation. Items that were closest to the equal difficulty line and also within the limit sets for the difficulty indices were used for anchor items. Additionally, these anchor items must be a miniature of the total test and about 20% of the total items. Figure 41 though Figure 54 in Appendix D are the diagrams of the delta plot for all 160 items for Form A and Form B. Since it was difficult to identify each item when they were clustered together, the actual anchor items were visually selected from the delta plots of the items categorized by each content specification of the test. There were in total of 37 items (24 items for V_1 and 13 items for NV_1) chosen as anchor items for Form A, and 37 items (24 items for V_2 and 13 items for NV_2) selected for Form B. A summary of the number of anchor items for anchor test two is presented in Table 18. See Table 36 in Appendix E for a summary of how items are allocated in each content specification for anchor test two.

Table 18

Spanish Language: A Summary of the Number of Anchor Items for Two Anchor Tests

Form / Anchor Test	Anchor One	Anchor Two
Form A	V_1: 24	V_1: 24
	NV_2: 13	NV_2: 13
Form B	V_1: 24	V_1: 24
	NV_2: 13	NV_2: 13

Results of Levine Linear Equating

Four Levine linear equating results for each anchor test were calculated: two for the verbal section and two for the non-verbal section of the test. In either the verbal or the non-verbal sections for each anchor test, an equating for the first link and an equating for the second link were conducted. Both of these two anchor tests used a common-item nonequivalent design. All detailed Levine Linear equating results are presented in Table 42 and Table 43, Appendix G for the verbal section and the non-verbal section of the test.

Table 19 provides a summary of all the statistics for the slopes and the intercepts pertaining to each anchor test. For the verbal section of anchor test one, the slope and the intercept were 1.13 and –8.96, respectively for the first link; and 1.10 and -6.22 for the second link. For the non-verbal section of anchor test one, the slope and intercept were 0.99 and 2.19, respectively for the first link; and 1.02 and 0.40 for the second link. The slopes of 1.05 and 1.02 and the intercepts of -2.91 and -1.05 for the first link and the second link, respectively were reported for the verbal section of anchor test two. In the

non-verbal section of anchor test two, the slope of 1.15 and the intercept of -4.40 were for the first link; and the slope of 0.92 and the intercept of 2.63 were for the second link.

Table 19

Spanish Language: A Summary of the Slopes and the Intercepts for Levine Equating of Anchor Item Design

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
Slope	1.13	1.10	0.99	1.02	1.05	1.02	1.15	0.92
Intercept	-8.96	-6.22	2.19	0.40	-2.91	-1.05	-4.40	2.63

Results of Mean-Sigma Equating Method

Four Mean-Sigma equating results for each anchor test were reported: two for the verbal section and two for the non-verbal section. In either the verbal or the non-verbal section for each anchor test, two equatings were for the first link and for the second link. Note that the equivalent groups design link for the first chain in the same for both anchor tests construction methods. The same is true for the second chain. The mean-sigma equating results can be found in Appendix G.

For the verbal section of both anchor test one and anchor test two, the slope and the intercept were the same: 0.98 and 1.25, respectively for the first link; and 1.04 and -3.80 for the second link. For the non-verbal section of two anchor tests, the slope and the

intercept were 1.00 and 0.10, respectively for the first link; and 1.05 and -2.25 for the second link. Table 20 provides a summary of all these statistics.

Table 20

Spanish Language: A Summary of the Slopes and the Intercepts for Mean-Sigma Equating of Equivalent Group Design

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
Slope	0.98	1.04	1.00	1.05	0.98	1.04	1.00	1.05
Intercept	1.25	-3.80	0.10	-2.25	1.25	-3.80	0.10	-2.25

Results of Double Linking Equating Evaluation Method

Double linking provides a built-in check on the equating process and leads to greater equating stability (Kolen & Brennan, 1995). In this method, both "within language" and "across language" links were examined for each anchor test. The "within language" was executed using the equivalent group design and the "across language" was examined using the anchor item design. Table 21 provides a summary of all the statistics for the slopes and the intercepts pertaining to each anchor test. In order to calculate the slopes and intercepts of the first chain or the second chain for each anchor test, we need to compose the two first links or the two second links from Table 19 and Table 20.

For the verbal section of anchor test one, the slope and intercept were 1.11 and -7.53, respectively for the first chain; and 1.14 and -10.40 for the second chain. For the non-verbal section of anchor test one, the slope and intercept were 0.99 and 2.29 for the first chain, respectively; and 1.07 and -1.90 for the second chain. The slopes of 1.03 and 1.06 and intercepts of -1.60 and -4.93 for the first chain and the second chain, respectively were reported for the verbal section of anchor test two. In the non-verbal section of anchor test two, slope of 1.15 and intercept of -4.30 were for the first chain; and slope of 0.97 and intercept of 0.56 were for the second chain.

Table 21

Spanish Language: A Summary of the Slopes and the Intercepts for Double Linking Equating of Two Anchor Tests

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Chain	Second Chain	First Chain	Second Chain	First Chain	Second Chain	First Chain	Second Chain
Slope	1.11	1.14	0.99	1.07	1.03	1.06	1.15	0.97
Intercept	-7.53	-10.40	2.29	-1.90	-1.60	-4.93	-4.30	0.56

Figure 12 illustrates the difference between the two different functions of the two anchor tests for the verbal section of the test. For the verbal section of anchor test one, the difference between the two chains was from 0 to 2.5 raw score points; for the verbal section of anchor test two, the difference between the two chains was from 0.25 to 2.75

raw score points. The absolute mean difference for anchor test one and anchor test two in the verbal section were about 1.30 and 1.73 raw score points, respectively. As can been seen from Figure 12, anchor test one was closer to the zero perfect line than anchor test two. See Appendix H and Appendix I for detailed double linking statistics. Also, Figure 77 through Figure 80 in Appendix J present more detailed diagrams of the double linking results of the equating function differences between the two chains for the verbal section of two anchor tests.

Figure 12

Spanish Language: Different Functions of Two Anchor Tests for Verbal Sections

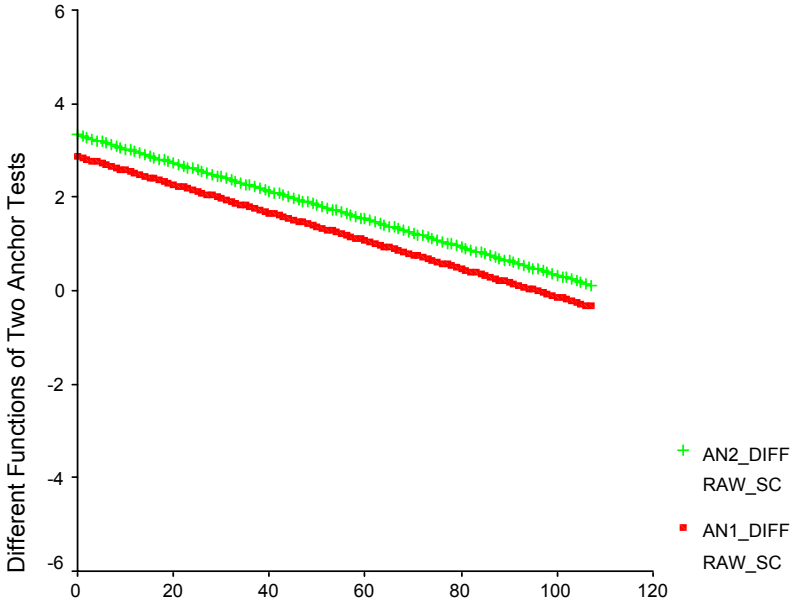
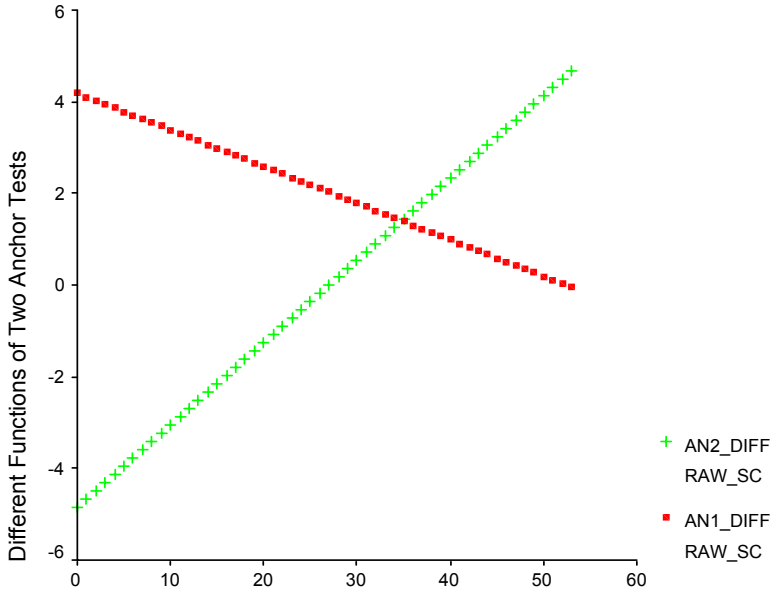


Figure 13 presents the difference between the two different functions of the two anchor tests for the non-verbal section of the test. For the non-verbal section of anchor test two, the difference between two chains was from -5 to 5 raw score points; for the

non-verbal section of anchor test one, the difference between the two chains was from 0 to 4 raw score points. The absolute mean difference for anchor test one and anchor test two in the non-verbal section were about 2.07 and 2.43 raw score points, respectively. As can be seen from Figure 13, anchor test one was closer to the zero perfect line than anchor test two. See Appendix H and Appendix I for detailed double linking statistics. Also, Figure 81 through Figure 84 in Appendix J present more detailed diagrams of the double linking results of the equating function difference between the two chains for the non-verbal section of two anchor tests.

Figure 13

Spanish Language: Different Functions of Two Anchor Tests for Non-Verbal Sections



Results of Standard Error of Equating Evaluation Method

The values of the mean standard errors of equating are shown in Table 22. For the first chain in the verbal section, the SEE for anchor test two (3.87) was higher than for anchor test one (3.45), and in the same order for the second chain (6.21, 5.60, respectively). If we averaged these two chains, anchor test two (5.04) scored higher the SEE than anchor test one (4.53). The MSEE for anchor test two (0.0910) was similar to anchor test one (0.0818). The diagram of SEE for the verbal section is presented in Figure 14.

For the first chain in the non-verbal section, anchor test two (1.97) had a slightly higher SEE than anchor test one (1.88), and in the same order for the second chain (3.14, 2.86, respectively) as well. If we averaged these two chains, anchor test two (2.56) scored higher the SEE than anchor test one (2.37). The MSEE for anchor test two (0.0462) was similar to anchor test one (0.0428) as well. The diagram of SEE for the non-verbal is presented in Figure 15. Detailed SEE results can be found in Table 56 and Table 57, Appendix L.

Table 22

Spanish Language: MSEE between Two Anchor Tests

	Anchor Test One		Anchor Test Two	
	Verbal	Non-verbal	Verbal	Non-verbal
First Chain	3.45	1.88	3.87	1.97
Second Chain	5.60	2.86	6.21	3.14
Average	4.53	2.37	5.04	2.56
MSEE	0.0818	0.0428	0.0910	0.0462

Figure 14

Spanish Language: SEE of the Verbal Section for Two Anchor Tests

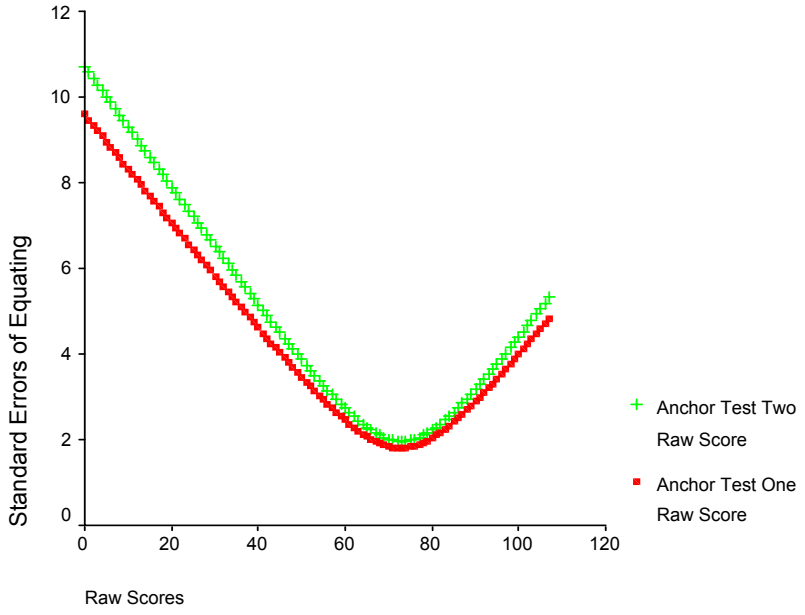
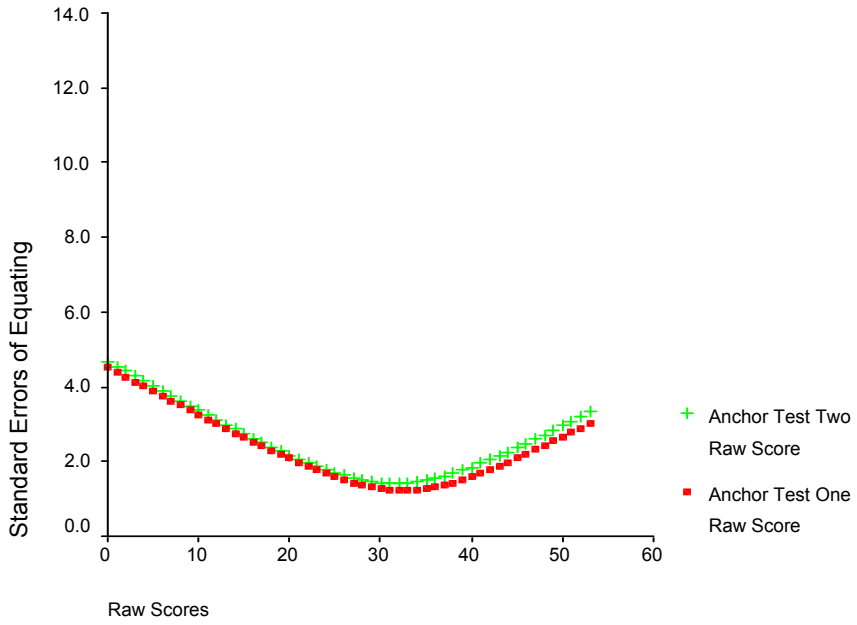


Figure 15

Spanish Language: SEE of the Non-Verbal Section for Two Anchor Tests



Target Language Three - Chinese Language

A Preliminary Study of the Data

The initial step in this phase of the analysis was a simple comparison of the means and variances of the scores distributions of the test across forms and languages. The summary statistics for these scores are presented in Table 23. In Form A, the mean of SL (M = 105.09) was about 3 points higher than TL (M = 102.38), however, the standard deviation of SL (SD = 16.31) was more than 3 points lower than TL (SD = 19.50). In Form B, the difference of mean was smaller than Form A, and in the reverse order for standard deviation. The mean of TL (M = 101.98) was 2 points higher than SL (M = 99.93), and the standard deviation of TL (SD = 17.21) was 5 points lower than SL (SD = 22.92). Within English language, mean of Form A (M = 105.09) scored five points higher than Form B (M = 99.93), however, the standard deviation of Form A (SD = 16.31) was scored six points lower than Form B (SD = 22.92). In TL, the means were about the same for Form A and Form B (M = 102.38, M = 101.98, respectively), and Form A (SD = 19.50) surpassed Form B (SD = 17.21) two points of their standard deviations.

Table 23

Chinese Language: The Statistics for Examinees Total Right Scores by Language and Test Forms

Statistics	English Language		Chinese Language	
	Form A	Form B	Form A	Form B
Mean	105.09	99.93	102.38	101.98
Standard Deviation	16.31	22.92	19.50	17.21
Count	1677	1463	128	116

A two-way ANOVA was also conducted at this preliminary stage with one factor being languages and the other being test form. The results are presented in Table 24. In this analysis, ANOVA was utilized to test the null hypotheses that there was no difference in the scores between languages or between forms. The interaction between the two factors (i.e., language groups and forms) was also of interest.

As can be seen from Table 24, both factors of language groups, F (1, 3370) = 0.064, p = 0.800 and interaction, F (1, 3370) = 3.340, p = 0.068 were not statistically significant, however, the form, F (1, 3370) = 4.531, p = 0.033 was statistically significant. Since each factor had only two levels, we can infer that the mean values for each level were not significantly different from each other for language groups and the interaction. The mean scores of both SL and TL in Form A were higher than that of SL and TL in Form B. The relationship between the means is presented graphically in Figure 16. The Effect size (eta squared) was also calculated to measure the magnitude of the

difference between the language and the form. Results showed that eta square was 0.001 for the language. This was a very small effect size, which indicated that the significant difference might be unimportant.

Table 24

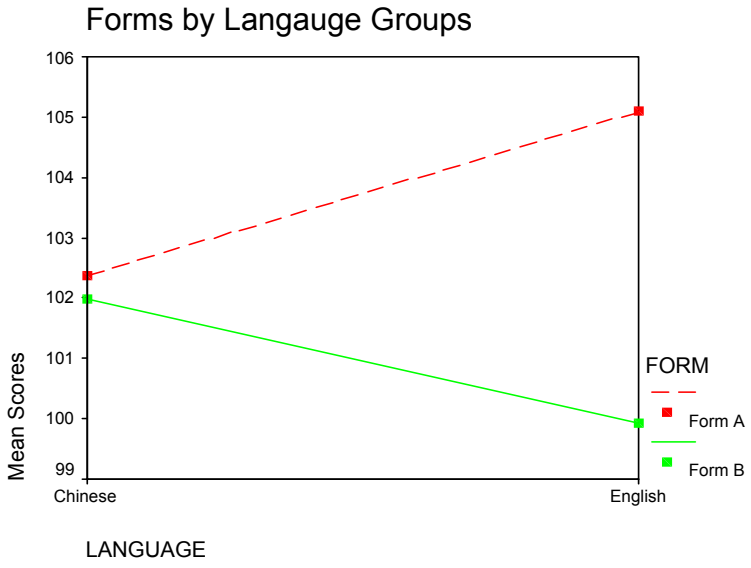
Chinese Language: ANOVA Results for Total Number Right Score

Source of Variation	Sum of Squares	DF	Mean Square	F	p	Effect Size
Language	24.684	1	24.684	0.064	0.800	0.000
Form	1739.09	1	1739.09	4.531	0.033*	0.001
Language * Form	1282.258	1	1282.258	3.340	0.068	0.001
Error	1293605.714	3370	383.859			
Corrected Total	1314403.213	3373				

* significant at the 0.05 level.

Figure 16

Chinese Language: Means Scores by Forms and Language Groups



Investigating Reliability and Validity of the Items

Results of Reliability Estimation of the Test Items

Internal consistency was examined for the languages and forms using a total of 160 items. The results are presented in Table 25. In Form A, the alphas were 0.8955 and 0.9247 for SL and TL, respectively for the verbal section of the test, and the alphas were 0.9037 and 0.9184 for SL and TL, respectively for the non-verbal section of the test. In Form B, the alphas were 0.9455 and 0.9021 for SL and TL, respectively for the verbal section of the test, and the alphas were 0.9256 and 0.9039 for SL and TL, respectively for the non-verbal section of the test. All these coefficient alphas for the languages and forms are very reliable, which means that the test items are quite homogeneous.

Table 25

Chinese Language: Cronbach's Alpha by Language and Form

	English Language		Chinese Language	
	Verbal	Non-Verbal	Verbal	Non-Verbal
Form A	0.8955	0.9037	0.9247	0.9184
Form B	0.9455	0.9256	0.9021	0.9039

Results of Validity Estimation of the Test Items

Construct validity was investigated using principle component analysis (PCA) and parallel analysis (PA) to determine the number of factors to extract. PA is a method that generates random data from the same mean and the same standard deviation of the item responses. Eigenvalues for both actual data using PCA method and generated data were computed and compared. If the eigenvalues of the real data exceed the eigenvalues of the random data then the factor would be retained (Thompson & Daniel, 1996).

For Form A, PCA was performed on all 148 items and extracted two components with eigenvalues of 11.05 and 3.13, respectively, and these two components accounted for 9.58% of the total variance. The eigenvalue of the third component that PCA produced was 2.04. The maximum eigenvalue that the PA produced was 4.23, and the next two eigenvalues were 3.08 and 1.92; therefore, two factors were retained. The Scree Plot can be found in Figure 25, Appendix B. The rotated component matrix using the Varimax method showed the majority of items loaded on the first component and the second component, with 76.35% (113/148) of the items having a correlation coefficient

larger than 0.10 and 20.95% (31/148) a correlation coefficient larger than 0.30. These two components represented the verbal section and non-verbal section of the test, respectively.

For Form B, PCA was performed on all 158 items and extracted two components with eigenvalues of 22.17 and 4.97, respectively, and these two components accounted for 17.18% of the total variance. The eigenvalue of the third component that PCA produced was 2.63. The maximum eigenvalue that the PA produced was 5.75, and the next two eigenvalues were 4.93 and 1.98; therefore, two factors were retained. The Scree Plot can be found in Figure 26, Appendix B. The rotated component matrix using the Varimax method showed that the majority of items loaded on the first component and the second component, with 89.24% (141/158) of the items having a correlation coefficient larger than 0.10 and 23.32% (40/158) a correlation coefficient larger than 0.30, and these two components represented the verbal section and non-verbal section of the test, respectively.

Choosing Anchor Items

Anchor Test One - Results of Combination of (1) (2) and (4)

Anchor items should be a miniature version of the total test (Pansy & Kromrey, 1993). In this section, item difficulty indices and item discrimination indices were analyzed in selection of proper anchor items. Table 34 in Appendix C presented the item difficulty indices and item discrimination indices. First, the items that discriminated well and had mild and similar difficulty indices were chosen as anchor items. Then, all these anchor items must be a miniature of the total test. Since 20% of the total items were larger than 20 items, 20% of the total items criterion was utilized. In addition, anchor

items should be administered in the same order for both language groups. For Form A, there were in total 38 items (24 items for V_1 and 14 items for NV_1) chosen for the anchor test; for Form B, a total of 38 items (24 items for V_2 and 14 items for NV_2) were selected. Here V_1 and V_2 stood for the verbal section of the test for Form A and Form B, respectively; NV_1 and NV_2 were the non-verbal of the test for Form A and Form B, respectively. A summary of the number of anchor items for anchor test one is presented below in Table 26. See Table 37 in Appendix E for a summary of how items are allocated in each content specification for anchor test one.

Anchor Test Two - Results of Combination of (1) (2) and (5)

Delta plot method based on item difficulty values and converted to a normal deviate with an arbitrary mean and standard deviation. For this method, items that were closest to the equal difficulty line and that also within the limits set for the difficulty indices were used for anchor items. In addition, these anchor items must be a miniature of the total test and about 20% of the total items. Figure 55 through Figure 68 in Appendix D are the delta plots for all 160 items for Form A and Form B, respectively. Since it was difficult to identify each item when they were clustered together, the actual anchor items were visually selected from the delta plots of the items categorized by each content specification of the test. There were in total of 38 items (24 items for V_1 and 14 items for NV_1) chosen as anchor items for Form A, 38 items (24 items for V_2 and 14 items for NV_2) selected for Form B. A summary of the number of anchor items for anchor test two is reported in Table 26. See Table 37 in Appendix E for a summary of how items are allocated in each content specification for anchor test two.

Table 26

Chinese Language: A Summary of the Number of Anchor Items for Two Anchor Tests

Form / Anchor Test	Anchor One	Anchor Two
Form A	V_1: 24	V_1: 24
	NV_2: 13	NV_2: 13
Form B	V_1: 24	V_1: 24
	NV_2: 13	NV_2: 13

Results of Levine Linear Equating

There were a total of four Levine linear equating results for each anchor test: two for the verbal section and two for the non-verbal section of the test. In either the verbal or the non-verbal section for each anchor test, there was an equating for the first link and an equating for the second link. Both of these two anchor tests used common-item nonequivalent design. All detailed Levine Linear equating results are presented in Table 44 and Table 45, Appendix G for the verbal section and the non-verbal section of the test.

Table 27 provides a summary of all the statistics for the slopes and the intercepts pertaining to each anchor test. For the verbal section of anchor test one, the slopes and the intercepts were 1.03 and -2.46, respectively for the first link; and 1.07 and -6.15 for the second link. For the non-verbal section of anchor test one, the slopes and intercepts were 0.98 and 0.65, respectively for the first link; and 1.34 and -14.41 for the second link. The slopes of 1.18 and 0.89 and the intercepts of -11.51 and 6.81 for the first link and the second link, respectively were reported for the verbal section of anchor test two. In the

non-verbal section of anchor test two, the slope of 1.01 and the intercept of -0.27 were for the first link; and the slope of 1.11 and the intercept of -4.25 were for the second link.

Table 27

Chinese Language: A Summary of the Slopes and the Intercepts for Levine Equating Method of Anchor Item Design

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
Slope	1.03	1.07	0.98	1.34	1.18	0.89	1.01	1.11
Intercept	-2.46	-6.15	0.65	-14.41	-11.51	6.81	-0.27	-4.25

Results of Mean-Sigma Equating Method

Four Mean-Sigma equating results for each anchor test were reported: two for the verbal section and two for the non-verbal section. In either the verbal or the non-verbal section for each anchor test, there was an equating for the first link and an equating for the second link. Note that the equivalent groups design link for the first chain in the same for both anchor tests construction methods. The same is true for the second chain. Detailed Mean-Sigma equating results can be found in Appendix G.

For the verbal section of both anchor test one and anchor test two, the slope and the intercept were the same: 1.46 and -34.06, respectively for the first link, and 1.06 and -3.34 for the second link. For the non-verbal section of two anchor tests, the slope and the

intercept were 1.11 and 6.50, respectively for the first link; and 0.66 and 19.93 for the second link. Table 28 provides a summary of all these statistics.

Table 28

Chinese Language: A Summary of the Slopes and the Intercepts for Mean-Sigma Equating Method of Equivalent Group Design

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
Slope	1.46	1.06	1.11	0.66	1.46	1.06	1.11	0.66
Intercept	-34.06	-3.34	6.50	19.93	-34.06	-3.34	6.50	19.93

Results of Double Linking Equating Evaluation Method

Double linking provides a built-in check on the equating process and leads to greater equating stability (Kolen & Brennan, 1995). In this method, both "within language" and "across language" links were examined for each anchor test. The "within language" was executed using the equivalent group design and the "across language" was examined using the anchor item design. Table 29 provides a summary of all the statistics for the slopes and the intercepts pertaining to each anchor test. In order to calculate the slopes and intercepts of the first chain or the second chain for each anchor test, we need to compose the two first links or the two second links from Table 27 and Table 28.

For the verbal section of anchor test one, the slope and intercept were 1.50 and -37.65, respectively for the first chain; and 1.13 and -9.72 for the second chain. For the non-verbal section of anchor test one, the slope and intercept were 1.09 and 7.22 for the first chain, respectively; and 0.88 and 12.30 for the second chain. The slopes of 1.72and 0.94 and intercepts of -50.86 and 3.84 for the first chain and the second chain, respectively were reported for the verbal section of anchor test two. In the non-verbal section of anchor test two, slope of 1.12 and intercept of 6.20 were for the first chain; and slope of 0.73 and intercept of 17.87 were for the second chain.

Table 29

Chinese Language: A Summary of the Slopes and the Intercepts for Double Linking Equating of Two Anchor Tests

	Anchor Test One				Anchor Test Two			
	Verbal		Non-Verbal		Verbal		Non-Verbal	
	First Chain	Second Chain	First Chain	Second Chain	First Chain	Second Chain	First Chain	Second Chain
Slope	1.50	1.13	1.09	0.88	1.72	0.94	1.12	0.73
Intercept	-37.65	-9.72	7.22	12.30	-50.86	3.84	6.20	17.87

Figure 17 illustrates the difference between the two different functions of the two anchor tests for the verbal section of the test. For the verbal section of anchor test two, the difference between the two chains was from -57 to 30 raw score points; for the verbal section of anchor test one, the difference between the two chains was from -33 to 10 raw

score points. The absolute mean difference for anchor test one and anchor test two in the verbal section were about 11.65 and 23.06 raw score points, respectively. As can be seen from Figure 17, anchor test one was closer to the zero perfect line than anchor test two. See Appendix H and Appendix I for more detailed double linking statistics. Figure 85 through Figure 88 in Appendix J present more detailed diagrams of the double linking results of the equating function differences between the two chains for the verbal section of two anchor tests.

Figure 17

Chinese Language: Different Functions of Two Anchor Tests for Verbal Sections

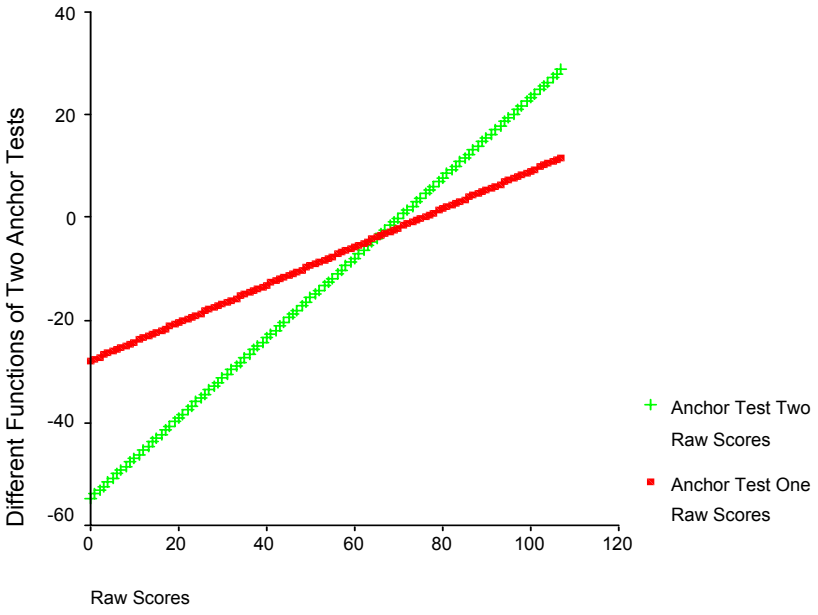
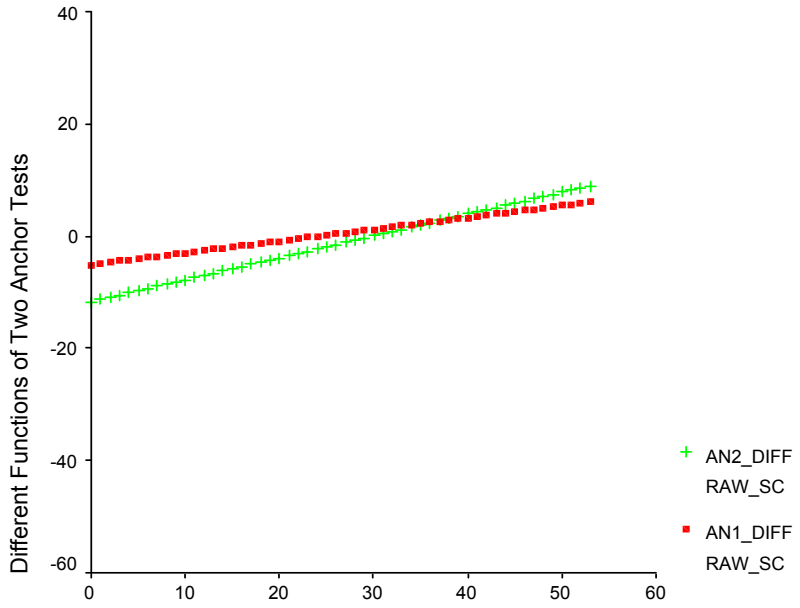


Figure 18 presents the difference between the two different functions of the two anchor tests for the non-verbal section of the test. For the non-verbal section of anchor test two, the difference between the two chains was from -10 to 4 raw score points; for

the non-verbal section of anchor test one, the difference between two chains was from -5 to 3 raw score points. The absolute mean difference for anchor test one and anchor test two in the non-verbal section were about 2.86 and 5.35 raw score points, respectively. As can be seen from Figure 18, anchor test one was closer to the zero perfect line than anchor test two. See Appendix H and Appendix I for more detailed double linking statistics. Also, Figure 89 through Figure 92 in Appendix J present more detailed diagrams of the double linking results of the equating function differences between the two chains for the non-verbal section of two anchor tests.

Figure 18

Chinese Language: Different Functions of Two Anchor Tests for Non-Verbal Sections



Results of Standard Error of Equating Evaluation Method

The values of the mean standard errors of equating are shown in Table 30. For the first chain in the verbal section, anchor test two (4.04) had higher SEE than anchor test one (2.51), and in the same order for the second chain (5.38, 5.27, respectively). If we averaged these two chains, anchor test two (4.71) scored higher SEE than anchor test one (3.89). The MSEE for anchor test two (0.0855) was similar to anchor test one (0.0670). The diagram of SEE for the verbal is presented in Figure 19.

For the first chain in the non-verbal section, anchor test two (1.34) had a slightly higher SEE than anchor test one (1.32), however, it was not in the same order for the second chain (2.73, 2.90, respectively). If we averaged these two chains, anchor test one (2.11) had larger SEE than anchor test two (2.04). The MSEE for anchor test one (0.0360) was similar to anchor test two (0.0351) as well. The diagram of SEE for the non-verbal is presented in Figure 20. Detailed SEE results can be found in Table 58 and Table 59, Appendix K.

Table 30

Chinese Language: MSEE between Two Anchor Tests

	Anchor Test One		Anchor Test Two	
	Verbal	Non-verbal	Verbal	Non-verbal
First Chain	2.51	1.32	4.04	1.34
Second Chain	5.27	2.90	5.38	2.73
Average	3.89	2.11	4.71	2.04
MSEE	0.0670	0.0363	0.0855	0.0351

Figure 19

Chinese Language: SEE of the Verbal Section for Two Anchor Tests

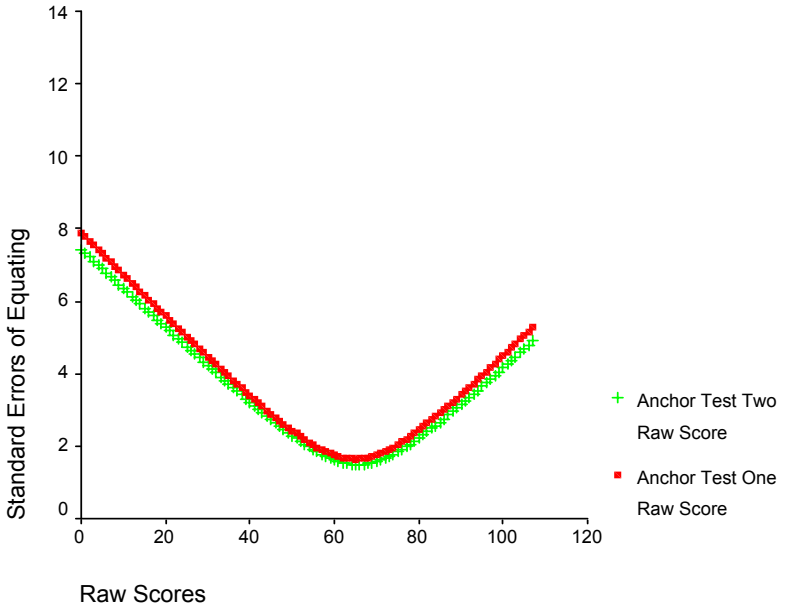
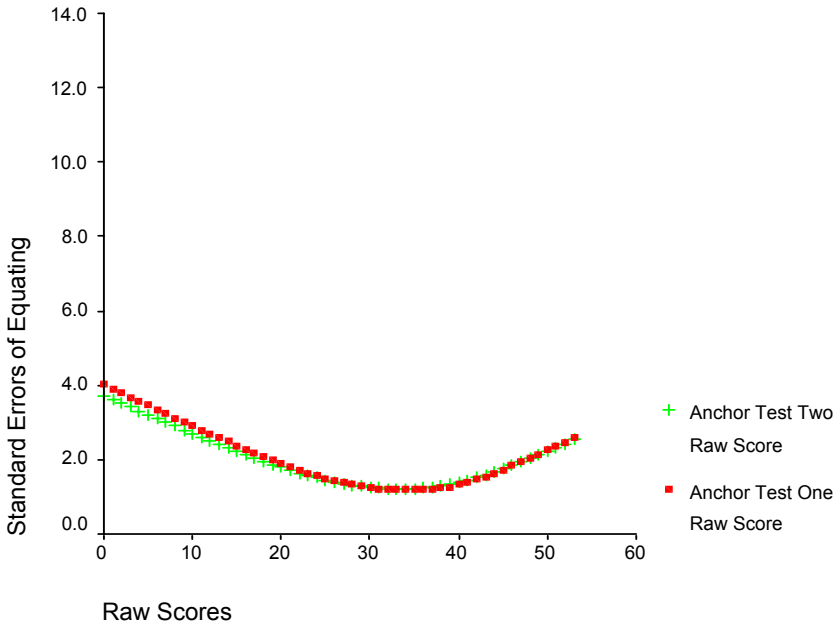


Figure 20

Chinese Language: SEE of the Non-Verbal Section for Two Anchor Tests



CHAPTER FIVE

Review of Results

Recently, adapting professional certification examinations from one language to another has increased. Equating with an anchor item design is an often utilized statistical procedure for adapted certification exams that adjusts test scores on different versions of the same exam so that the scores can be interpreted interchangeably. Because the set of anchor items is used in the adapted exams, the items that eventually chosen as anchor items are very important for the accuracy of equating. The researcher of this study argues that the accuracy of equating does not only rely on number of anchor items or location of anchor items or content representative anchor items. The anchor items should be a combination of both practical considerations and statistical considerations. However, this combination cannot be accomplished in one test. Therefore, the primary research questions addressed in this study concern how different approaches of selecting anchor items (hereafter referred to as anchor tests) on the accuracy of equating for test adaptation. Moreover, identifying the more effective anchor test in terms of evaluating the accuracy of equating is the major concern of this study.

In this study, the double linking method and the mean standard errors of equating method were used to evaluate the accuracy of equating for different anchor tests. In total three sets of data were analyzed for this research study. For each set of the data, the verbal section of the test and the non-verbal section of the test were investigated. Scoring outcomes of an actual certification examination with a sample of nearly 9,000 examinees taking both SL and TL versions of the test data set were utilized for this research study.

The Levine Linear equating method and Mean-Sigma equating method were employed for anchor item design and equivalent group design, respectively. The within language equating link in each of the anchor tests was assumed to be fairly stable. Therefore, the differences found between the equating results in the two chains mainly reflected the instability of the cross-lingual equating links for different anchor tests.

For the verbal section of the test, the results of the double linking method presented a very interesting case. This method supported the notion that different choices for anchor items can result in different equatings and anchor test one was a better choice among all three language groups for the cross-lingual equating. In the Korean language, the absolute mean difference for anchor test one and anchor test two were about 4.28 and 12.52 raw score points, respectively. The absolute mean difference for the Spanish language were 1.30 and 1.73 raw score points for anchor test one and anchor test two respectively, which were the smallest mean absolute differences among all three language groups. In the Chinese language, the mean absolute difference were 11.65 and 23.06 raw score points for anchor test one and anchor test two, respectively. A summary of these statistics for anchor test two are presented in Table 31.

For the non-verbal section of the test, the results of the double linking method presented a very interesting case as well. This method also supported the notion that different choices for anchor items can result in different equatings and anchor test one was a better choice among all three language groups for the cross-lingual equating. In the Korean language, the absolute mean difference for anchor test one and anchor test two were about 1.35 and 2.00 raw score points, respectively. The absolute mean difference for the Spanish language were 2.07 and 2.43 raw score points for anchor test one and

anchor test two, respectively. In the Chinese language, the absolute mean difference were 2.86 and 5.35 raw score points for anchor test one and anchor test two, respectively. A summary of these statistics for anchor test two are presented in Table 31.

Table 31

All Language Groups: A Summary of the Absolute Mean Differences for Two Anchor Tests

		Korean Language	Spanish Language	Chinese Language
Verbal	Anchor Test One	4.28	1.30	11.65
	Anchor Test Two	12.52	1.73	23.06
Non-Verbal	Anchor Test One	1.35	2.07	2.86
	Anchor Test Two	2.00	2.43	5.35

The MSEE statistics for the verbal section of all three language groups were in the same direction as the double linking method. Overall, anchor test two demonstrated a larger MSEE than anchor test one. For the Korean language, the MSEE for anchor test one and anchor test two were 0.108 and 0.118, respectively, and the difference was 0.010. The MSEE were 0.082 for anchor test one and 0.091 for anchor test two for the Spanish languages, which was 0.009 difference between two anchor tests. For the Chinese language, the MSEE for anchor test one was 0.067 and was 0.086 for anchor test two, and the difference between these two was 0.019. However, the MSEE statistics for all language groups did not show large differences between the two anchor tests.

The MSEE statistics for the non-verbal section of all three language groups also were in the same direction of the double linking method. Overall, anchor test two demonstrated a larger MSEE than anchor test one. For the Korean language, the MSEE for anchor test one and anchor test two were 0.047 and 0.050, respectively, and the difference was 0.003. The MSEE were 0.043 for anchor test one and 0.046 for anchor test two for the Spanish languages, which was 0.003 difference between two anchor tests. For the Chinese language, the MSEE for anchor test one was 0.036 and 0.035 for anchor test two, and the difference between these two was 0.001. However, the MSEE statistics for all language groups did not show large differences between the two anchor tests.

Conclusions

The importance of this research study lies in finding the impact of different anchor tests on the accuracy of the cross-lingual equating process by comparing the verbal section and the non-verbal section of the test using different anchor tests. The findings indicated that using double linking method as an evaluation tool produces a convergence of equating results across the verbal section and non-verbal section of the test for all three sets of the data. The key in using this method is that if the equating process is free from error, the equating relationship resulting from the two links could be expected to be similar.

The findings from the study are very encouraging. The most important finding of this study is that for each set of the data, anchor test one and anchor test two did not produce the same results. In fact, the results of double linking showed that anchor test one was a better choice of selecting anchor items than anchor test two. Further, based on

the MSEE, anchor test one had smaller mean standard errors of equating than anchor test two.

The findings confirmed the Rapp and Allalouf (2002) study in that utilizing double linking method is a useful tool for evaluating the cross-lingual process. This study is an extension of their study and found that the instability of cross-lingual procedure is mitigated by using most stable parameter method to choose anchor items and could be very useful tools for cross-lingual studies.

The findings also indicated that the differences between the conversion functions in the two alternative links were significant for both the verbal section and the non-verbal section of the test for all three data sets. Obviously, the differences are caused by a real and systematic problem that underlines the cross-lingual equating process. This problem was greater in the verbal section of the test than the nonverbal section of the test. Of all the verbal sections and the non-verbal sections of the test, the verbal sections of the test showed more instability than the non-verbal sections of the test. Out of the three language groups, only the Spanish language showed reverse tendency in that the non-verbal section of the test in fact presented more instability than the verbal section of the test.

What sort of problems could create such differences for alternative anchor tests across language groups of the same test? The following four factors could interfere in the cross-lingual equating comparisons and introduce differences: (1) different anchor tests; (2) ability difference; (3) test adaptation process; (4) accurate equating.

The implication of this dissertation study finding is that we can not assume that the different approaches of selecting anchor items are the same, especially for the cross-lingual studies because the anchor items are chosen from the adapted items. Aside from

the inevitable and uncontrollable problem of content distortion that stems from the translation process, the other dimension that can be controlled to a certain extent is the manner of choosing proper anchor items. This is very critical when we attempt to equate groups that are from different language groups. Items eventually chosen as anchor items might discriminate well, be appropriately difficulty, adapt well from one language to another and show no DIF. Moreover, these items need to be content representative of the total test, and 20 items or 20% of the total items, whichever is larger. The data examined would indicate that choosing items with appropriate parameters is more important than choosing items that show less differential functioning. Furthermore, the results showed that choosing anchor items with one criterion may be quite different from those choosing anchor items using another criterion. See Appendix E for reference on the impact of different anchor tests.

It is likely that part of the differences between two anchor tests found in this dissertation study are caused by the problems of instability in the group performances on the test. Items that perform satisfactorily in one language may multifunction in another, and vice versa. Consequently, for each language we would expect to end up with a different test, composed of different items and with different characteristics.

All results pointed to the conclusion that we cannot simply assume that an adapted version of an exam is psychometrically equivalent to the original version of the test for different anchor tests. In this study, the overall exam instability across languages was large across the different versions of the exam. The very act of test adaptation may change the items in some fundamental way. Further, different items maybe affected in different ways. One cannot reliably predict the characteristics and behavior of an adapted

item without knowing the characteristics and behavior of the original item. The literature identified many differences between cultures and languages that could affect items that are adapted. Different factors will affect the adapted items in different ways. Therefore, a review of the exam adaptation procedures currently employed by the test developer and investigations to identify specific causes of these differences (for an example of how this might be done, see Allalouf, Hambleton, & Sireci, 1999) is very much in order. If these steps are taken, improvements in the equivalence of different anchor tests for the original and adapted exams can be obtained and verified using statistical analyses. As the exam stands now in three language versions, a serious error would result if these language versions of the exam were considered to be equivalent and scores used interchangeably. As indicated in the introduction, adapting an exam into multiple languages is a complex process, and many aspects of an exam, including the content specifications, the exam directions, and the exam administration conditions (van de Vijver & Tanzer, 1997), need to be taken into account. However, important advances have been made in exam adaptation methodologies (Hambleton, in press), and practical guidelines for adapting educational, psychological, and credentialing exams have been made available (e.g., Hambleton, 2000, in press; van de Vijver & Hambleton, 1996). By following these exam adaptation guidelines, including statistical analysis of adapted exams, cross-lingual test developers can facilitate valid interpretations of the performance of examinees that take different language versions of an exam.

By following the test adaptation guidelines, one of the statistical analysis procedures is by using equating. Equating is often necessary and anchor test design seems the most appropriate for the adapted tests. However, choosing items for the anchor will

impact the equating accuracy. The current data supported the notion that different choices for anchor items can result in different equatings and using items with the most appropriate parameters is a better choice.

Recommendations and Suggestions for Future Study

This study compared different anchor tests using a double linking method and estimated the mean standard errors of CTT equating for the common-item nonequivalent groups design. In addition, the findings of this study are based on largely descriptive results. Further investigations that study different anchor tests using IRT equating methods would be useful. Likewise, because only one test and three sets of sample sizes were used in this study, examining which anchor test is better than the other should be investigated in other testing situations. Also, the present study needs to be replicated with different language groups and more than two forms of the test. Equal sample sizes for both SL groups and TL groups should be examined. Further, in order to prevent large estimation errors, large sample sizes for TL groups need to be investigated as well.

This study focused on estimating random equating error using CTT equating methods. However, total equating error is comprised of random equating error and systematic equating error. The relative performance of the methods studied in this research study with regard to random error could change if total equating errors were considered. Thus, the behavior of systematic error should continue to be explored in conjunction with random error.

This study did not deal with issues regarding the adaptation process of the test. We should always keep in mind that some improvement in measuring and equating for

different anchor tests can be attained by improving the adaptation techniques and by carefully controlling the adaptation process (Hambleton & Patsula, 1999).

To reduce the computational burden in estimating standard errors of CTT equating, it would be beneficial for analytic expressions to be derived. Analytic expressions also would be more useful for evaluating the sample sizes needed to reach a certain precision in equating. However, such analytic expressions likely would be very difficult to derive. The only such expressions that do exist are those derived by Lord (1982) for chained true-score equating for the external common-item case. Even for this situation, Lord made simplifying assumptions that the standard errors are underestimated.

The use of CIPE software for Levine Linear equating procedure could be a problem. This software worked well in regard to the random error component investigated in this study but might not perform as well with regard to systematic error. When systematic error is of concern, this software might not work well for multiple-group procedures than the single-group procedures (Bock & Zimowski, 1996). Multiple-group procedures allow the distribution of ability to differ among the groups, which is more appropriate in a non-equivalent group design.

Based on the results and conclusions of this research study, the following suggestions are encouraged for improving the different anchor tests on the accuracy of anchor item equating practice and research in the future:

1. Given more information on the actual items, issues regarding best translation items as another way of finding anchor test should be investigated.
2. From a psychological viewpoint, making cross-lingual comparisons of a certification exam is highly complex. We cannot automatically assume that the

adapted exams will have the same meaning and same difficulty for the various language groups as they had on the original English version. This assumption needs to be carefully checked. Therefore, the researcher of this study suggests that investment of time, effort and money in the process of adapting certification exams may produce satisfactory results in terms of usability, reliability and validity of the test. Adherence to the test adaptation guidelines currently promoted by the ITC should reduce the likelihood of introducing biasing factors into the test adaptation process.

3. The issue of equating different language versions of the test clearly requires further research which may reveal whether the equating procedure has been adapted is satisfactory, or whether different equating procedures should be used.
4. To help reduce the risk of introducing bias due to the use of few criteria for evaluating equating accuracy, the applicability of using multiple criteria should be considered for future equating study.
5. Although the findings of this study emphasized the importance of a stable parameter method for anchor items selection, the researcher of this study suggests that the importance DIF method for mitigation of validity concerns should also take into consideration. That is, the items that are eventually chosen as anchor items must be from among those having the most stable parameters and at the same time these items must contain no large DIF. In this study, the items with the most stable parameters showed no large DIF.
6. The current study supported the notion that different choices for anchor items can results in different equatings. Further, items choices based on different criteria can

also result in different anchor tests. If the results of this study can be replicated with different tests and different language groups, the conclusion that different anchor tests produced different results will be strengthened. In fact, such a study is planned by the researcher.

A Brief Summary of This Study

The importance of this research study lies in finding the impact of different approaches of selecting anchor items on the accuracy of the cross-lingual equating process by comparing the verbal section and the non-verbal section of the test. In this study, two methods of selecting anchor test items were utilized. One was the parameters (appropriate item difficulty indices and item discrimination indices) method, and the other one was the DIF (least differential functioning between languages) method. The purpose of this study was to determine whether the parameter method was better than the DIF method or vice-versa. There were two ways to evaluate which method was better than the other: double liking method and MSEE method, and the double linking method was the primary method and the MSEE was the secondary method.

The results of double linking method supported the notion that different choices for anchor items can result in different equatings and using items with the most appropriate parameters method was a better choice. The results of MSEE did not show large difference between the parameter and the DIF method of anchor item selection, however, these differences were in the same direction as the primary method as well. That is, the parameter method was superior to the DIF method in virtually every situation.

REFERENCES

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

Akagi, A. (1991). Difficulties in adapting test items for cross-cultural studies. *Bulletin of the International Test Commission, 18*, 65-71.

Allalouf, A. (1998). *The development and application of a new method of calibrating the difficulty indices of items on the Psychometric Entrance Test* (Tech. Rep. No. 79). Jerusalem: National Institute for Testing and Evaluation.

Allalouf, A. (1999). *Scoring and equating at the National Institute for Testing and Evaluation (NITE Research Report 269).* Jerusalem: National Institute for Testing and Evaluation.

Allalouf, A. (2003a). Using DIF in test adaptation for cross-lingual assessment. *International Journal of Testing,10,* 6-7.

Allalouf, A. (2003b). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16* (1), 55-73.

Allalouf, A., & Rapp, J. (2002). *Equating translated verbal test forms using multiple channels.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

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

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., p.508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore: Johns Hopkins University.

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H.Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, INC.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Report No. 88–2). New York: College Entrance Examination Board.

Barnard, J. J. (1996). *In search of equity in educational measurement: traditional versus modern equating methods.* Paper presented at ASEESA's national conference at the HSRC Conference Centre, Pretoria, South Africa.

Bartram, D. (1995). The development of standards for the use of psychological tests in occupational settings: The competence approach. *The Psychologist*, *8* (4), 219–223.

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment, 12*, 62–71.

Bartram, D., & Coyne, I. (1998). Variations in national patterns of testing and test use. *European Journal of Psychological Assessment, 14,* 249–260.

Beller, M. (1996). Translated versions of Israel's inter-university Psychometric Entrance Test (PET). *International Perspectives on Academic Assessment*, *10*, 207-218.

Beller, M., Gafni, N., & Hanani, P. (2002). Constructing, adapting, and validating admissions tests in multiple languages: The Israeli case. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross cultural assessment.* Hillsdale, NJ: Erlbaum.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: theory and applications.* New York: Springer-Verlag.

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

Brennan, R. L., (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice*, *4*, 6-18.

Brennan, R. L., & Kolen, M. J. (1987a). Some practical issues in equating. *Applied Psychological Measurement*, *11*, 279–290.

Brennan, R. L., & Kolen, M. J. (1987b). A reply to Angoff. *Applied Psychological Measurement, 11*, 301–306.

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

Budgell, G.R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19, 309-321.

Bullinger, M., Anderson, R., Cella, D., & Aaronson, N. (1993). Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Quality of Life Research*, *2*, 451-459.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. *The Personnel and Guidance Journal*, *56* (5), 472-475.

Byrne, B.M. (1994). *Structural equation modeling with EQS and EQS-Windows: Basic concepts, applications, and programming.* Thousand Oaks, CA: Sage.

Byrne, B. M., Shavelson, R., & Muthen, B. (1988). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466.

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

Canadian Psychological Association (1987). *Guidelines for educational and psychological testing.* Ottawa, Canada: Canadian Psychological Association.

Casagrande, J. (1954). The ends of translation. *International Journal of American Linguistics*, *20*, 335-340.

Cook, L. L. (2000). *Factors affecting the validity of scores obtained on tests given in different languages to examinees of different cultural backgrounds*. Paper presented at the annual meeting of the International Association for Educational Assessment, Jerusalem.

Cook, L. L. & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R.K. Hambleton (Ed.), *Applications of item response theory to equate achievement tests* (Research Rep. No. RR-85-31). Princeton, NJ: Educational Testing Service.

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*, *11*, 225-244.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Crouse, J.D. (1991). *Comparing the equating accuracy from three data collection designs using bootstrap estimation methods.* Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA.

Cummins, J., Munoz-Sandoval, A. F., Alvarado, C. G., & Ruef, M. L. (1998). *The Bilingual Verbal Ability Tests.* I tasca, IL: Riverside Publishing.

Curley, W. E., & Schmitt, A. P. (1993). *Revising SAT-verbal items to eliminate differential functioning* (College Board Report No. 93–2). Princeton, NJ: Educational Testing Service.

Davison, M. L. (1985). Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin, 97,* 94–105.

Davison, M. L. (1991). *Multidimensional scaling*. Malabar, FL: Krieger.

Divgi, D. R. (1981). *Two direct procedures for scaling and equating tests with item response theory.* Paper presented at the Annual Meeting of National Council on Measurement in Education.

Dorans, N.J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*, 3-17.

Dorans, N. J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing the unexpected differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23* (3), 355–368.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, *3*, 245-254.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Eignor, D. R., & Cook, L. L. (1991). *The effect of sample and test variation on achievement test equatings.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Ellis, B. B. (1995). A partial test of Hulin´s psychometric theory of measurement equivalence in translated tests. *European Journal of Psychological Assessment*, *11*, 184-193.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, *77*, 177–184.

Ellis, B. B., & Mead, A. (1998). *Measurement equivalence of a16PF Spanish translation: An IRT differential item and test functioning analysis.* Paper presented at the 24th meeting of the International Association of Applied Psychology, San Francisco.

Eyde, L. D., & Robertson, G. J. (1994). *Improving test use in the United States: A brief history.* Paper presented at the 23rd International Congress of Applied Psychology, Madrid, Spain.

Eyde, L. D., Moreland, K. L., & Robertson, G. J. (1988). *Test user qualifications: A data-based approach to promoting good test use* (Report for the Test User Qualifications Working Group). Washington, DC: American Psychological Association.

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shewan, C. M., Harrison, P. L., Porch, B. E., Hammer, A. L., & Primoff, E. S. (1993). *Responsible test use: Case studies for assessing human behavior.* Washington, DC: American Psychological Association.

Fernandez-Ballesteros, R., Hambleton, R. K., & O'Neil, T. (2001). *European Survey on Aging Protocol: Empirical translation and adaptation results from seven countries* (Laboratory of Psychometric and Evaluative Research Report No. 404). Amherst, MA: University of Massachusetts, School of Education.

Figueroa, R. (1989). Psychological testing of linguistic-minority students: Knowledge gaps and regulations. *Exceptional Children*, *56,* 145-148.

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

Fouad, N. A. (1993). Cross-cultural vocational assessment. *The Career Development Quarterly, 42*, 4-13.

Foxcroft, C. D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment*, *13* (3), 229–235.

Fremer, J. (1996). Promoting high standards for test use: Developments in the United States. *European Journal of Psychological Assessment, 12* (1), 160–168.

Fremer, J., Diamond, E. E., & Camara, W. J. (1989). Developing a code of fair testing practices in education. *American Psychologist*, *44,* 1062–1067.

Gafni, N., & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation*, *3*, 309-319.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, *6*, 304–312.

Gierl, M. J., Rogers, W. T., & Klinger, D. A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *Alberta Journal of Educational Research, 4* (3), 353-376.

Gregoire, J. (1997). Regulation of testing practice in European French-speaking countries. *The International Test Commission Newsletter*, *7* (1), 6–13.

Gronlund, N.E. (1998). *Assessment of Student Achievement.* New York: MacMillan.

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

Harris, D. J., Welch, C. J., & Wang, T. (1994, April). *Issues in equating performance assessments.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Hambleton, R. K. (1992). *Translating achievement tests for use in cross-national Studies.* Paper presented for the Third International Mathematics and Science Study (TIMSS).

Hambleton, R. K. (1993). Translating 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. (2000). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*, 164–172.

Hambleton, R. K. (2001). Issues, designs, and technical guidelines for adapting test in multiple languages and cultures. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates, INC.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: technical issues and methods. *Bulletin of the International Test Commission, 18*, 3-32.

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. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1*, 1–12.

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

Hambleton, R. K., Merenda, P., & Spielberger, C. (2002). *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Erlbaum.

Hambleton, R. K., Sireci, S. G., & Robin, F. (1999). Adapting credentialing exams for use in multiple languages. *CLEAR Exam Review*, *10*, 24–28.

Hambleton, R. K., Yu, J., & Slater, S.C. (1999). Field-test of the ITC Guidelines for Adapting Psychological Tests. *European Journal of Psychological Assessment*, *15*, 270–276.

Harkness, J. (Ed.). (1998a). *Cross-cultural equivalence*. Mannheim, Germany: ZUMA.

Harkness, J. (1998b). *Response scales in cross-national survey research*. Paper presented at the meeting of the American Psychological Association, Toronto.

Harvey, R. J., & Murry, W. D. (1994). Scoring the Myers-Briggs Type Indicator: Empirical comparison of preference score versus latent-trait methods. *Journal of Personality Assessment, 62*, 116-129.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.), *Educational Measurement* (p. 141). Washington DC: American Council on Education.

Hofstede, G. (1997). *Cultures and organizations: Software of the mind* (Rev.ed.). New York: Mcgraw-Hill.

Holland, P. W., & Rubin, D. B. (1982). *Test equating.* New York: Academic.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hu, S., & Oakland, T. (1991). Global and regional perspectives on testing children and youth: An empirical study. *International Journal of Psychology, 26,* 329–344.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology, 18,* 115–142.

Hulin, C. L., & Candell, G. L. (1986). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 4,* 417-440.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement.* Homewood, IL: Dow Jones Irwine.

International Test Commission. (1995). *ITC directory.* Leuven, Belgium: Author.

Jeanrie, C., & Bertrand, R. (1999). *Translating tests with the International Joint Committee on Testing Practices: Code of fair testing practices in education.* Washington, DC: Author.

Joint Committee on Testing Practices. (2000). *Rights and responsibilities of test takers: Guidelines and expectations.* Washington, DC: Author.

Kendall, I., Jenkinson, J., De Lemos, M., & Clancy, D. (1997). *Supplement to guidelines for the use of psychological tests.* Sydney: Australian Psychological Society.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*, 131-143.

Koene, C. J. (1997). Tests and professional ethics and values in European psychologists. *European Journal of Psychological Assessment, 13,* 219–228.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*(3), 197–206.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 9,* 25-44.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 4,* 29-36.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3,* 97-104.

Kolen, M.J. (1995). *CIPE user's guide*. Iowa: Iowa Testing Programs.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: methods and practices.* New York: Springer.

Kolen, M. J., & Harris, D. J. (1990). Comparison of item pre-equating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement, 27,* 27-39.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling.* Newbury Park, CA: Sage.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23,* 4–14.

Liou, M., & Bond, L. (1985). *A theoretical investigation of error components in item response theory equating.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

LISA Frequently Asked Questions. *Geneva, Switzerland: Localization Industry Standards Association.* [On-line]. Available: http://www.lisa.unige.ch/infor/faqs.html

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

Lord, F. M. (1982). Standard error of an equating by item response theory. *Applied Psychological Measurement, 6,* 463–472.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equating. *Applied Psychological Measurement, 8,* 453-461.

Lonner, W. J., & Berry, J. W. (Eds.) (1986). *Field methods in cross-cultural research.* Beverly Hills: Sage.

Loret, P.G. (1975). *Implementing, evaluating, and using a statewide assessment program: logistics and contracted services.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

Marco, G. L., Peterson, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating methods. In D.White (Ed.), *New Horizons in Testing* (pp. 147-177). New York: Academic.

Marks, E., & Lindsay, C. A. (1972). Some results relating to test equating under relaxed test form equivalence. *Journal of Educational Measurement, 9,* 45-55.

McBride, J. R., & Weiss, D. J. (1974). *A word knowledge item pool for adaptive ability measurement* (Research Report, 74-2). University of Minnesota, Psychometric Methods Program.

Mckinley, R. L., & Reckase, M. D. (1981). *A comparison of procedure for constructing large item pools.* Columbia, No.: Missori University, Tailored Testing Research Laborary.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3* (2nd ed.) [Computer program]. Mooresville, IN: Scientific Software International.

Miura, I. T., Okamoto, Y., Kim, C. C., Steere, M., & Fayol, M. (1993). First graders' cognitive representation of number and understanding of place value: Cross-national comparisons. *Journal of Educational Psychology, 1,* 24-30.

Morante, E. (1987). A primer on placement testing. In D. Bray & M. Belcher (Eds.), *Issues in student assessment: New directions for community colleges, 59*, 55-63. San Francisco& London: Jossey-Bass Inc. & ERIC clearinghouse for Junior colleges.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S., & Most, R. B. (1995). Assessment of test user qualifications: A research-based measurement procedure. *American Psychologist, 50,* 14–23.

Morris, C. N. (1982). On the foundations of test equating. In P.W. Holland & D.B. Rubin (Eds.), *Test equating* (pp. 160-191). New York: Academic.

Muñiz, J., & Hambleton, R. K. (1997). *Directions for the translation and adaptation of tests.* Papeles del Psicologo, August, 63–70.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1,* 115–135.

Newmark, P. (1988). *A textbook of translation.* New York: Prentice Hall.

Nunnally, J. C. (1972). *Educational measurement and evaluation* (2nd ed.). New York: McGraw-Hill.

Oosterhof, A. (2001). *Classroom Applications of Educational Measurement (3rd ed.).* Columbus, Ohio: Merrill Prentice Hall.

Pearson, B. Z., Fernandez, M. C., & Oller, D. K. (1992). Measuring bilingual children's receptive vocabularies. *Child Development, 63,* 1012-1221.

Pearson, B. Z., Fernandez, M. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning, 43,* 93-120.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics*, *8*, 137-156.

Pena, E. D., Bedore, L. M., & Zlatic-Giunta, R. (2002). Development of categorization in young bilingual children. *Journal of Speech, Language, and Hearing Research, 45* (5), 938-947.

Poortinga, Y. H., & van de Vijver, F. J. R. (1991). Testing across cultures. In R.K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277–308). Boston: Kluwer.

Popham, W. J. (1981). *Modern educational measurement*. Englewood Cliff, NJ: Prentice-Hall.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983). *The effect of anchor test size in vertical equating with Rasch and three-parameter models.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Rapp, J. & Allallouf, A., (2002). *Evaluating cross-lingual equating*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Restrepo, M. A., & Silverman, S. W. (2001). Validity of the Spanish preschool language scale-3 for use with bilingual children. *American Journal of Speech-Language Pathology, 10,* 382-393.

Robertson, G. J., & Eyde, L. D. (1993). Improving test use in the United States: The development of an interdisciplinary casebook. *European Journal of Psychological Assessment, 9,* 137–146.

Robin, R., Sireci, S.G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing, 3* (1), 1-20.

Saville & Holdsworth Ltd. (1993). *Best practice in the management of psychometric tests: Guidelines for developing policy.* Surrey, England: Author.

Shackleton, V., & Newell, S. (1994). European management selection methods: A comparison of five countries. *International Journal of Selection and Assessment, 2,* 91–102.

Simner, M. L. (1996). Recommendations by the Canadian Psychological Association for improving the North American safeguards that help protect the public against test misuse. *European Journal of Psychological Assessment, 12,* 72–82.

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

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5,* 299–321.

Sireci, S. G., & Berberoglu, G. (2001). Using bilingual respondents to evaluate translated – adapted items. *Applied Measurement in Education, 13* (3), 229-248.

Sireci, S. G., Bastari, B., & Allalouf, A. (1998). *Evaluating construct equivalence across adapted tests* (Laboratory of Psychometric and Evaluative Research Report No. 340). Amherst: University of Massachusetts, School of Education.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). *Adapting credentialing examinations for international uses.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Sireci, S. G., Harter, J., Yang. Y., & Bhola, D. (2000). *Evaluating the construct equivalence of international employee opinion surveys* (Laboratory of Psychometric and Evaluative Research Report No. 379). Paper presented at the

Annual Meeting of the National Council on Measurement in Education, New Orleans.

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

Skaggs, G., & Lissitz, R. W. (1986b). An exploration of the robustness of four tests equating models. *Applied Psychological Measurement, 10,* 303-317.

Spray, J. A. (1990). One-parameter item response theory models for psychomotor tests involving repeated, independent attempts. *Research Quarterly for Exercise and Sport, 61,* 162-168.

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics (3rd ed.).* New York: HarperCollins Publishers, Inc.

Tamayo, J. (1987). Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation. *Educational and Psychological Measurement, 47,* 893-902.

Tanzer, N. K. (1999). *Using instruments in multicultural and multilingual applications.* Paper presented at the Cross-Cultural Conference, Graz, Austria.

Tanzer, N. K., & Sim, C.O.E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for test adaptations. *European Journal of Psychological Measurement, 15,* 258–269.

Test Publishers Association. (1994a). *Responsible test use. Guidelines for test publishers and test users.* London: The Publishers Association.

Test Publishers Association. (1994b). *Responsible educational test use. Guidelines for test publishers and test users.* London: The Publishers Association.

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software International.

Thompson, B. & Daniel, L. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197-208.

Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement, 8,* 64.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). New York: MacMillan.

Toury, G. 1978. The Nature and Role of Norms in Translation. In Venuti, L. *The Translation Studies Reader.* London: Routledge.

Tsai, T-H, Hanson, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for common-item nonequivalent groups design. *Applied Measurement Education, 1,* 17-30.

Tyler, B. (1991). *Using tests responsibly and effectively - The role of professional judgment in assessment.* Paper presented at the 2nd European Congress of Psychology, Budapest, Hungary.

Tyler, B., & Miller, K. M. (1990). *Test user qualification: A data base for preventing test misuse: Setting the standards.* Paper presented at the 23rd International Congress of Applied Psychology, Kyoto, Japan.

Umbel, V. M., Pearson, B. Z., Fernandez, M. C., & Oller, D. K. (1992). Measuring bilingual children's receptive vocabularies. *Child Development, 63,* 1012-1020.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross cultural research.* Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In Hambleton, R. K. & Zaal, J. N. (Eds.), *Advances in educational and psychological testing: Theory and applications.* Boston: Kluwer Academic Publishers.

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

Vermeer, H. 1989. Skopos and Commission in Translational Activity. In Venuti, L. *The Translation Studies Reader.* London: Routledge.

Waldron, B. (1988). *LEQUATE program information. [Computer Program].*

Wang, Z-M. (1993). Psychology in China: a review. *Annual Review of Psychology, 44,* 87-116.

Walter, S. (1997). *Globalization, adult education and training: impacts and issues.* New York: Zed Books.

Wiersma, W. & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Boston, MA: Allyn and Bacon.

Wingersky, M. S., Barton, M. A., & Lord, R. M. (1982). *LOGIST user's guide* [Computer program]. Princeton NJ: Educational Testing Service.

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

Woldbeck, T. (1998). *Basic Concepts in Modern Methods of Test Equating.* Paper presented at the Annual Meeting of the Southwest Psychological Association, New Orleans.

Wood, D. A. (1960). *Test construction: Development and interpretation of achievement tests.* Columbus, OH: Charles E. Merrill Books, Inc.

Wood, D. J. (1989). A comparison of three linear equating methods for the common-item nonequivalent populations design. *Applied Psychological Measurement*, *13*, 257-261.

Wright, B. D. (1997). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-116.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press University of Chicago.

Wright, N. K. & Dorans, N. J. (1990). *Using the selection variable for matching or equating.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.

Yen, W. M. (1983). Tau-equivalence and equipercentile equating. *Psychometrika, 48,* 353-369.

Zeng, L. (1993). A numerical approach for computing standard errors of linear equating. *Applied Psychological Measurement*, *17*, 177-186.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed- score equating of number-correct scores. *Applied Psychological Measurement, 19,* 231-240.

Zhang, H-C. (1988). Psychological measurement in China. *International Journal of Psychology, 23,* 101-117.

Zhu, W-M. (1998). Test equating: what, why, how? *Research Quarterly for Exercise and Sport, 1, 11-23.*

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). Bilog MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items. Chicago: Scientific Soft-ware International, Inc.

APPENDICES

APPENDIX A

Exemption Letter from the Institutional Review Board

OHIO
UNIVERSITY
Office of the Vice President
for Research

04E037

Office of Research Compliance
Research and Technology
Grosvenor 117
Athens OH 45701-2979

T. 740.593.0664
F. 740.593.9838
www.ohiou.edu/research

A determination has been made that the following research study is exempt from IRB review because it involves:

Category 4 . research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens if publicly available or recorded without identifiers

Project Title: The Effect of Different Anchor Selection Approaches on the The Accuracy of Test Equating for Test Adaptation

Project Director: Hua Gao

Department: Education - Research and Evaluation

Advisor: George Johanson

Rebecca Cale

Rebecca Cale, Associate Director, Research Compliance
Institutional Review Board

9/26/04
Date

APPENDIX B

Scree Plots of the Principle Component Analysis

Figure 21

Korean Language: Scree Plot of the Principle Component Analysis for Form A

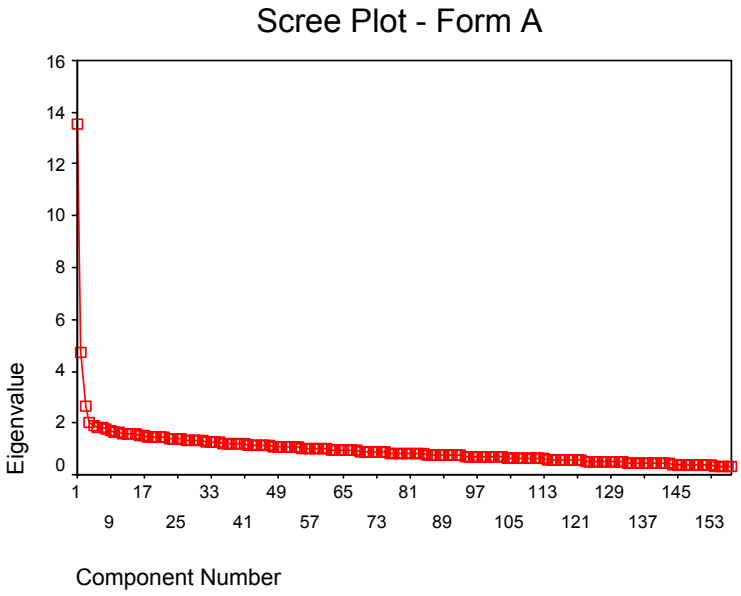


Figure 22

Korean Language: Scree Plot of the Principle Component Analysis for Form B

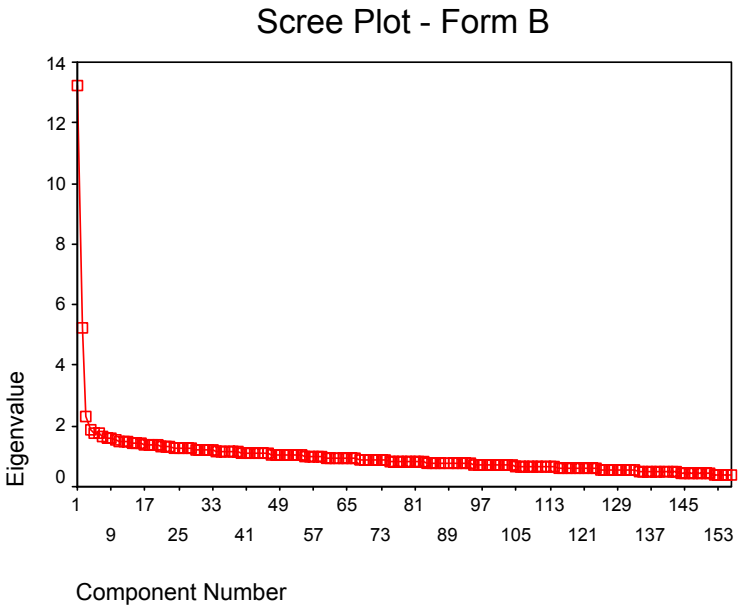


Figure 23

Spanish Language: Scree Plot of the Principle Component Analysis for Form A

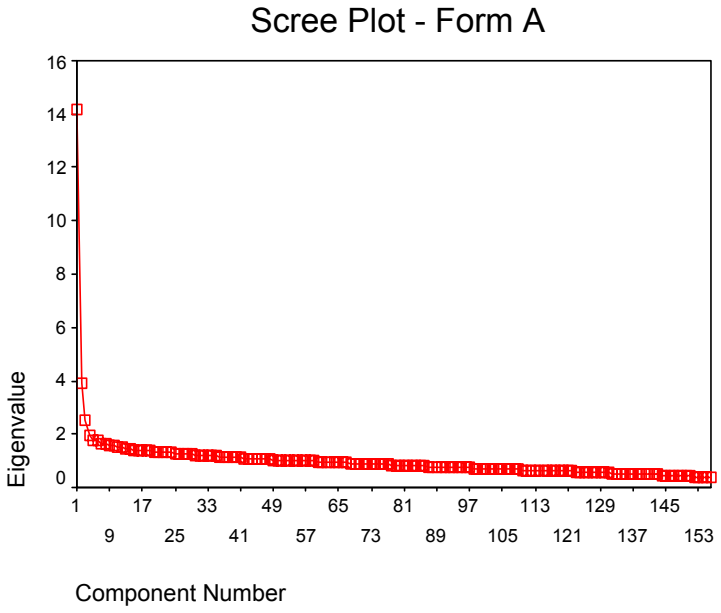


Figure 24

Spanish Language: Scree Plot of the Principle Component Analysis for Form B

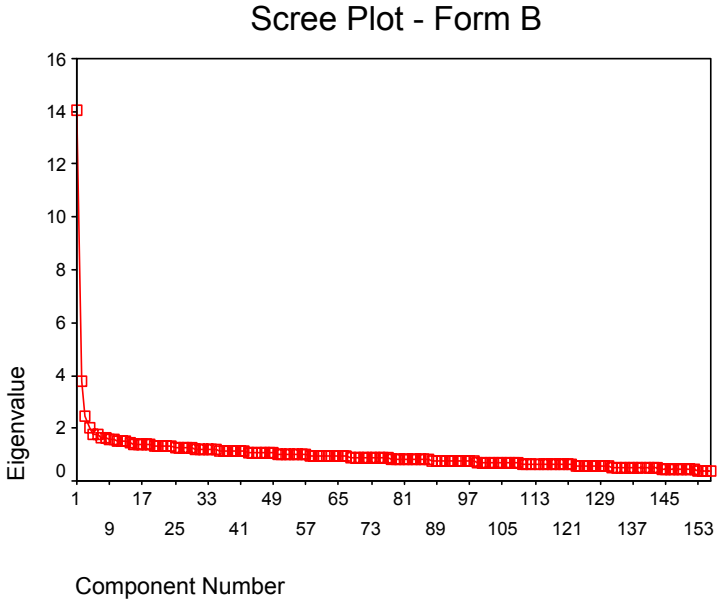


Figure 25

Chinese Language: Scree Plot of the Principle Component Analysis for Form A

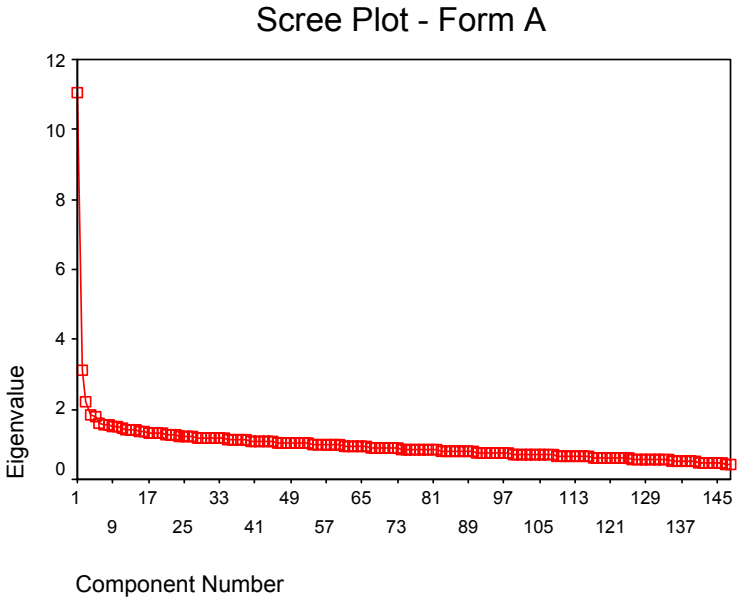
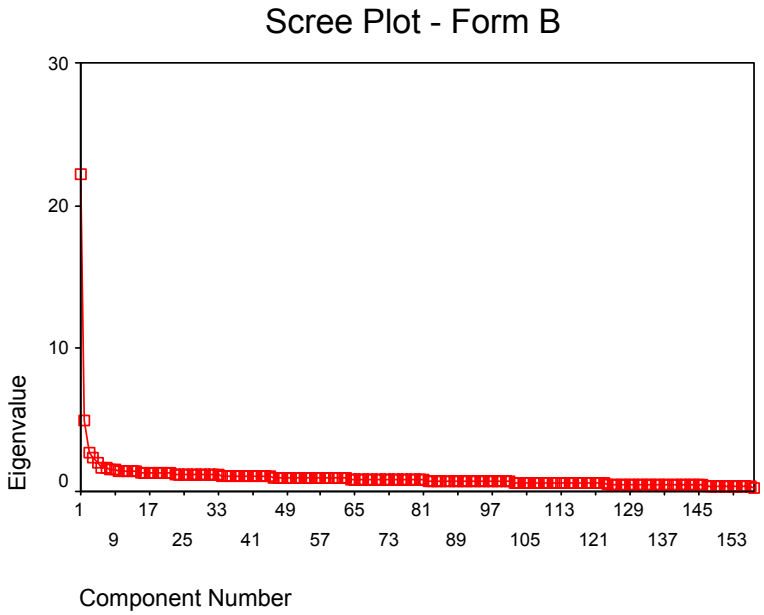


Figure 26

Chinese Language: Scree Plot of the Principle Component Analysis for Form B



APPENDIX C

Item Difficulty Indices and Item Discrimination Indices

Table 32

Item Difficulty Indices and Item Discrimination Indices by Languages (Korean and English) and Forms (From A and Form B)

	Form A				Form B			
Items	Korean		English		Korean		English	
	Diff.	Disc.	Diff.	Disc.	Diff.	Disc.	Diff.	Disc.
1.	.9859	.0492	.9166	.2765	.9431	.3063	.9712	.1820
2.	.8028	.3681	.6617	.3163	.9675	.1232	.9705	.1898
3.	.4648	.4572	.4537	.1388	.8455	.1379	.8108	.1288
4.	.7183	.1637	.7691	.1674	.9106	.3976	.9128	.2812
5.	.8028	-.0923	.8160	.1548	.0000	.0000	.0000	.0000
6.	.9437	.2925	.9280	.2875	.8374	.2059	.8769	.2413
7.	.9437	.1785	.9269	.2453	.8780	.3315	.9072	.3235
8.	.3521	.1056	.6046	.0350	** .5203	.2213	.4114	.2138
9.	.6761	-.0080	.9314	.3122	.7480	.1437	.8812	.3326
10.	.8873	.2236	.7840	.3285	.8293	.1906	.7883	.1076
11.	.9437	.1744	.6549	.4011	.8130	.3095	.7131	.3437
12. *	.8592	.2843	.7874	.2753	.7642	.1826	.7651	.3787
13.	.9014	.4738	.7771	.3858	.9106	.3685	.7911	.3448
14.	.9859	.1601	.8629	.3435	.9675	.2507	.9015	.3594
15.	.4648	.0465	.4869	.2376	.4065	.0675	.5218	.2687
16.	.7887	.1909	.5314	.2783	.8943	.2077	.8446	.2231
17.	.6479	.0268	.7177	.1884	.5285	.2294	.6006	.1856
18.	.9718	.0759	.8869	.3532	.9024	.2142	.9015	.2945
19.	.8310	.1809	.9166	.2032	.5528	.2158	.8207	.1412
20.	.7042	-.0299	.8777	.1656	.8943	.2850	.8586	.1851
21. *	.7183	.2624	.6663	.3705	.8049	.2308	.6850	.3615
22.	.8310	.3828	.7440	.2773	.8211	.1921	.7707	.3564

23.	*	.5634	.2917	.7086	.3314	**	.7805	.3294	.7989	.2002
24.		.7465	.0281	.9074	.2252		.8374	.1751	.9233	.1575
25.	*	.6479	.2554	.7211	.2992		.3171	.0593	.6287	.1433
26.		.2958	.5022	.2640	.3138		.2846	.3159	.3165	.3047
27.		.2535	.0494	.3954	-.0776		.4472	.2011	.5436	.1667
28.		.5352	.1746	.7177	.2905	**	.5854	.2000	.7518	.3244
29.		.1972	.0657	.1646	.2565		.7317	.3056	.5921	.2441
30.	*	.6197	.5507	.5326	.4253		.5854	.2453	.8221	.2863
31.		.6901	.3633	.5691	.1797		.8049	.4397	.8622	.3011
32.		.3662	.2939	.2766	.1443		.3902	-.0464	.4571	.0211
33.		.8732	.1798	.7429	.2748		.8618	.0679	.9255	.2315
34.		.6056	.3739	.8686	.3194	**	.5528	.2807	.6498	.2546
35.		.8169	.2791	.8057	.3543		.8130	.3041	.7806	.3076
36.		.8873	.1671	.6377	.2765		.6179	.1148	.5703	.2001
37.		.6479	.0935	.8091	.0573		.8537	.1677	.8945	.2715
38.		.9014	.3244	.9040	.2413		.0000	.0000	.0000	.0000
39.		.7183	.1156	.9063	.2694		.7967	.4080	.9198	.3432
40.	*	.6620	.2181	.5726	.2793	**	.6829	.3277	.5316	.2990
41.		.3803	.0326	.3920	.1110		.4878	.2195	.7918	.3603
42.		.6901	.1055	.7943	.3479		.7724	.0304	.6842	.1926
43.		.8873	.3879	.8389	.2443	**	.7236	.2457	.7321	.2042
44.		.7465	.0733	.8251	.1218		.7805	.2290	.5802	.2137
45.		.5070	.2315	.3806	.1763	**	.6667	.3058	.6885	.3332
46.		.6056	.2590	.7897	.1814	**	.6423	.3860	.6821	.2830
47.		.8028	.0952	.8091	.2000		.5285	.1889	.8101	.2446
48.		.0000	.0000	.0000	.0000		.4146	.0417	.6793	.1808
49.		.7746	-.0362	.6034	.1524		.5935	.1277	.6350	.2684
50.	*	.7465	.3139	.5783	.3050		.9431	.3732	.9508	.2400
51.	*	.7746	.2105	.7303	.2509		.5854	.1642	.6231	.1914
52.	*	.6197	.3038	.5543	.3858		.8943	.2161	.8052	.1331

53.	*	.7183	.3765	.6949	.2764	** .5854	.3406	.6428	.3502
54.		.7042	.2977	.3531	.0735	.9024	.2577	.8537	.2871
55.		.5352	.1822	.5531	.1941	.0000	.0000	.0000	.0000
56.		.9296	.2501	.7303	.4259	.9024	.2420	.7883	.4199
57.		.8873	.1286	.8137	.2379	** .6341	.2208	.5851	.2126
58.		.4366	.1938	.6469	.1785	** .7317	.2036	.7961	.2629
59.		.4225	.1342	.4880	.0112	** .7317	.3503	.5942	.2806
60.		.1972	.0540	.6526	.3452	** .7886	.3712	.8158	.3361
61.		.7746	.0039	.7040	-.0379	.3415	.0407	.2890	.1096
62.		.9155	-.0613	.8629	.3368	** .7154	.2457	.6723	.2100
63.		.5352	-.0022	.6583	.3339	.3252	.0121	.3439	.0636
64.		.8169	.2254	.7360	.3226	.8211	.0564	.8917	.2600
65.	*	.7606	.4772	.6914	.2387	.7398	.2459	.9233	.2693
66.		.9437	.1947	.9691	.1667	.7805	.1191	.8270	.2494
67.		.9437	.3987	.9611	.2688	.4634	-.0020	.4346	.2192
68.	*	.8169	.2816	.7463	.3749	.9106	.3395	.9480	.3200
69.		.4225	.3085	.3154	.0928	.8374	.3028	.7482	.2565
70.		.9155	.0560	.8914	.1996	.9106	.1693	.8235	.1650
71.	*	.6056	.4346	.5097	.3195	.5285	.2513	.5218	.3731
72.	*	.7465	.3795	.6891	.3208	.7480	.0146	.7672	.1105
73.		.4930	.0624	.3897	.2568	** .7317	.2071	.6624	.2889
74.	*	.4648	.2543	.4571	.3164	.7886	.2121	.8551	.2485
75.	*	.7465	.2203	.5931	.3274	** .7480	.3085	.6287	.2825
76.		.0000	.0000	.0000	.0000	** .4228	.2439	.3368	.2537
77.		.5915	.2652	.7189	.1780	.9593	.3046	.9585	.1817
78.	*	.4648	.2676	.6194	.2619	.8211	.0457	.7665	.2781
79.		.4507	.1452	.5006	.1702	.9024	.4816	.8854	.2719
80.		.6761	.1120	.6400	.2819	.4472	.1792	.6744	.3208
81.		.8451	-.0929	.8023	.1988	.8618	.1007	.8530	.3471
82.	*	.5915	.4204	.4251	.2702	.9431	.1107	.8762	.3509

83.	.6338	.3795	.3817	.2692	.5610	.3831	.4184	.2819
84.	.5352	.1048	.5497	.1248	.9756	.3903	.9536	.2655
85.	.4225	.3471	.3097	.3370	.8049	.2531	.9001	.2654
86.	.3662	.2015	.4160	.2834	.4472	.1469	.4740	.2773
87.	1.0000	.0000	.8069	.2907	** .6504	.3014	.5331	.3352
88.	.8169	.4191	.4800	.3916	** .6992	.3496	.4930	.3316
89.	.9155	-.0680	.7989	.2340	.3171	.0482	.3762	.0730
90.	.2817	.3429	.5691	.2331	.5203	.1445	.4529	.0543
91.	.9155	.1941	.8880	.3102	.2927	.0147	.3291	.0434
92.	.6761	.3913	.8914	.3252	.8618	.3433	.9030	.2990
93. *	.7746	.2965	.7840	.3493	** .7561	.4130	.7672	.3436
94. *	.3662	.4665	.4309	.2982	.9268	.4775	.8854	.2641
95.	.9296	.0706	.8983	.3054	.2033	.1465	.1624	.1682
96.	.2535	.3312	.2274	.1691	** .6098	.2888	.6857	.3271
97. *	.8310	.2715	.7909	.3559	.6423	.2781	.8122	.1259
98.	.9859	-.0062	.9406	.3175	.4553	.2765	.4480	.4493
99.	.7606	.3563	.5657	.3066	.7236	.2236	.6512	.3404
100.	.9577	.2171	.8320	.3261	.7236	.2086	.8980	.2677
101.	.2958	.1624	.4811	.0512	** .4878	.3154	.6167	.3842
102.	.7465	.0410	.6491	.2139	** .7317	.3668	.6871	.3529
103. *	.8732	.2477	.8069	.2178	.2439	.1907	.3003	.1783
104.	.6197	.1362	.6571	.2587	** .4878	.2914	.4522	.2917
105.	.2113	.0691	.1440	.0280	.6260	.1899	.5823	.1693
106.	.9577	.0451	.6571	.1737	.4715	-.1331	.3481	.0471
107. *	.6761	.3628	.5589	.2878	** .6667	.3968	.6034	.3570
108.	.9155	.1536	.8286	.1110	** .6667	.2925	.5816	.3269
109.	.8592	.1139	.8183	.1868	.9106	.2687	.9212	.2862
110. *	.7746	.2421	.6069	.4505	** .7805	.2641	.6793	.3743
111. *	.7746	.2829	.6411	.2898	.8618	.1156	.8882	.2401
112.	.9437	-.2176	.9189	.2167	.8780	.3457	.9023	.3823

113.	*	.8732	.4470	.8183	.2745	.6748	.2377	.7693	.1969
114.		.9014	.2201	.5646	.3655	** .7236	.3598	.6006	.3442
115.		.4366	.0592	.2903	.2311	** .5041	.2499	.5274	.2116
116.		.8873	.3729	.7943	.4130	.8780	.4906	.8101	.3916
117.		.4507	.2116	.7291	.1869	.9024	.2682	.9198	.2859
118.		.3380	.2351	.5017	.1639	.7642	.4019	.7166	.2912
119.		.8451	.1030	.8777	.1671	.7236	.0966	.7412	.1388
120.		.2535	.2288	.7269	.2719	.5610	.1596	.6857	.3137
121.	*	.5915	.4771	.4011	.4534	.3089	.0694	.5120	.2027
122.	*	.6901	.2319	.8377	.2070	** .4959	.2669	.4740	.2690
123.		.8169	.1572	.5440	.3406	** .5691	.2603	.7180	.2549
124.		.6338	.2924	.4149	.3298	.8211	.3003	.9030	.2913
125.		.8873	.2326	.8274	.3817	.9512	.4869	.8805	.3617
126.		.8732	.1742	.8343	.3389	.4715	.2884	.5661	.2156
127.		.8732	.2931	.7177	.2139	.7236	.2619	.7039	.3447
128.	*	.3803	.4225	.4137	.2800	.8943	.1172	.9149	.1910
129.		.9014	.2770	.7257	.1542	.9756	-.0491	.9058	.3224
130.		.8873	.0486	.7120	.2599	** .8537	.2130	.7419	.3034
131.		.9296	.2391	.9086	.3204	.9268	.0374	.9269	.1269
132.		.9437	.2151	.9246	.2592	.7073	.3122	.7820	.3481
133.	*	.5634	.3455	.6286	.3857	.8374	.1374	.8861	.2742
134.	*	.7183	.3025	.5977	.3549	** .4472	.3741	.4719	.3303
135.		.9577	-.1770	.8571	.1969	.8780	.2618	.8727	.3522
136.		.9155	.0627	.9406	.2384	.5772	.0409	.6273	.2264
137.	*	.8028	.2513	.8343	.2170	.7805	.1752	.7215	.1854
138.		.7887	.0277	.8011	.3016	.9187	.4769	.9128	.3944
139.		.8451	.4372	.8057	.3424	.8780	.2681	.8622	.1896
140.		.6338	.1863	.7371	.2865	.3171	-.0079	.3150	.0319
141.		.5211	.2428	.7154	.0979	.6911	.1860	.8010	.2637
142.	*	.7465	.2529	.6137	.2774	.0000	.0000	.0000	.0000

143.	.8873	.3579	.7349	.3489	** .8293	.4628	.7574	.3132
144.	.8732	.2789	.9029	.2794	.5691	.1565	.4402	.2433
145.	.6056	.3076	.3497	.2411	.8862	.1436	.9487	.2848
146.	.6761	.2470	.4217	.3967	.5122	.1945	.4831	.3303
147. *	.7606	.4211	.8880	.2021	.7073	.1628	.6786	.1528
148. *	.7042	.3143	.5703	.3666	.6260	.3264	.6744	.3842
149.	.9296	.0632	.5897	.4124	.8374	.2649	.6132	.2817
150.	.8451	.3139	.6091	.2285	** .6260	.3254	.6259	.2874
151.	.7746	.2829	.4857	.2402	** .8293	.4017	.7152	.3771
152.	.8592	.3333	.6983	.4271	.8211	.2786	.8298	.1862
153.	.5775	.1129	.5417	.2948	.5610	.2254	.4114	.2605
154.	.6197	-.0698	.6594	.0720	.9675	.2913	.9494	.3179
155.	.6338	.1179	.4697	.1907	.4878	.2528	.6617	.2148
156.	.9014	.0817	.8571	.2908	.8211	.3328	.9015	.3044
157. *	.7324	.2223	.6914	.2405	.8862	.2085	.9339	.2582
158.	.9296	.2685	.8434	.2039	.5447	.0303	.4543	.1126
159.	.8310	.0558	.9051	.2866	.8780	.2570	.9191	.2908
160.	.9577	.0311	.9086	.2735	.5041	-.0368	.5949	.1434

Note. Based on item difficulty indices and item discrimination indices, * indicates that the items will be chosen as anchor items for Form A; ** indicates that the items will be chosen as anchor items for Form B.

Table 33

Item Difficulty Indices and Item Discrimination Indices by Languages (Spanish and English) and Forms (From A and Form B)

		Form A					Form B			
Items		Spanish		English			Spanish		English	
		Diff.	Disc.	Diff.	Disc.		Diff.	Disc.	Diff.	Disc.
1.		.8226	.1866	.9285	.2365		.8333	.2132	.9306	.2189
2.		.9194	.2409	.9457	.2619		.9123	.2295	.9466	.2453
3.		.5968	.2325	.7311	.2754	**	.6228	.3230	.7321	.2786
4.		.8548	.0975	.8453	.3426		.8509	.1682	.8466	.3342
5.		.5968	.1704	.5578	.2230		.5439	.2883	.5593	.2210
6.	*	.7097	.4746	.8143	.3796	**	.7018	.4946	.8154	.3721
7.	*	.5484	.4018	.6747	.3426	**	.5789	.4180	.6766	.3367
8.		.8387	.1942	.7957	.2703	**	.7895	.2772	.7974	.2643
9.		.9677	.1028	.9759	.1724		.9649	.1549	.9771	.1427
10.		.5323	-.1974	.5550	.0543		.5439	.0070	.5538	.0532
11.		.7581	.1658	.8177	.2834	**	.7719	.3024	.8189	.2842
12.	*	.5161	.4263	.6355	.3812	**	.5351	.5192	.6371	.3795
13.		.7903	.3545	.8927	.2723		.8421	.2894	.8917	.2772
14.		.9677	.2250	.9347	.3318		.9561	.2663	.9362	.3182
15.	*	.5161	.4105	.5997	.2310		.4737	.3039	.6031	.2215
16.		.3871	.0341	.4450	.1542		.4298	.1693	.4462	.1429
17.		.7419	.3256	.8549	.2286		.7632	.2594	.8577	.2210
18.		.8548	.4558	.8618	.3826		.8596	.4350	.8619	.3802
19.	*	.6774	.2605	.6499	.3779		.5965	.3339	.6495	.3781
20.		.6613	-.0516	.6960	.0638		.6842	.0513	.6960	.0640
21.		.6290	.2279	.6733	.4029		.6053	.2917	.6759	.3980
22.	*	.6774	.3691	.7730	.3237	**	.7018	.3530	.7745	.3116

23.		.8710	.2012	.8184	.2593		.8596	.0624	.8196	.2476
24.		.5645	.3264	.7813	.1438		.5877	.3187	.7828	.1297
25.		.7258	.3823	.8590	.2247		.7368	.3103	.8591	.2174
26.		.5968	.1102	.6224	.2999		.6053	.2334	.6253	.3053
27.		.4516	.1122	.6403	.3048		.5088	.2060	.6412	.3089
28.	*	.6774	.3016	.7338	.3014		.5702	.2385	.7349	.2999
29.		.6774	.2437	.6795	.3460	**	.6404	.4085	.6794	.3411
30.	*	.5645	.3035	.6637	.2929	**	.5088	.3075	.6627	.2962
31.		.8871	.0771	.8384	.1260		.8684	.0692	.8383	.1184
32.		.9032	.2621	.9312	.2766		.9123	.2390	.9299	.2770
33.		.9677	-.0046	.9133	.2470		.9211	.1207	.9126	.2518
34.	*	.5968	.2165	.6094	.2694		.5000	.1517	.6079	.2714
35.		.9355	.2965	.9333	.1347		.9211	.1140	.9334	.1278
36.	*	.7903	.3717	.8329	.3648		.7719	.3705	.8328	.3608
37.		.6935	.1055	.7173	.2377		.6228	.1252	.7189	.2242
38.		.7581	.4227	.8184	.1471		.6842	.1470	.8175	.1376
39.		.6452	.1551	.7503	.3177		.6754	.1330	.7495	.3134
40.		.6290	.6272	.5578	.3369	**	.5877	.4687	.5559	.3379
41.		.8226	-.0305	.8583	.2046		.8509	-.0907	.8605	.2029
42.		.8065	.0859	.8054	.2732		.7982	.2104	.8071	.2608
43.		.2903	.1873	.2785	.2096		.2719	.3310	.2762	.2121
44.		.3871	.1194	.4470	.1976		.3421	.1635	.4455	.2013
45.		.8710	.3621	.9023	.3080		.8860	.1820	.9035	.2966
46.		.5161	.2176	.4127	.2597		.4912	.2655	.4122	.2624
47.		.6290	.5646	.8391	.3002		.6228	.5435	.8390	.2956
48.		.7903	.0980	.8122	.0826		.7544	.1351	.8119	.0848
49.		.5161	.1428	.4828	.2129		.4912	.2049	.4830	.2118
50.		.8065	.3345	.8631	.3173	**	.7193	.4111	.8626	.3193
51.		.9677	.0686	.9574	.3102		.9649	.0551	.9577	.2969
52.		.0000	.0000	.0000	.0000		.0000	.0000	.0000	.0000

53.		.5806	.1787	.5674	.1929		.5351	.2859	.5690	.1804
54.	*	.5645	.6043	.6376	.3748	**	.5526	.5554	.6378	.3713
55.		.8065	.1780	.7290	.2654	**	.7895	.2650	.7300	.2676
56.	*	.6613	.3667	.7008	.3170		.6754	.2660	.7023	.3160
57.	*	.8226	.4149	.7937	.2944		.8421	.4286	.7932	.2944
58.		.0000	.0000	.0000	.0000		.0000	.0000	.0000	.0000
59.	*	.5645	.3530	.6747	.3053	**	.5439	.3467	.6759	.2950
60.	*	.7419	.4499	.7820	.2666	**	.7018	.4137	.7835	.2562
61.		.6935	.0230	.7847	.1323		.7105	.1839	.7856	.1371
62.		.3387	.1711	.5089	.2629		.3333	.2596	.5073	.2699
63.		.2097	.3373	.2724	.2690		.2105	.3221	.2734	.2688
64.	*	.7258	.4433	.7517	.2774	**	.6754	.3367	.7516	.2747
65.		.3387	.3686	.4477	.3047		.3333	.3848	.4476	.3063
66.		.4839	.2805	.4807	.1700		.5702	.1418	.4809	.1679
67.		.8387	.3599	.9037	.2030		.8509	.3455	.9056	.1930
68.	*	.7419	.5204	.7173	.4655	**	.7368	.5379	.7169	.4616
69.		.4355	.4720	.4746	.1869		.4561	.4079	.4754	.1867
70.	*	.7097	.3237	.6685	.3516	**	.6842	.3886	.6683	.3519
71.		.0000	.0000	.0000	.0000		.0000	.0000	.0000	.0000
72.		.5323	.3627	.4897	.1497		.4474	.2531	.4899	.1429
73.		.6774	.2531	.7428	.3788		.6930	.2239	.7425	.3715
74.	*	.7742	.2240	.8294	.2640		.8070	.3006	.8307	.2524
75.		.8387	.4289	.8109	.4028		.8333	.3928	.8099	.4063
76.		.4677	.3768	.5660	.4622	**	.4649	.3863	.5635	.4801
77.		.4839	.4281	.5289	.2173		.4386	.3533	.5295	.2143
78.		.5161	.2805	.6850	.2944		.5175	.2996	.6829	.3005
79.	*	.5161	.2456	.6451	.2053		.5877	.2312	.6468	.2053
80.		.8871	.2029	.8845	.2442		.8860	.1367	.8848	.2477
81.		.4677	-.0568	.5805	.1926		.5526	.1028	.5829	.1914
82.	*	.8226	.4264	.7696	.2807		.7456	.2935	.7703	.2650

83.		.3226	.4255	.5447	.3120		.4298	.2835	.5475	.3093
84.		.5968	.0961	.6142	.1371		.6140	.1549	.6155	.1302
85.	*	.7097	.3314	.7407	.2698	**	.7018	.3560	.7405	.2688
86.		.7258	.4650	.9078	.2754		.7544	.3672	.9091	.2703
87.		.8710	.2349	.9085	.2630		.8684	.2327	.9105	.2472
88.		.6452	.3851	.7180	.2673		.5965	.4389	.7189	.2594
89.	*	.6613	.4783	.6107	.4483	**	.6140	.4974	.6086	.4490
90.	*	.7419	.2139	.8026	.4144	**	.7018	.3620	.8043	.4131
91.		.4677	.3030	.6912	.3139	**	.5614	.3686	.6947	.3097
92.		.6613	.2871	.6410	.3859	**	.6316	.4107	.6412	.3815
93.		.3871	.2050	.4388	.2564		.3246	.2297	.4407	.2557
94.		.4839	.1272	.4966	.2250		.4211	.1570	.4969	.2204
95.		.1935	.2682	.1520	.1438		.1667	.2003	.1527	.1415
96.		.9032	.3975	.9326	.1869		.8947	.2294	.9327	.1814
97.		.6290	.4146	.8061	.2587		.6930	.3073	.8078	.2492
98.		.5323	.2872	.5997	.0552		.5000	.2593	.6003	.0462
99.		.9516	.2320	.9801	.2497		.9474	.1517	.9813	.2210
100.		.8710	.1779	.8686	.3088		.8684	.2167	.8681	.3184
101.		.0000	.0000	.0000	.0000		.0000	.0000	.0000	.0000
102.		.7742	.0932	.7318	.2107		.7632	.1809	.7300	.2176
103.		.3065	.3854	.3453	.2404		.2632	.3441	.3442	.2374
104.		.8548	.1221	.8033	.2287		.7544	.1979	.8050	.2169
105.	*	.4839	.2351	.4684	.2432	**	.4649	.2260	.4691	.2321
106.		.8871	.3981	.8900	.2827		.8684	.2968	.8897	.2718
107.		.5645	.2523	.6190	.1908		.5526	.1707	.6225	.1820
108.		.9677	.3377	.9436	.3075		.9211	.2573	.9445	.2916
109.		.1613	.1447	.1857	.1703		.2281	.1517	.1860	.1749
110.		.9194	.0535	.9567	.1967		.9211	.0924	.9563	.1939
111.		.3710	.1900	.4072	.2146		.3860	.2022	.4067	.2139
112.		.6290	.6088	.8459	.4394		.6053	.5205	.8480	.4310

113.		.6774	.1692	.7345	.2794		.6754	.1657	.7356	.2827
114.		.8710	.3491	.8597	.3568		.8596	.2773	.8598	.3476
115.		.8065	.2131	.8824	.1209		.8158	.1289	.8848	.1111
116.		.5968	-.0236	.5729	.0538		.6228	-.0320	.5725	.0495
117.		.7419	.2657	.8707	.1397		.7632	.2000	.8723	.1179
118.		.5484	.2521	.6623	.1725		.5965	.2719	.6634	.1635
119.		.5161	.0009	.4862	.0263		.5263	-.0104	.4851	.0149
120.		.3710	.2514	.4966	.2272		.3947	.1680	.4983	.2281
121.	*	.6613	.3186	.6926	.3248	**	.6140	.3951	.6933	.3233
122.		.1452	.2057	.2469	.1730		.2281	.0660	.2477	.1722
123.	*	.6935	.2715	.6774	.3543	**	.6579	.3075	.6780	.3488
124.		.7419	.2418	.8267	.3937		.7368	.3659	.8265	.4122
125.		.7258	.1748	.7607	.2352	**	.7105	.2317	.7627	.2319
126.	*	.7581	.2166	.7758	.2623	**	.7807	.2069	.7765	.2552
127.		.4355	.3601	.5922	.2191		.4298	.3965	.5940	.2230
128.		.5161	.5325	.5433	.3073		.5000	.3619	.5441	.3132
129.	*	.5161	.2298	.6210	.2536		.5263	.1704	.6190	.2580
130.		.2258	.0065	.2992	.1526		.2368	.0985	.2991	.1674
131.	*	.6129	.3503	.7166	.4035		.6228	.2995	.7183	.3971
132.		.6452	.4716	.7483	.3040		.6140	.4890	.7502	.2984
133.		.6129	.1372	.5949	.3307	**	.5614	.2852	.5982	.3259
134.		.7258	.1826	.8343	.0938		.7895	.1489	.8369	.0923
135.		.6935	.1450	.6534	.1838		.6053	.1816	.6530	.1854
136.	*	.5806	.4342	.5406	.3210	**	.5439	.2974	.5399	.3349
137.		.8065	.2087	.7820	.3120	**	.7719	.3553	.7835	.3096
138.		.5161	-.2406	.5179	.1949		.4561	-.1024	.5170	.1953
139.		.6129	.0448	.6699	.2668		.6053	.1392	.6711	.2647
140.	*	.7097	.5077	.8129	.3193		.6579	.5133	.8119	.3238
141.		.6290	.2640	.7056	.1745		.6491	.1707	.7065	.1722
142.		.7903	.2901	.7785	.4315		.7719	.3640	.7779	.4368

143.	*	.5645	.3583	.5763	.3194	**	.5439	.3650	.5781	.3143
144.		.5645	-.1932	.5509	.0909		.5965	-.1812	.5531	.0844
145.		.9355	.2648	.9697	.3572		.9211	.3292	.9695	.3484
146.		.6129	.0448	.5798	.1984		.5702	.0719	.5781	.2055
147.	*	.7742	.2427	.7820	.3390	**	.7719	.3824	.7814	.3289
148.		.5968	.0151	.6953	.3077		.5614	.0950	.6960	.3064
149.		.4839	.2945	.4697	.2175		.4825	.2665	.4698	.2137
150.		.9032	.0863	.8858	.3245		.9035	.0892	.8883	.3125
151.	*	.6613	.3001	.6059	.4845		.5526	.3337	.6058	.4893
152.		.7742	.2635	.8074	.1962		.7719	.3640	.8105	.1888
153.	*	.7903	.2644	.7744	.3089	**	.6930	.3851	.7745	.3047
154.		.7419	.1762	.6534	.3518	**	.6842	.2986	.6537	.3545
155.	*	.6129	.2587	.6864	.2917		.5789	.2137	.6863	.2929
156.		.5161	.4422	.6348	.4395		.5175	.4354	.6384	.4346
157.		.6129	.2695	.6919	.1866		.5526	.1888	.6926	.1812
158.		.8226	.3142	.7840	.3625		.8596	.3346	.7835	.3583
159.		.6935	.2639	.7882	.3827	**	.7018	.3998	.7897	.3881
160.		.9355	.1027	.8783	.2692		.8684	.3208	.8786	.2659

Note. Based on item difficulty indices and item discrimination indices, * indicates that the items will be chosen as anchor items for Form A; ** indicates that the items will be chosen as anchor items for Form B.

Table 34

Item Difficulty Indices and Item Discrimination Indices by Languages (Chinese and English) and Forms (From A and Form B)

		Form A				Form B			
Items		Chinese		English		Chinese		English	
		Diff.	Disc.	Diff.	Disc.	Diff.	Disc.	Diff.	Disc.
1.	*	.7266	.3785	.7954	.3140	.9569	.1666	.8606	.5389
2.		.7578	.4373	.8110	.2059	.8621	.1536	.8715	.3502
3.	*	.7891	.3605	.8320	.3017	.9138	.3527	.8660	.6463
4.		.7344	.0545	.6653	.2085 **	.7069	.3091	.7785	.4788
5.		.5313	.2644	.5333	.2371	.6207	.1165	.5468	.1046
6.		.4766	.2679	.4301	.0894	.8362	.3641	.8079	.2601
7.		.2891	.1486	.2196	.0829	.8707	.2844	.7505	.4063
8.	*	.6875	.3966	.7618	.2994	.5776	.2378	.4983	.0708
9.		.8672	.1791	.8116	.2410	.8707	.4167	.8407	.6152
10.		.9219	.1450	.9370	.2657	.9655	.3321	.9029	.5067
11.		.5859	.1664	.6623	.1028	.5172	.0243	.5229	.2658
12.		.8125	.4299	.9124	.2805	.2155	.0439	.2331	.1434
13.		.4609	.2372	.6251	.1301 **	.7241	.3284	.6794	.4697
14.		.7188	.1686	.7684	.1856 **	.5517	.3748	.6794	.3978
15.		.0000	.0000	.0000	.0000	.8621	.3141	.8332	.3568
16.		.7734	.4018	.8566	.2666	.8276	.1941	.7949	.5312
17.		.6641	.2646	.7888	.1755	.9052	.2238	.8168	.5931
18.	*	.6641	.2879	.7271	.2670 **	.6207	.3889	.5913	.4318
19.		.8594	.1430	.8908	.1831	.7241	.2724	.7211	.3879
20.	*	.7188	.4345	.7588	.3398	.4483	.1384	.4067	.2411
21.	*	.7188	.5728	.7936	.3374	.3276	.1260	.3465	.1463
22.		.8594	.0710	.8380	.2619	.8103	.1530	.7321	.4510

23.		.4766	.1384	.5327	.2038	.6293	.0513	.6576	.2875
24.		.9297	.2542	.9202	.2156	.9655	.4489	.8906	.6882
25.	*	.5625	.2141	.5717	.2204	.4224	.1746	.4115	.0712
26.		.5234	.1350	.4811	.2847	.8362	.1634	.7826	.4506
27.		.6563	.3483	.7139	.1937 **	.6897	.4130	.7485	.3249
28.	*	.5938	.3902	.5927	.3875	.6207	.1656	.5735	.3394
29.		.3828	.4431	.3299	.3272 **	.8017	.3411	.8086	.5032
30.	*	.5234	.3742	.6149	.2801	.8621	.2403	.9009	.5334
31.	*	.7500	.4515	.7918	.2831	.5431	.2158	.6398	.2563
32.		.2891	.0674	.2795	.0242	.5000	.2545	.5407	.0114
33.		.0000	.0000	.0000	.0000	.8793	.2038	.8257	.5276
34.	*	.5391	.3604	.6257	.2834 **	.5086	.4394	.6036	.4359
35.		.6563	.2443	.6479	.2495	.7414	.1282	.6644	.3516
36.		.3828	.1691	.4997	.2004	.5172	.0849	.5981	.0682
37.	*	.7734	.3852	.7996	.3442 **	.6897	.4074	.7027	.4143
38.	*	.8125	.4016	.8488	.3017	.8966	.2756	.8510	.1983
39.		.4219	.2654	.4307	.2531	.6121	.2074	.4361	.1463
40.		.8750	.3229	.9238	.2465	.4655	.0860	.4621	.2387
41.		.3438	.2743	.4397	.3619	.5690	.2969	.4648	.3862
42.		.5391	.1225	.6383	.2186 **	.8448	.4538	.7731	.5282
43.		.8203	.0365	.7355	.0783 **	.7069	.3631	.7088	.4982
44.		.5391	.4263	.5909	.1640	.5776	.0846	.6787	.0884
45.		.4219	.3556	.4205	.2169	.8966	.3725	.8271	.4960
46.		.4219	.2613	.3545	.2608 **	.8362	.3586	.7628	.4539
47.	*	.7031	.3826	.8086	.3586 **	.7586	.5003	.7478	.5346
48.		.5938	.2139	.6605	.1918	.3190	.2756	.3336	.2177
49.		.6406	.3791	.5705	.2530 **	.5690	.3041	.5352	.3521
50.		.8125	.2856	.8104	.2716	.7155	.2761	.6391	.4425
51.		.4141	.0409	.4151	.1038 **	.5603	.2857	.5598	.3813
52.	*	.7109	.3080	.7726	.3077	.3017	.2107	.2761	.2118

53.	*	.6406	.2123	.7223	.2269	.7328	.3253	.6822	.4112
54.		.5625	.2445	.5591	.0751	.1983	-.0076	.2659	-.1054
55.		.0000	.0000	.0000	.0000	.7328	.2469	.6808	.2686
56.		.8281	.2767	.8704	.2491 **	.7931	.2405	.8079	.5465
57.		.5234	.2946	.4553	.1118 **	.7931	.4551	.8120	.5708
58.		.6797	.2363	.7349	.3022	.5690	.1394	.4880	.2764
59.	*	.4531	.3471	.5069	.3059	.8103	.3568	.8421	.3065
60.		.2500	.0521	.2567	.0163	.3793	.2017	.4641	.2253
61.		.0000	.0000	.0000	.0000 **	.4052	.4810	.5830	.4163
62.	*	.6016	.3612	.6959	.3509 **	.8103	.3594	.7341	.5026
63.	*	.6328	.2637	.6593	.2486	.8793	.3036	.9057	.1948
64.	*	.7109	.4318	.7493	.2726 **	.7155	.3862	.7239	.3158
65.		.4688	-.0912	.4961	.0380	.8017	.2656	.8565	.5239
66.		.4688	.2603	.4091	.1424	.4310	.3007	.4381	.1475
67.	*	.5547	.4635	.5417	.2731	.3793	.1118	.4498	-.0102
68.	*	.5078	.3924	.4199	.3245	.9741	.3370	.9132	.5472
69.		.6406	.1083	.6773	.1905	.1897	.0611	.2543	.0321
70.		.4688	.1608	.2753	.0412	.0000	.0000	.0000	.0000
71.		.0000	.0000	.0000	.0000	.6638	.3521	.6254	.2210
72.	*	.5234	.3167	.5261	.3047	.4655	-.0181	.4935	.1265
73.		.6953	.2414	.7175	.1272 **	.8879	.3988	.8681	.6235
74.		.4688	.3537	.5699	.2384	.2069	.0216	.3158	.0716
75.		.4219	.3482	.4439	.2012	.4052	.1822	.4074	.3251
76.		.8047	.4107	.7534	.2640	.4310	.0880	.4087	.2629
77.	*	.6172	.5465	.6353	.3600 **	.6121	.2681	.5591	.2703
78.	*	.6719	.3819	.7672	.3229	.3707	.3192	.3438	.1982
79.		.6250	.3141	.7067	.1812	.7931	.2782	.7061	.2871
80.	*	.5859	.3876	.5549	.4344	.6293	.1677	.7177	.4158
81.		.0000	.0000	.0000	.0000	.3707	.0675	.3254	.1920
82.		.0000	.0000	.0000	.0000 **	.5948	.4771	.7259	.4472

83.	.7344	.1644	.8092	.1482	.3707	.3638	.4696	.1476
84.	.0000	.0000	.0000	.0000	.5776	.1955	.4853	.3084
85.	.4609	.2896	.4559	.2301	.1897	.1217	.2112	.1985
86.	.3516	.4904	.4547	.3881	.3103	.1713	.3971	.2217
87. *	.7266	.4116	.6713	.3579	.7155	.1139	.7286	.4382
88.	.4922	.1417	.5021	.1892	.1466	.1209	.1572	.0439
89. *	.7656	.3294	.7409	.2809	.8621	.2418	.7457	.4245
90.	.7188	.2998	.7445	.2020	.7845	.2833	.6876	.3956
91.	.7578	.3933	.7427	.2304 **	.7759	.3288	.6678	.4345
92.	.5547	.2213	.4667	.0699 **	.7500	.3006	.6494	.3884
93. *	.6172	.3517	.7205	.3619	.6034	.0540	.6152	.2758
94.	.8984	.4523	.9616	.2951	.5603	.0859	.5414	.2595
95.	.4844	.2022	.5831	.2333 **	.8103	.3320	.7949	.3186
96.	.4766	.0276	.5339	.0926	.2845	-.1001	.2215	-.0134
97.	.0000	.0000	.0000	.0000	.8276	.1793	.8209	.4629
98. *	.5391	.3489	.4865	.3663	.5862	.4157	.7697	.1773
99.	.6797	.3966	.6593	.2387	.8362	.0160	.8148	.2381
100.	.4844	.0595	.5183	.0964	.3707	.2599	.4545	.2111
101.	.3359	.1628	.3473	.0934	.5259	.1682	.4921	.1636
102.	.3516	.4060	.3863	.3228 **	.4655	.4539	.5878	.5292
103.	.7734	.3247	.8332	.1885	.6810	.1901	.6883	.2279
104.	.0000	.0000	.0000	.0000	.6638	.1599	.6528	.3357
105.	.8516	.4357	.8896	.2401	.6207	.0020	.5386	.1712
106.	.4063	.1593	.4793	.1648	.5086	.2025	.4757	.2521
107.	.7109	.0690	.6737	.1478	.4655	.1185	.4730	.1515
108.	.3906	.2814	.3779	.1629	.2586	.0192	.3390	.1702
109.	.2500	.2645	.3479	.2455 **	.7414	.3434	.7239	.4379
110. *	.6172	.4599	.6047	.3470	.1810	.0761	.2064	.1275
111.	.4766	.0671	.4523	.2141 **	.7069	.3755	.7779	.3593
112.	.8594	.2363	.8380	.2545 **	.8190	.4320	.7847	.5437

113.	.5313	.2219	.5807	.1934	.5776	.2864	.6008	.1485
114.	.7656	.3795	.9190	.2350	.6034	.0375	.4839	.1411
115.	.7031	.3423	.8836	.1307	.7241	.0815	.6644	.1932
116.	.4375	.2922	.3899	.2370	.5517	.0775	.4682	.1884
117.	.2500	.1249	.3041	.1225	.4483	-.1549	.4826	.1899
118.	.3906	.3325	.3893	.2964	.4138	.3975	.3896	.1619
119.	.3359	.4447	.2292	.2919	.2586	.3809	.1818	.2935
120. *	.6172	.5993	.6539	.4594	.4224	.0250	.4293	.1464
121.	.3203	.2130	.3209	.0623	.6207	.2064	.4395	.3929
122. *	.7266	.3583	.7049	.3272	.2759	.3480	.2160	.2558
123.	.8516	.3668	.9424	.2302	.8362	.4582	.9412	.4156
124.	.0000	.0000	.0000	.0000	.8966	.2673	.8804	.4251
125.	.5703	-.0809	.4283	-.0618 **	.7069	.3350	.7642	.4676
126.	.4453	.2395	.4229	.1006 **	.6466	.4280	.6487	.3755
127.	.8750	.1696	.8818	.2561	.3966	.2612	.5318	.2380
128.	.5156	.1847	.6269	.2097	.5948	.2802	.4874	.2610
129.	.8594	.3392	.8566	.3422 **	.7500	.4474	.8011	.5579
130.	.9063	.1115	.8974	.2906	.8448	.1023	.7601	.4343
131. *	.7109	.2667	.6509	.2877	.8793	.0796	.8407	.4668
132. *	.5156	.4087	.5813	.3962	.8362	.4028	.8455	.5895
133.	.4063	.2387	.3893	.2289 **	.6897	.2944	.7430	.4951
134.	.5313	.1064	.4427	.1276	.5948	.1412	.4272	.2329
135.	.7109	.1073	.8188	.1408	.7069	.2429	.6671	.4290
136.	.8359	.0221	.8170	.0992	.6638	.3619	.4990	.3455
137.	.8203	.0638	.7744	.1930 **	.7241	.3399	.6576	.3851
138.	.7500	.0912	.7313	.2075 **	.7069	.3361	.6288	.3178
139.	.5625	.4650	.4883	.2560	.7672	.1888	.6671	.3892
140.	.4531	.2306	.5531	.2385 **	.5431	.2598	.5571	.3333
141. *	.7656	.3217	.8080	.3434	.7500	.2241	.7820	.4967
142.	.5391	.2244	.5321	.2177	.7500	.2017	.7047	.4097

143.	.8125	.0294	.8254	.1657		.7069	.1803	.6569	.2814
144.	.0000	.0000	.0000	.0000		.8966	.1259	.8339	.4958
145. *	.5469	.3220	.5345	.2713		.8879	.0291	.8059	.3558
146. *	.5859	.4218	.5609	.3309		.4828	.0787	.4880	.2984
147.	.3984	.2841	.4385	.2269		.0000	.0000	.0000	.0000
148.	.7344	.2490	.6959	.2878	**	.5948	.3826	.6391	.4490
149.	.6719	.2054	.7079	.1580		.6638	.2860	.6630	.4011
150.	.6641	.1819	.7157	.0041		.5948	.0474	.5564	.1736
151.	.9609	.0268	.9316	.1895	**	.8190	.3164	.8059	.4836
152.	.8828	.0894	.8170	.0910		.6379	.3793	.6794	.2040
153.	.7734	.3969	.8410	.2459		.6379	.1076	.6104	.2103
154.	.8438	.3038	.8440	.2001		.9310	.0028	.8489	.1843
155.	.4141	.2500	.3221	.2934		.4224	.2923	.3575	.2457
156.	.8047	.1990	.8134	.2145		.7759	.1470	.7204	.3935
157.	.6641	.0996	.6905	.1129		.2759	.0849	.3042	.1239
158.	.9688	.1844	.9190	.2637		.6552	.1212	.8025	.1501
159.	.6172	.3627	.7469	.1870		.6983	.2627	.6370	.2340
160.	.4297	.3378	.4385	.1196		.5431	.0128	.4730	.2268

Note. Based on item difficulty indices and item discrimination indices, * indicates that the items will be chosen as anchor items for Form A; ** indicates that the items will be chosen as anchor items for Form B.

APPENDIX D

Delta Plots for All Target Language Groups

Figure 27

Korean Language: Delta Plot for All Content Specifications in Form A

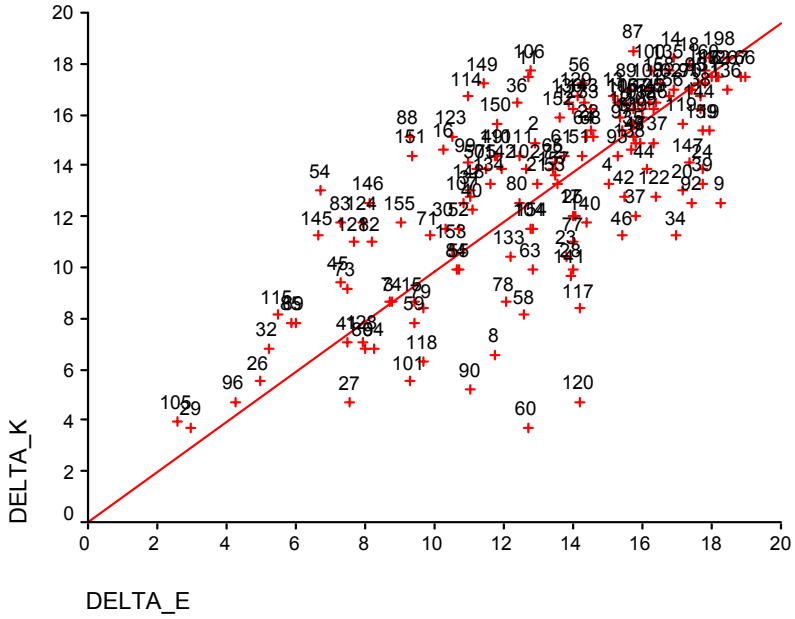


Figure 28

Korean Language: Delta Plot for Content Specification One in Form A

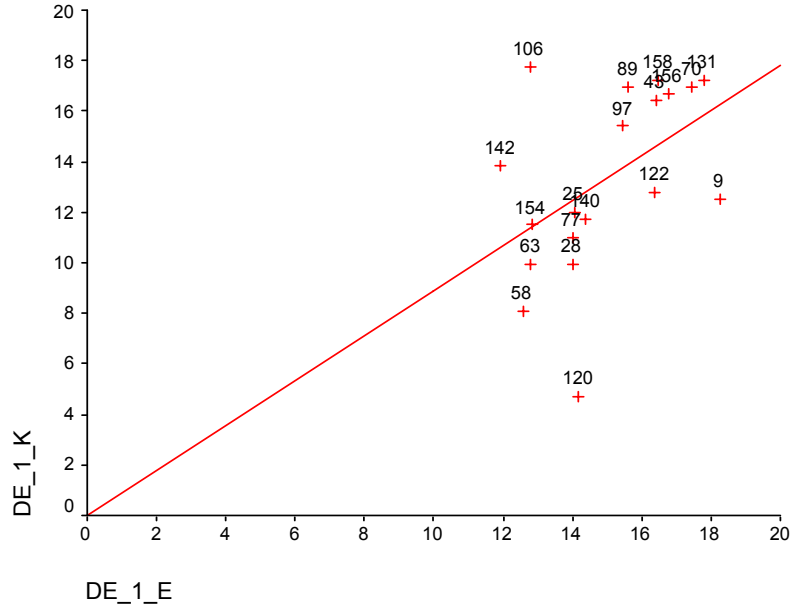


Figure 29

Korean Language: Delta Plot for Content Specification Two in Form A

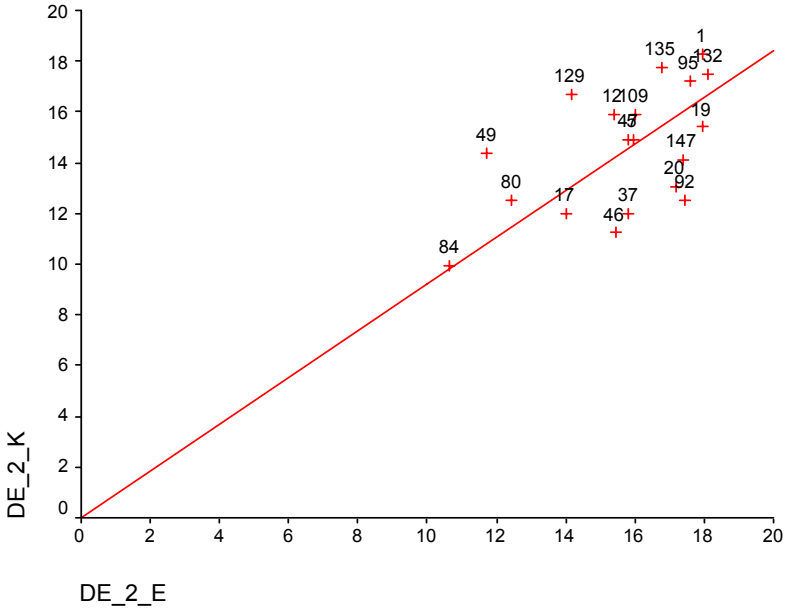


Figure 30

Korean Language: Delta Plot for Content Specification Three in Form A

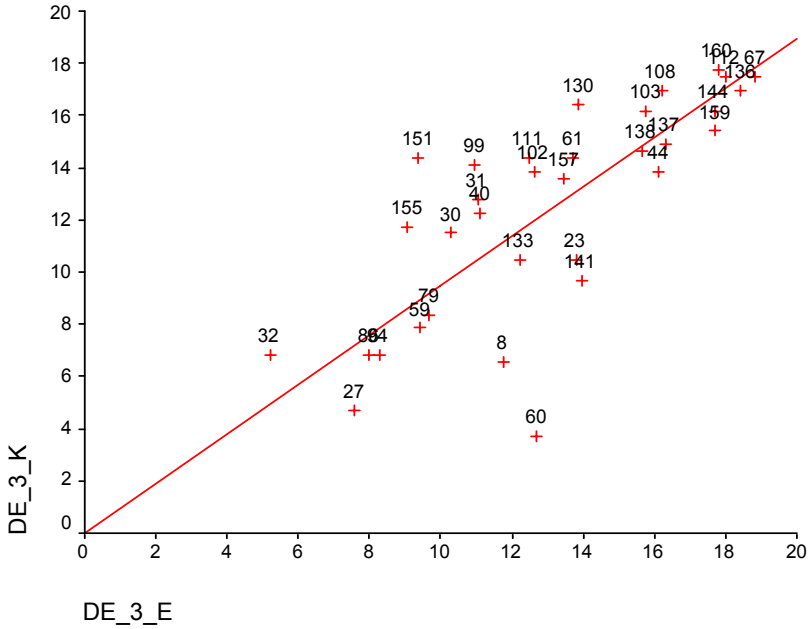


Figure 31

Korean Language: Delta Plot for Content Specification Four in Form A

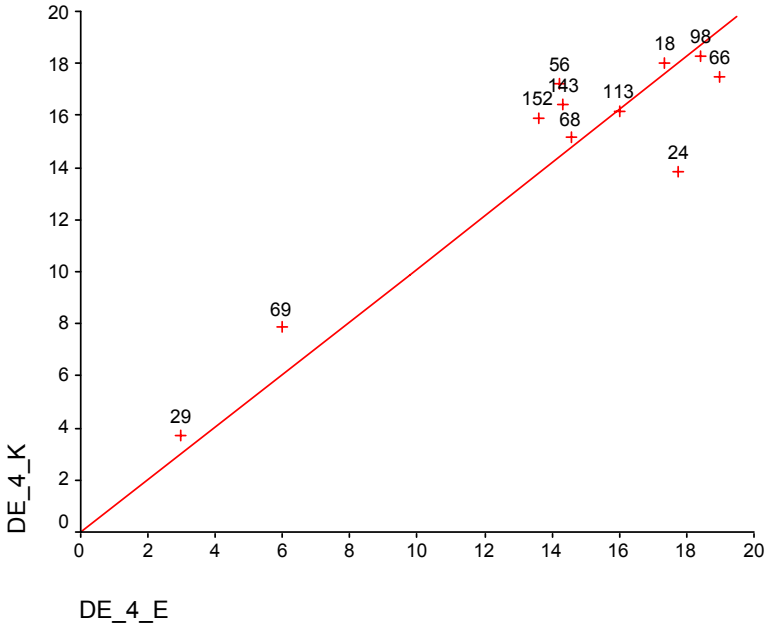


Figure 32

Korean Language: Delta Plot for Content Specification Five in Form A

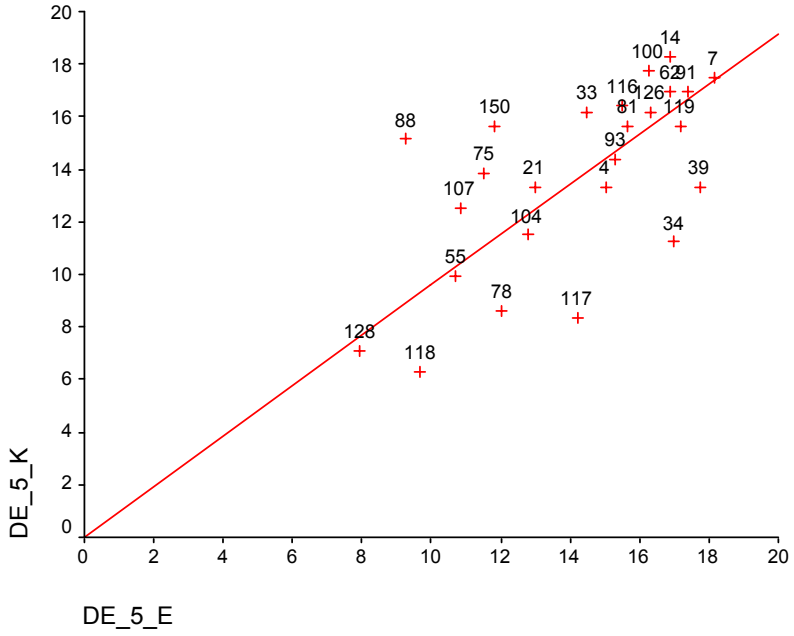


Figure 33

Korean Language: Delta Plot for Content Specification Six in Form A

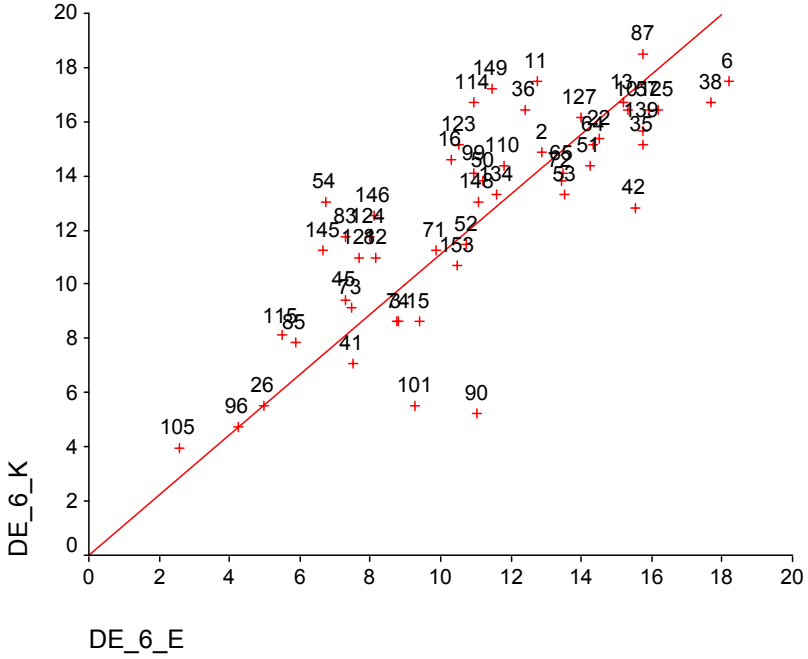


Figure 34

Korean Language: Delta Plot for All Content Specifications in Form B

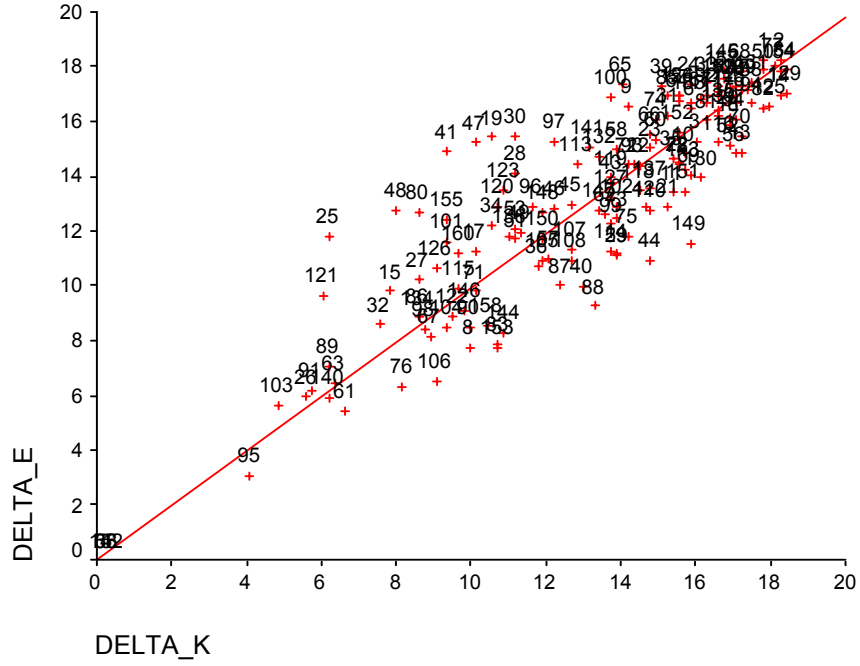


Figure 35

Korean Language: Delta Plot for Content Specification One in Form B

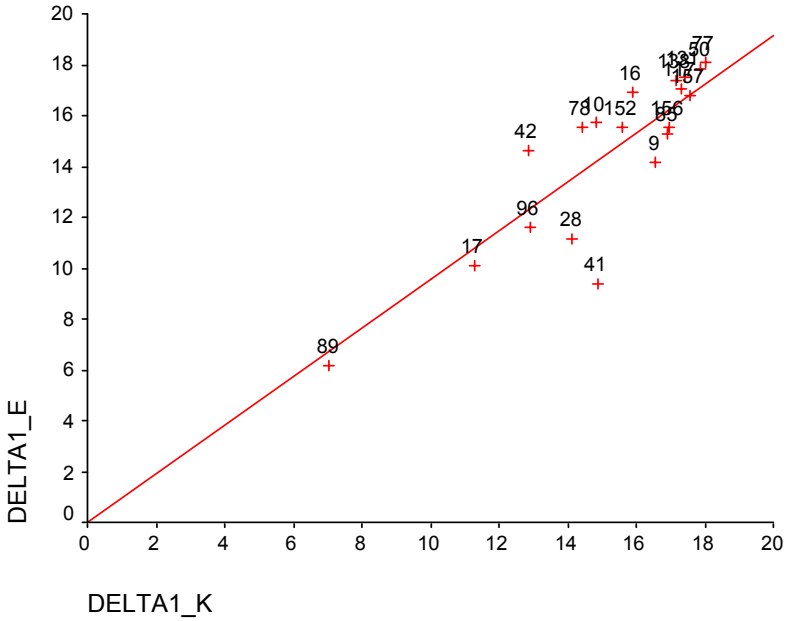


Figure 36

Korean Language: Delta Plot for Content Specification Two in Form B

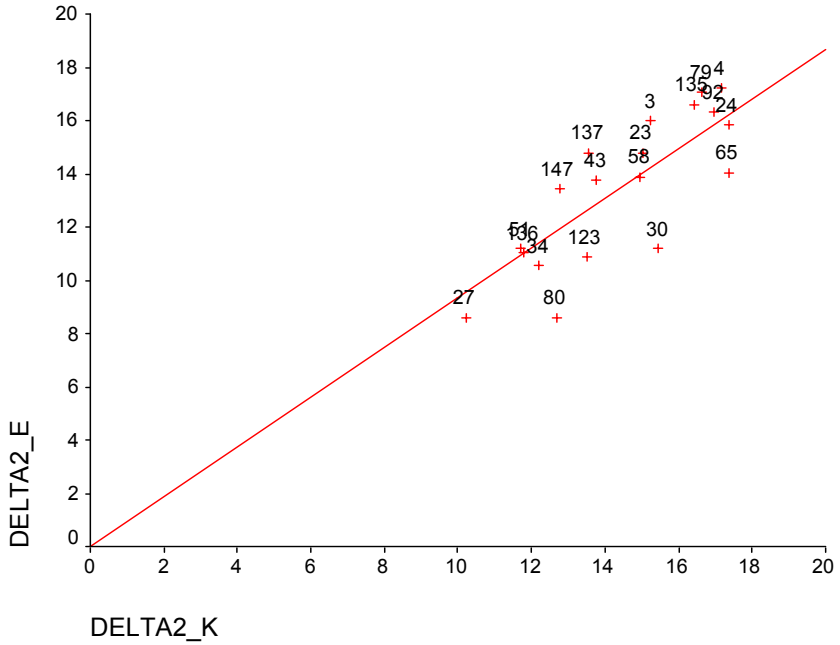


Figure 37

Korean Language: Delta Plot for Content Specification Three in Form B

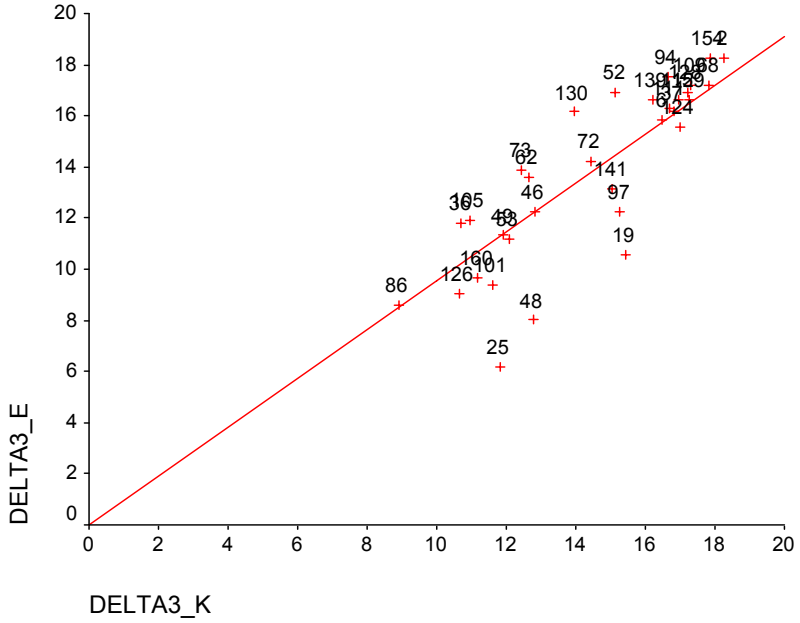


Figure 38

Korean Language: Delta Plot for Content Specification Four in Form B

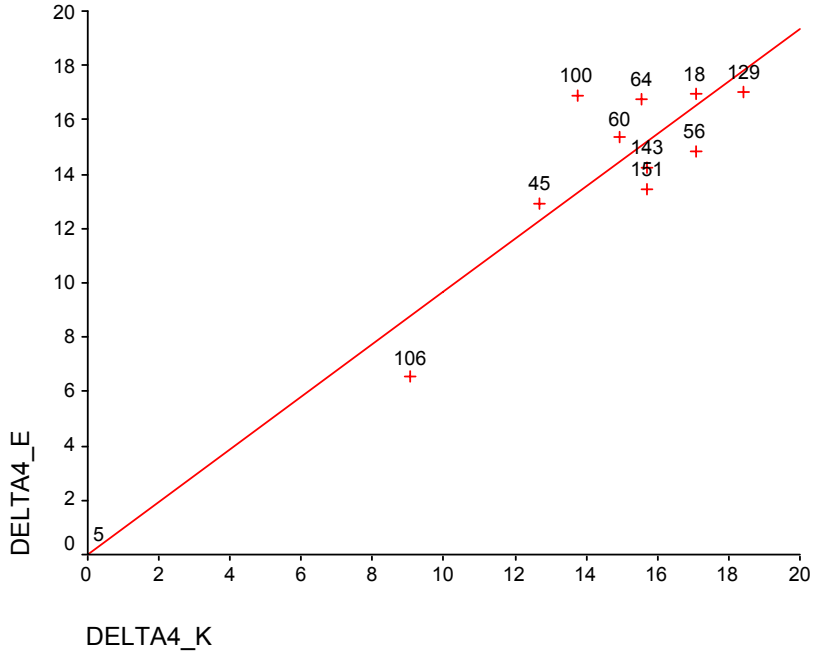


Figure 39

Korean Language: Delta Plot for Content Specification Five in Form B

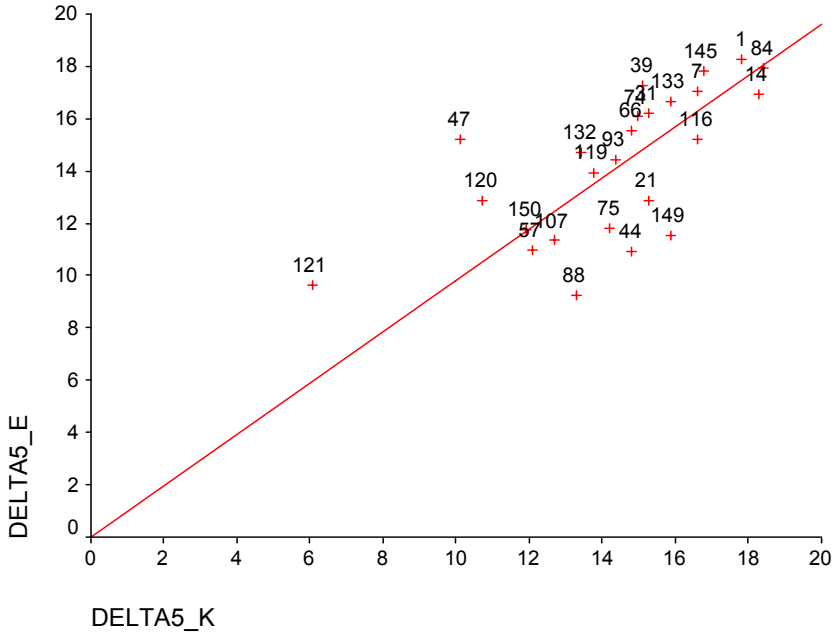


Figure 40

Korean Language: Delta Plot for Content Specification Six in Form B

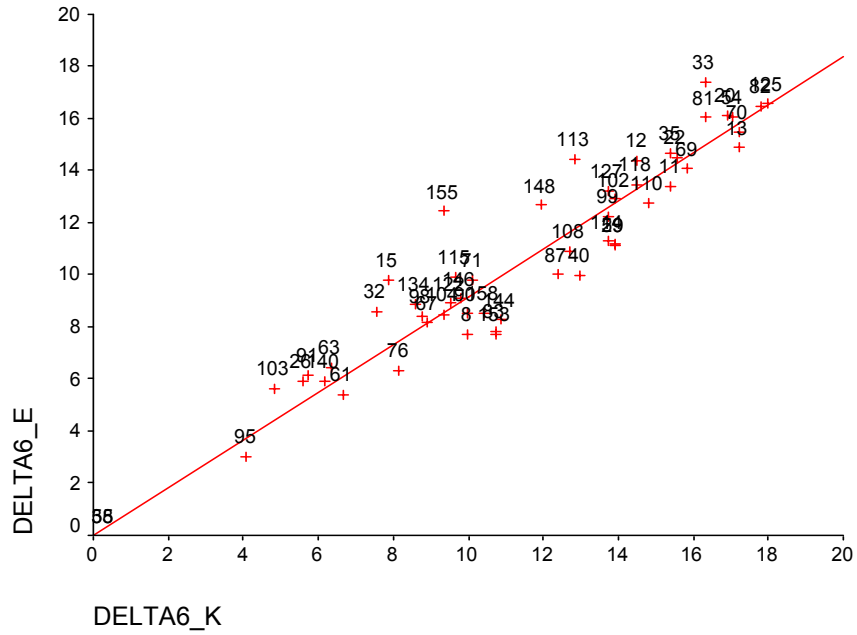


Figure 41

Spanish Language: Delta Plot for All Content Specifications in Form A

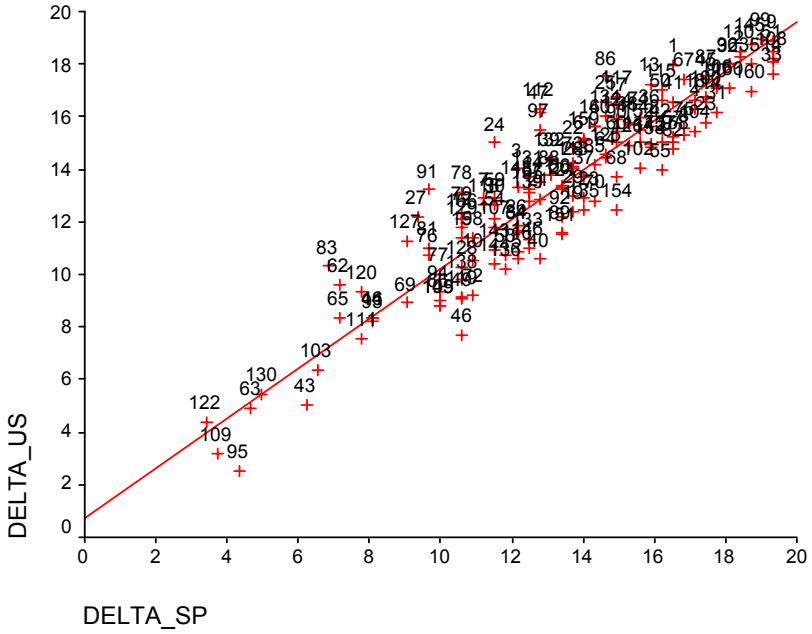


Figure 42

Spanish Language: Delta Plot for Content Specification One in Form A

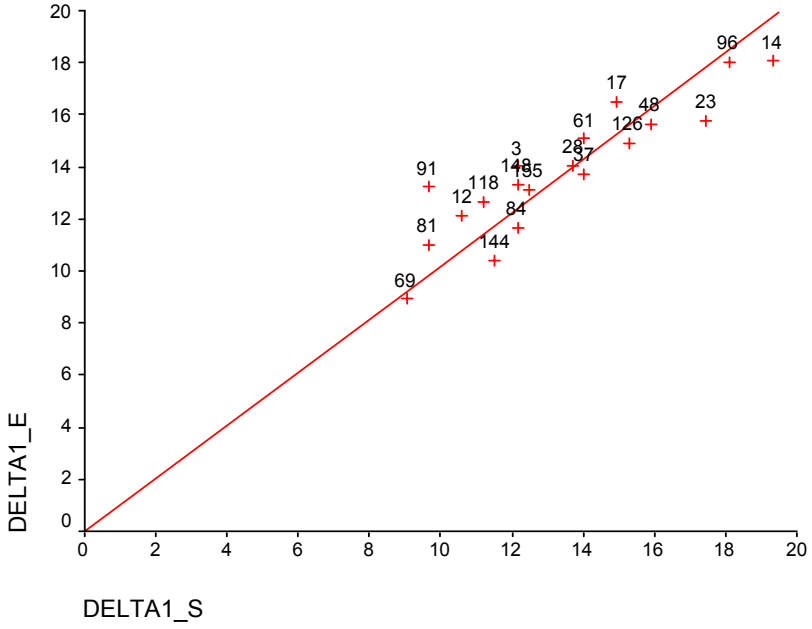


Figure 43

Spanish Language: Delta Plot for Content Specification Two in Form A

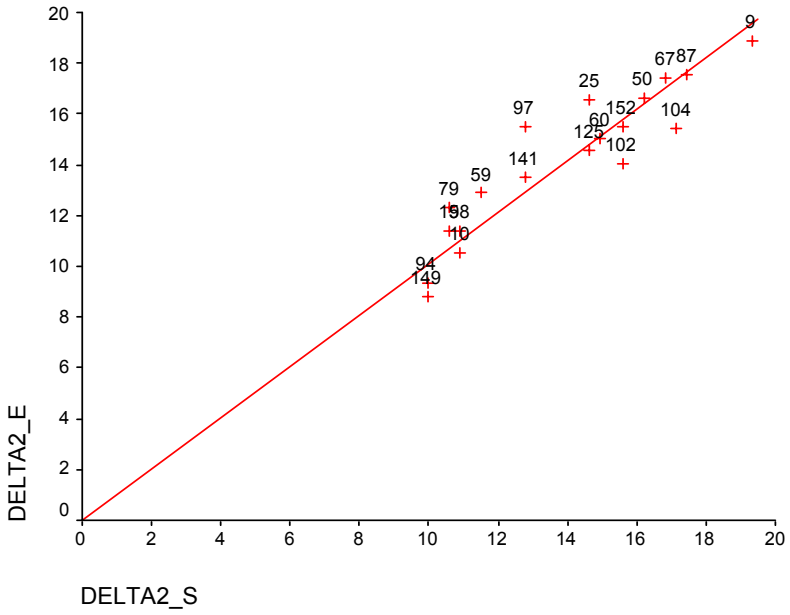


Figure 44

Spanish Language: Delta Plot for Content Specification Three in Form A

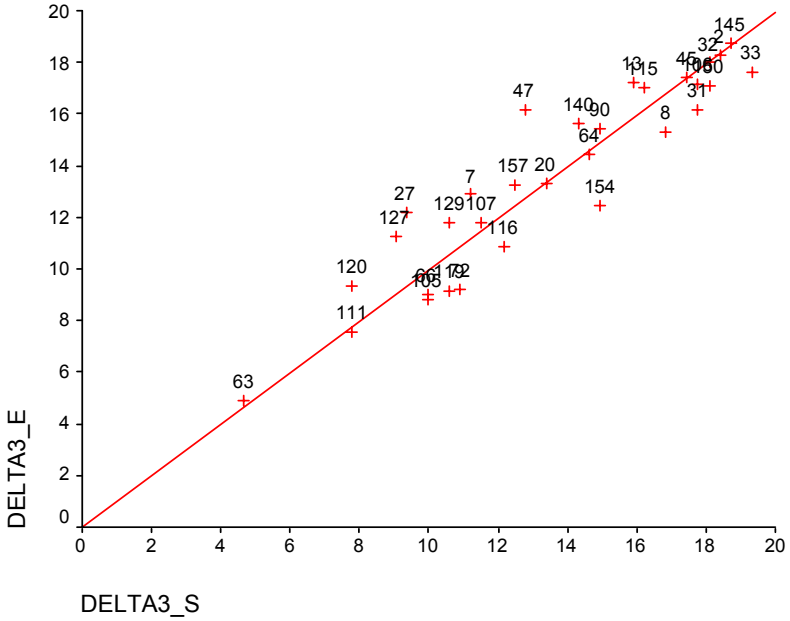


Figure 45

Spanish Language: Delta Plot for Content Specification Four in Form A

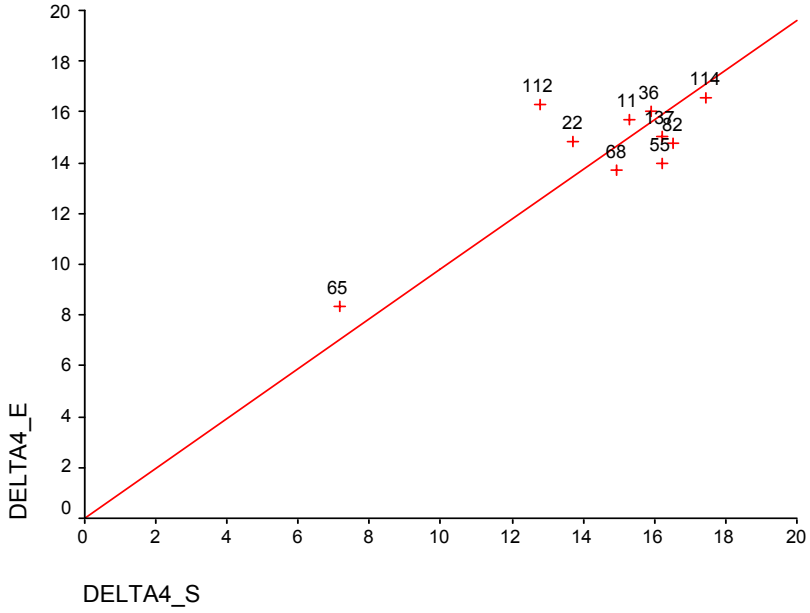


Figure 46

Spanish Language: Delta Plot for Content Specification Five in Form A

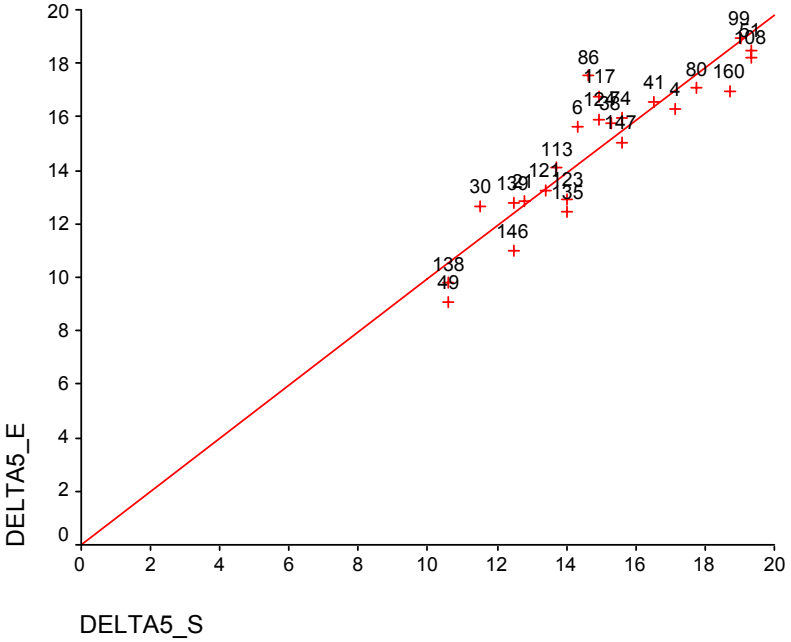


Figure 47

Spanish Language: Delta Plot for Content Specification Six in Form A

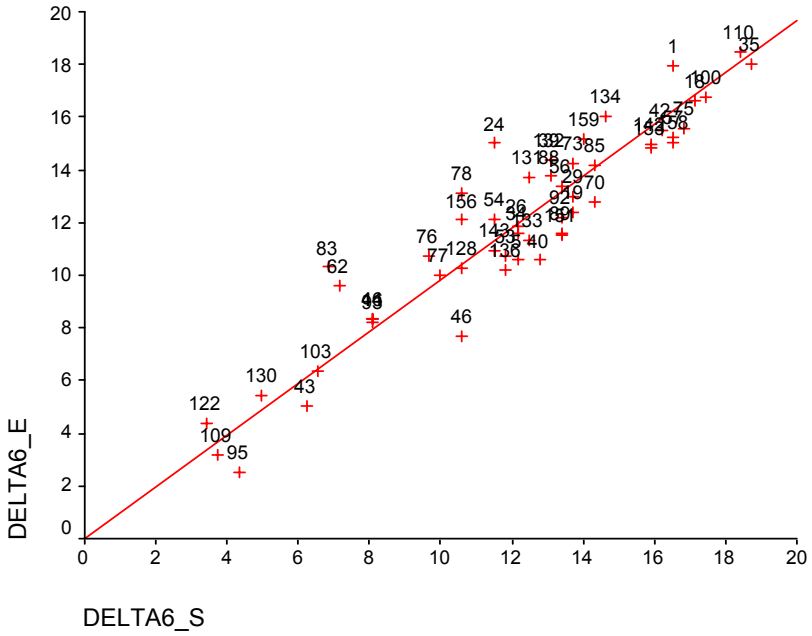


Figure 48

Spanish Language: Delta Plot for All Content Specifications in Form B

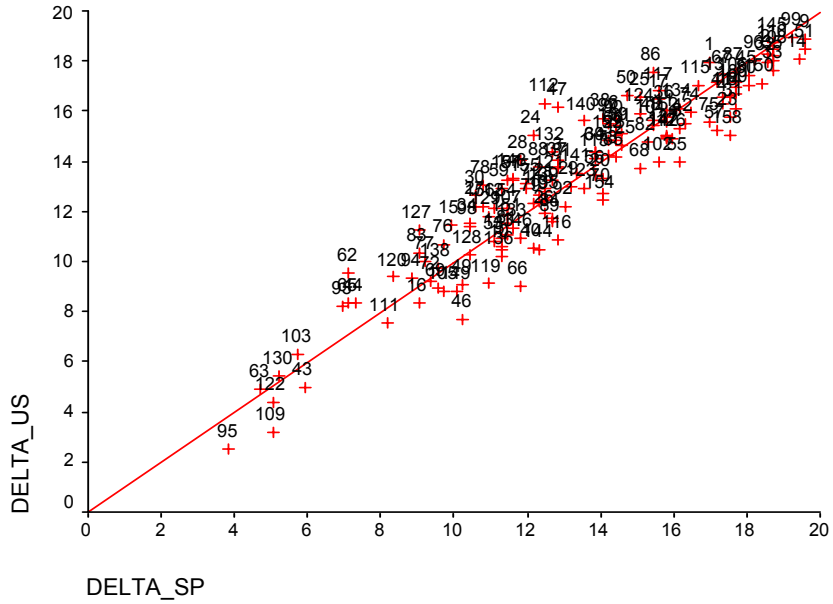


Figure 49

Spanish Language: Delta Plot for Content Specification One in Form B

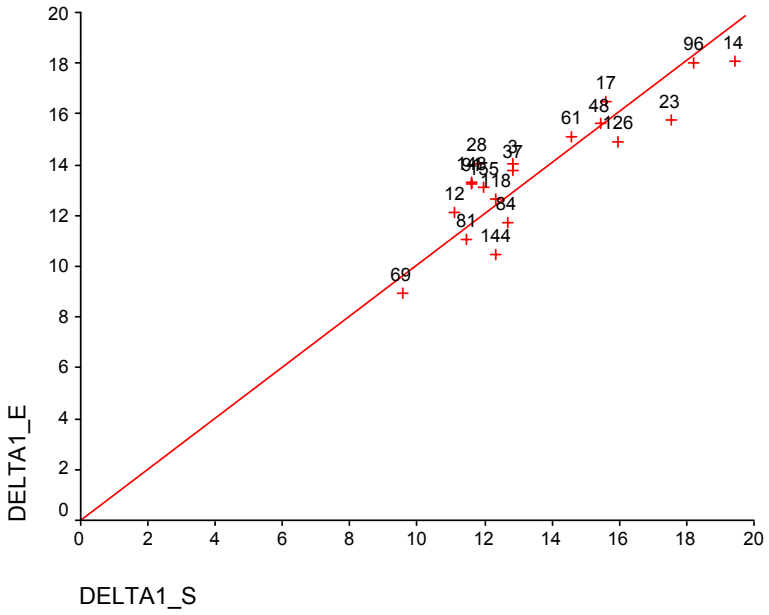


Figure 50

Spanish Language: Delta Plot for Content Specification Two in Form B

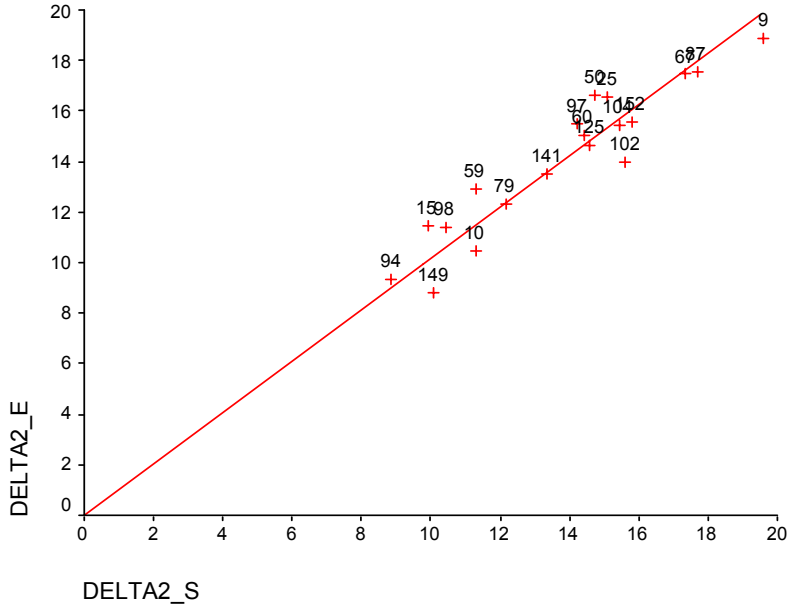


Figure 51

Spanish Language: Delta Plot for Content Specification Three in Form B

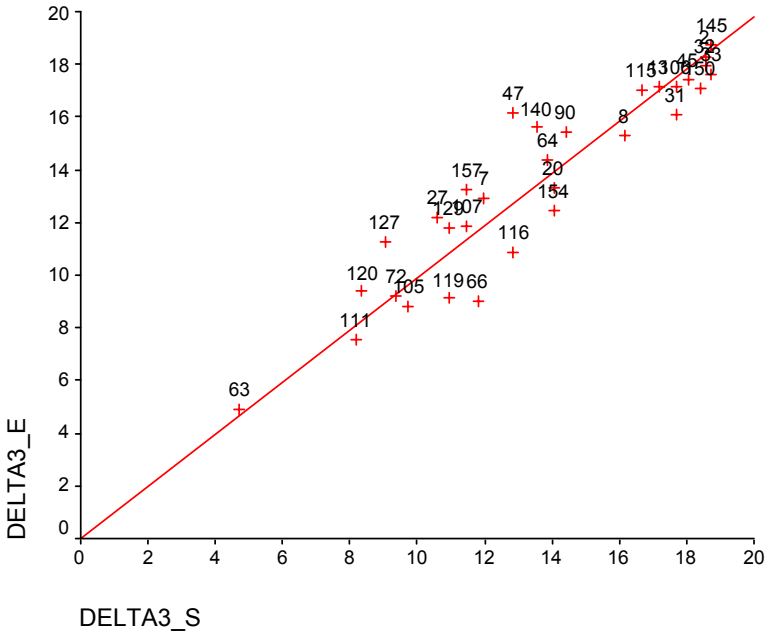


Figure 52

Spanish Language: Delta Plot for Content Specification Four in Form B

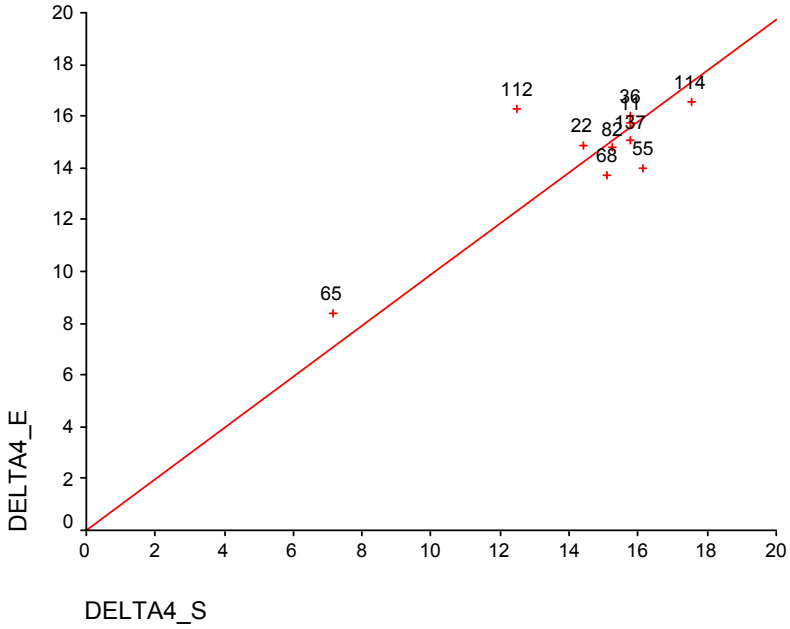


Figure 53

Spanish Language: Delta Plot for Content Specification Five in Form B

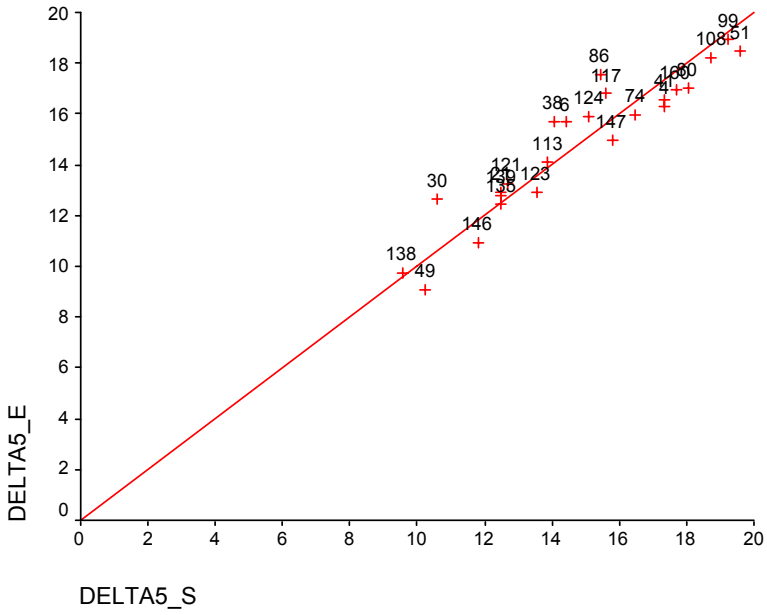


Figure 54

Spanish Language: Delta Plot for Content Specification Six in Form B

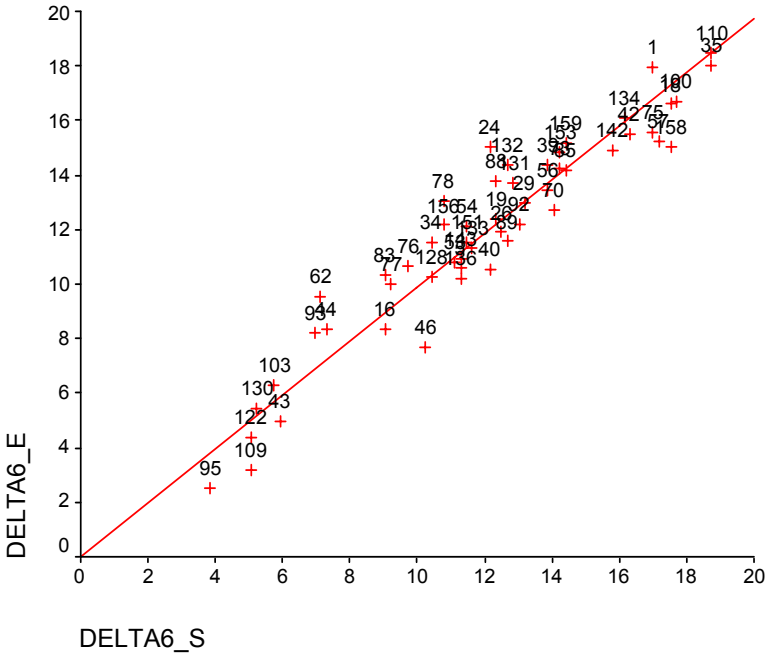


Figure 55

Chinese Language: Delta Plot for All Content Specifications in Form A

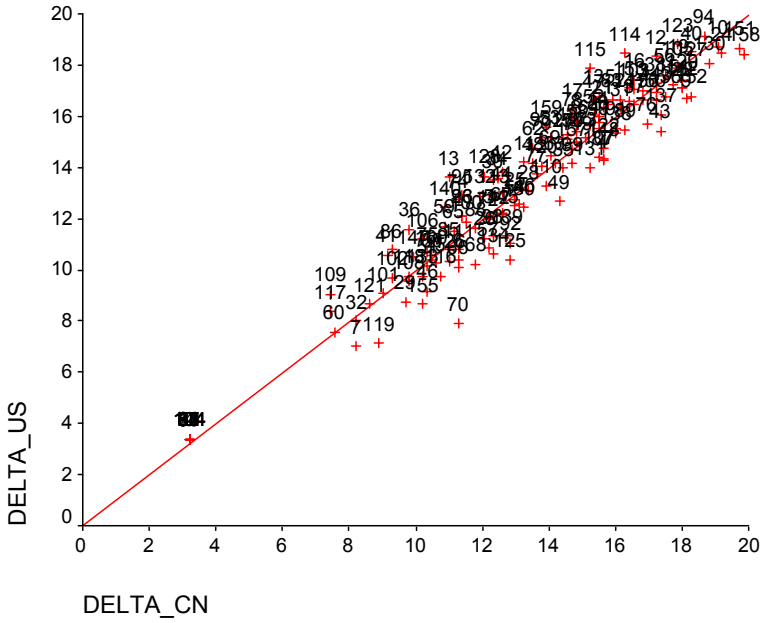


Figure 56

Chinese Language: Delta Plot for Content Specification One in Form A

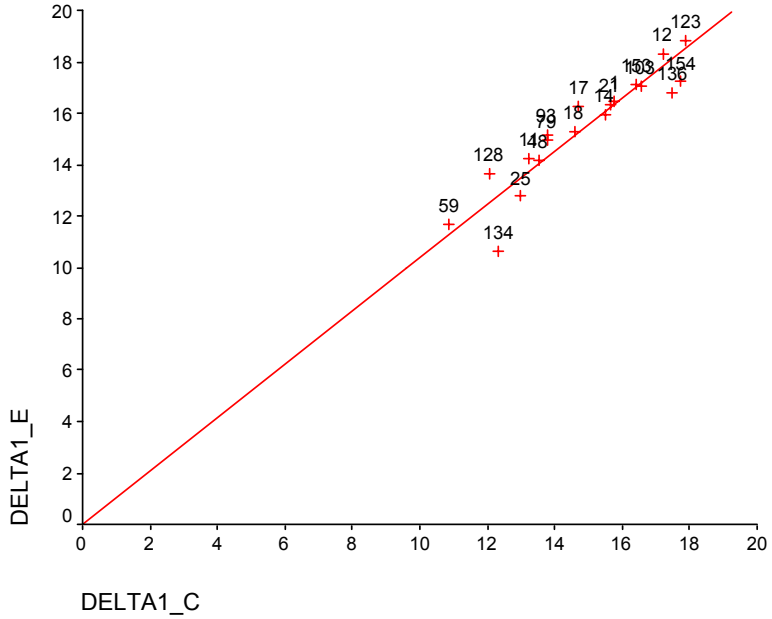


Figure 57

Chinese Language: Delta Plot for Content Specification Two in Form A

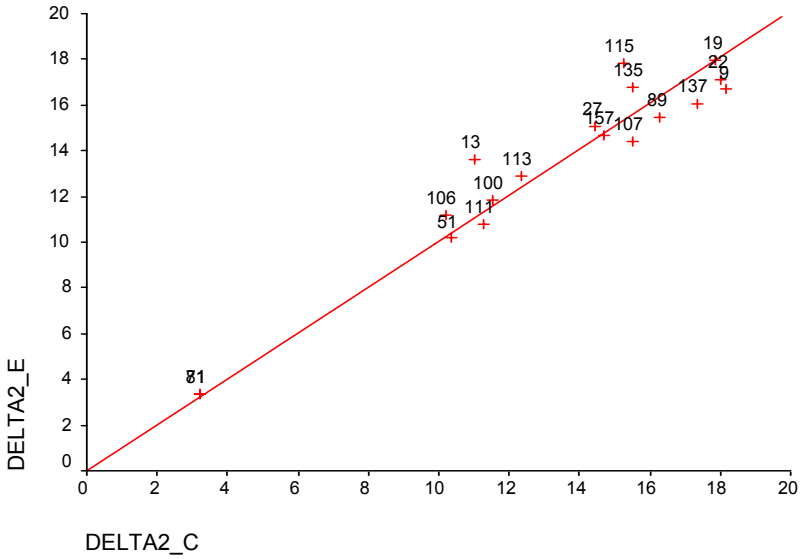


Figure 58

Chinese Language: Delta Plot for Content Specification Three in Form A

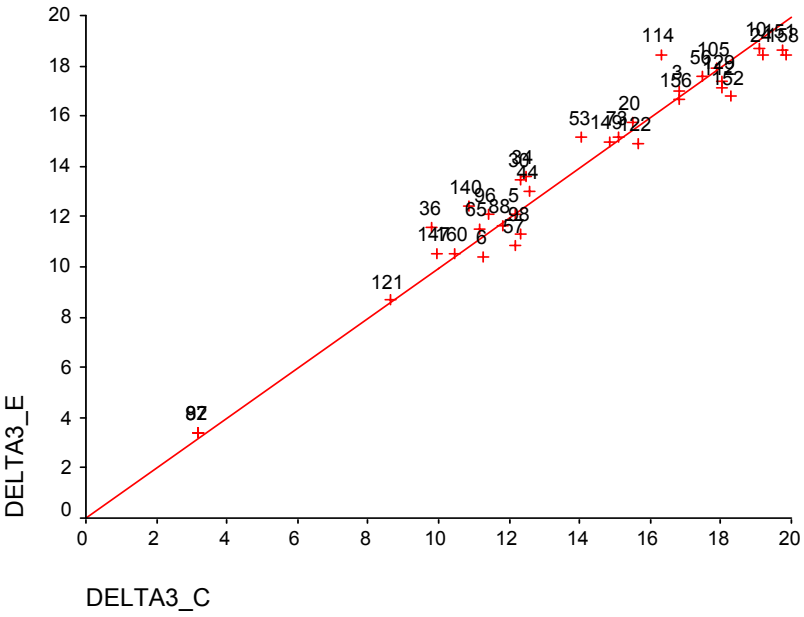


Figure 59

Chinese Language: Delta Plot for Content Specification Four in Form A

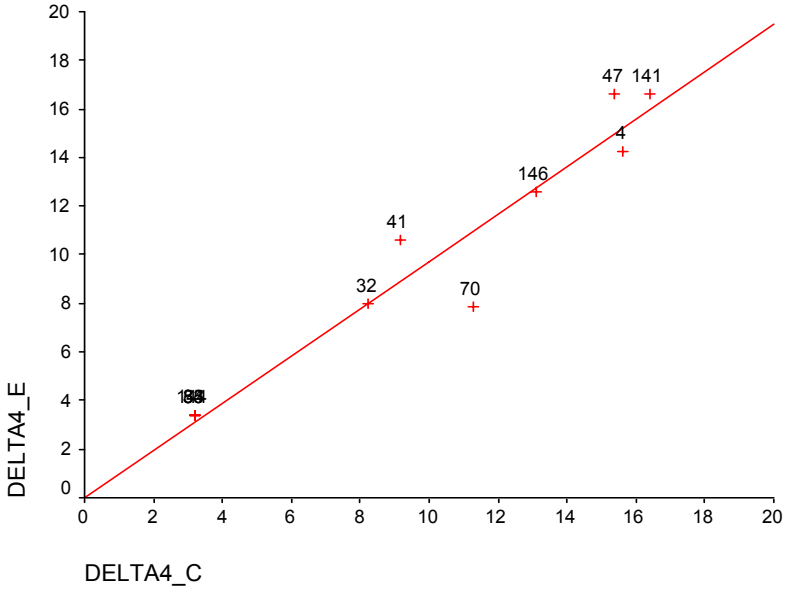


Figure 60

Chinese Language: Delta Plot for Content Specification Five in Form A

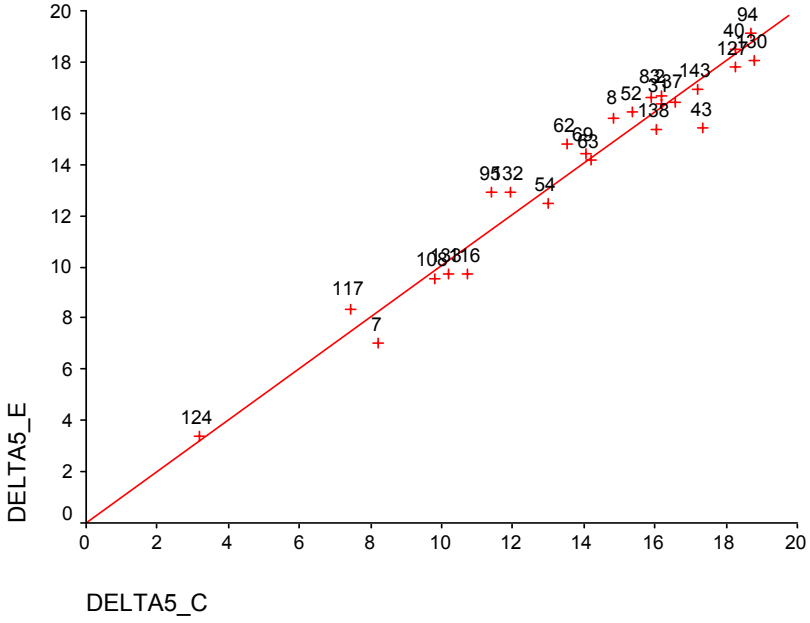


Figure 61

Chinese Language: Delta Plot for Content Specification Six in Form A

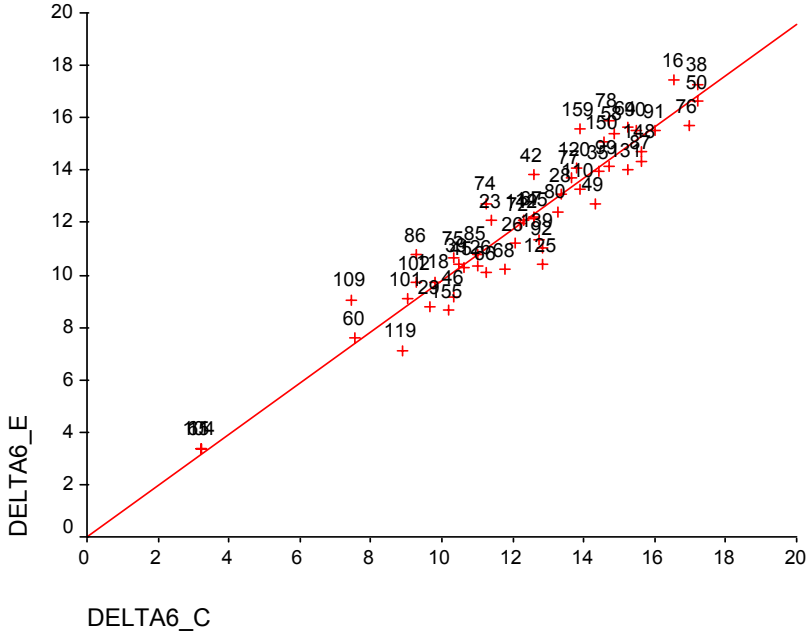


Figure 62

Chinese Language: Delta Plot for All Content Specifications in Form B

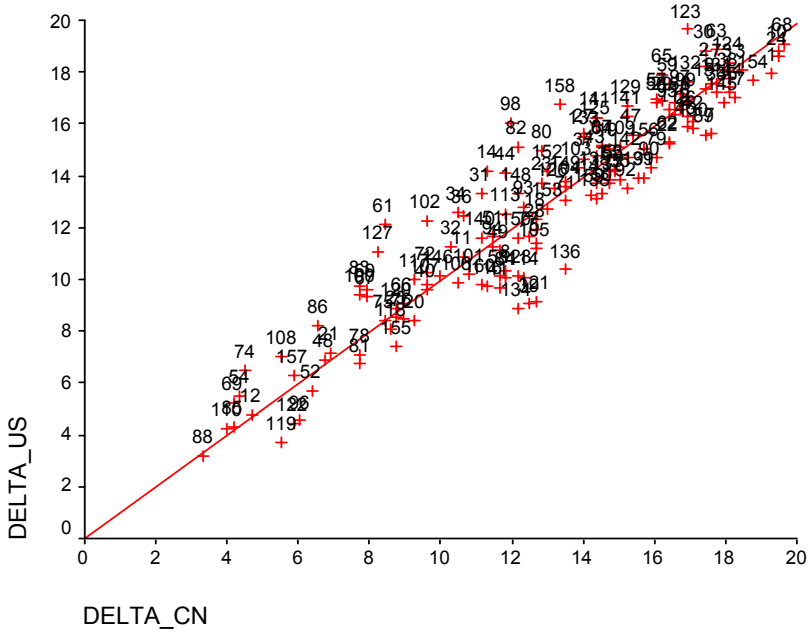


Figure 63

Chinese Language: Delta Plot for Content Specification One in Form B

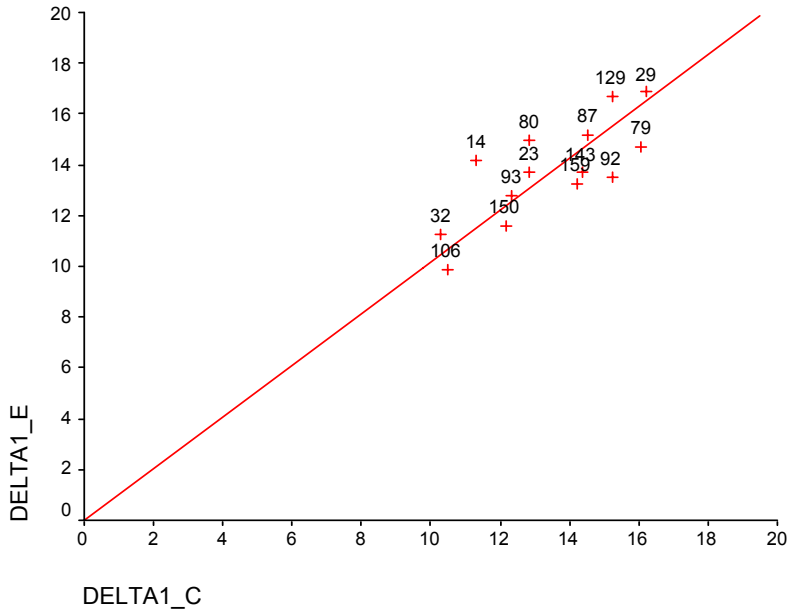


Figure 64

Chinese Language: Delta Plot for Content Specification Two in Form B

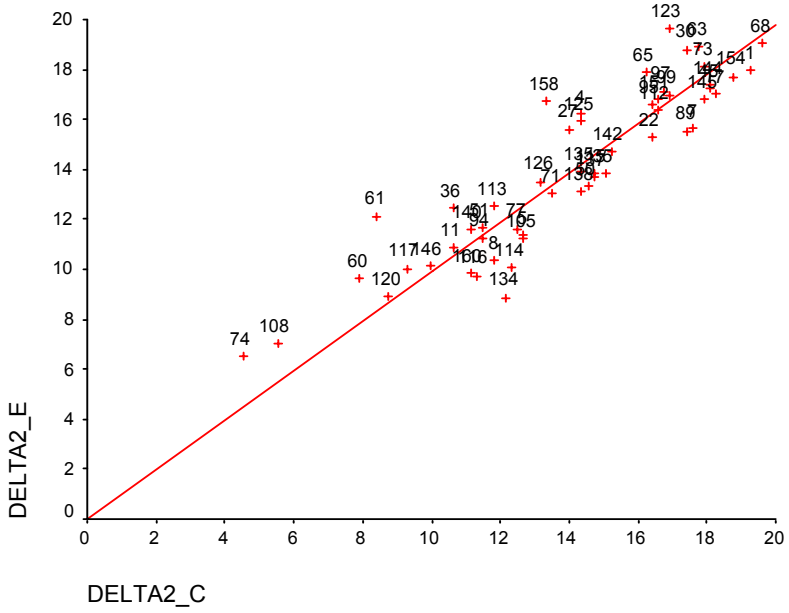


Figure 65

Chinese Language: Delta Plot for Content Specification Three in Form B

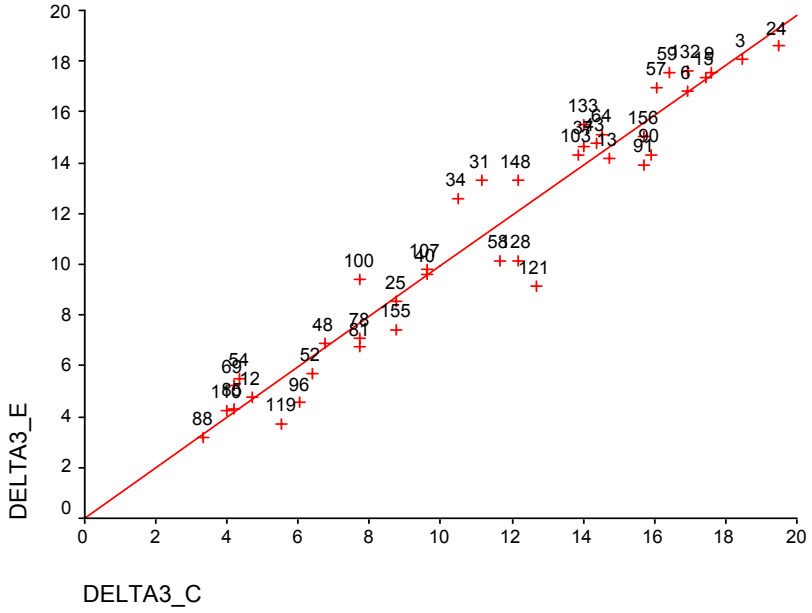


Figure 66

Chinese Language: Delta Plot for Content Specification Four in Form B

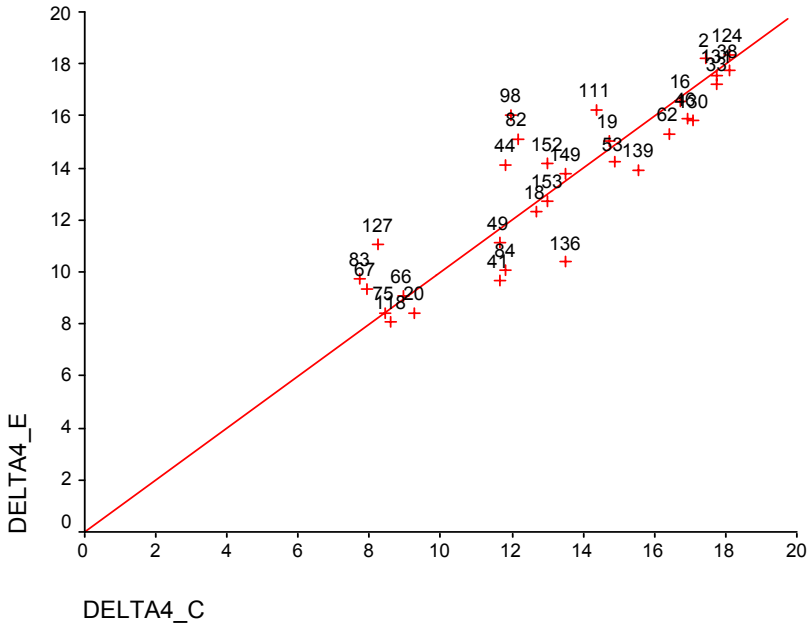


Figure 67

Chinese Language: Delta Plot for Content Specification Five in Form B

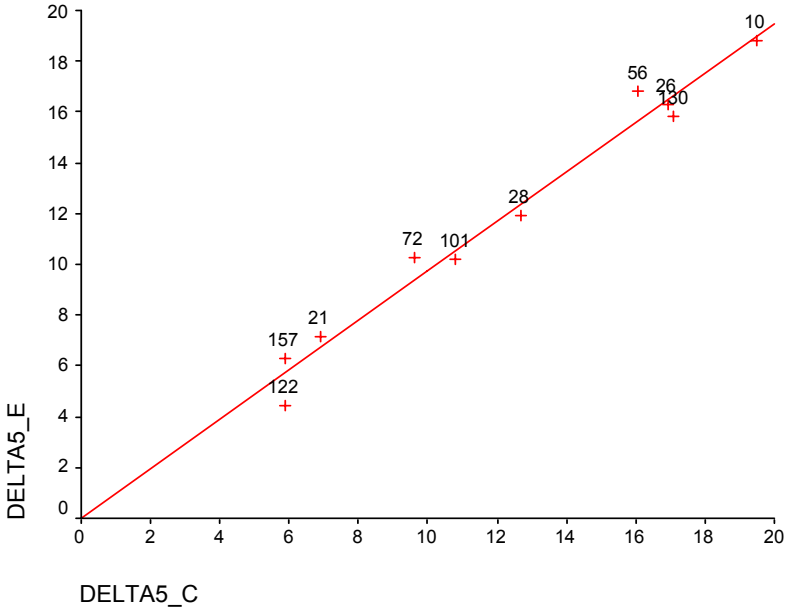
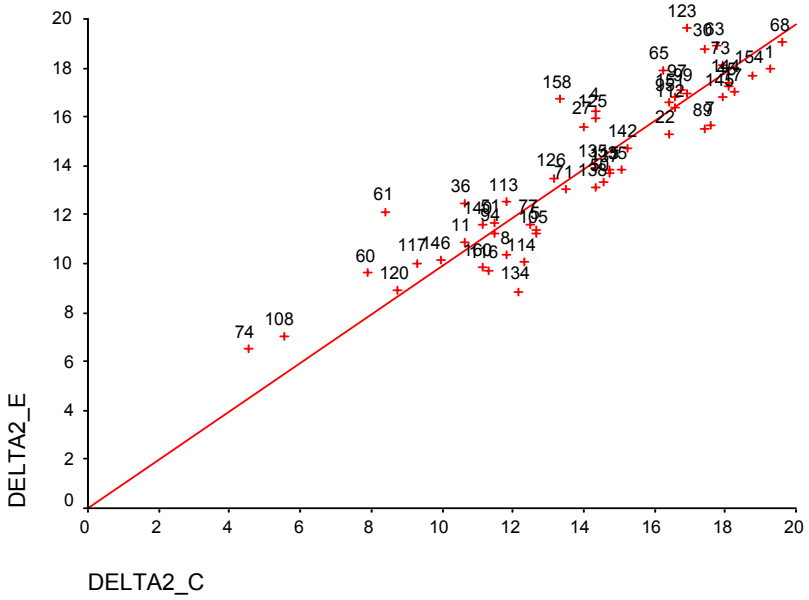


Figure 68

Chinese Language: Delta Plot for Content Specification Six in Form B



APPENDIX E

Items That Chosen as Anchor Items

for Two Anchor Tests

Table 35

Korean Language: Items That Chosen as Anchor Items for Two Anchor Tests

	Form A		Form B	
Content Area	Anchor One	Anchor Two	Anchor One	Anchor Two
Content Area 1	25 * 97 122 142	25 * 140 154	28 96	89 117 152 157
Content Area 2	12 147	5 47 84	23 34 43 58** 123	51 58** 92 136 142
Content Area 3	23 30 40 94 103 111 133 137* 157*	67 79 86 112 136 137* 138 157*	46** 53** 62 73 101 130	6 37 46** 49 53** 68 86 159
Content Area 4	68* 113*	18 68* 113*	45 60 143 151	5 18

(Table continues)

Table 35. (continued)

Content Area 5	21	7	57	84
	75	55	75	93**
	78	91	88	119
	93*	93*	93**	150
	107	126	107	
	128*	128*	150	
Content Area 6	50	2	8	22
	51	10	40	38
	52*	13	59	54
	53	22	76	55
	65	26	87	67
	71*	52*	102**	70
	72	64	104**	82
	74	71*	108	102**
	82	76	110	104**
	110	96	114	118
	121	127	115	122**
	134*	134*	122**	125
	148*	148*	134	146
Total	36	36	36	36

Note. * indicates that the items that chosen as anchor items for both anchor tests in Form A; ** indicates that the items chosen as anchor items for both anchor test in Form B.

Table 36

Spanish Language: Items That Chosen as Anchor Items for Two Anchor Tests

	Form A		Form B	
Content Area	Anchor One	Anchor Two	Anchor One	Anchor Two
Content Area 1	12 28* 126 131	28* 69	3 12 91 126	48 118
Content Area 2	15 59 60* 79	50 60* 87 125 152	50 59 60 125**	67 79 125** 141
Content Area 3	7 64* 90 105 129 140	2 20 32 45 63 64* 107 111 145	7 8 61 90 105 154	5 13 32 37 67 72 106 123 156
Content Area 4	22 36* 68 82	36* 114	11** 22 55 68 137	11** 82

(Table continues)

Table 36. (continued)

Content Area 5	6	21	19	74
	30	41	30	85
	74	99	121	99
	121*	121*	132	108
	123	139	154	113
	147*	147*		125
Content Area 6	19	16	29**	19
	34*	18	40	29**
	54*	26	54	53
	56*	34*	70	56
	57	54*	76	73
	70	56*	85**	85**
	85	77	89	110
	89	85	92	128
	131	93	133**	130
	136	100	136	133**
	143	103	143**	134
	151	128	153	143**
	153	143	159	151
Total	37	37	37	37

Note. * indicates that the items that chosen as anchor items for both anchor tests in Form A; ** indicates that the items chosen as anchor items for both anchor test in Form B.

Table 37

Chinese Language: Items That Chosen as Anchor Items for Two Anchor Tests

	Form A		Form B	
Content Area	Anchor One	Anchor Two	Anchor One	Anchor Two
Content Area 1	1* 18 21 25 59 93	1* 14 103 153	14 29** 92 129	29** 87 93 147
Content Area 2	89	19 71 81 157	42 47 102 109**	104 109**
Content Area 3	3 20 30 34 53 98 122	65 73 82 88 97 105 121 156 160	13 34 37 43 57 73 91 133 148	3 6 9 12 13 15 25 40 48 85 88 110
Content area 4	47 141 146*	32 146*	56	10 26

(Table continues)

Table 37. (continued)

Content Area 5	8	31*	18	16
	31*	37*	46	66
	37*	40	49	75
	52	63	62	149
	62	124	82	
	63		111	
	132			
Content Area 6	28*	28*	4	11
	38	35	27	51**
	64	39	51**	68
	67*	45	61	71
	68	50	73**	73**
	72*	60	77	94
	77	67*	95**	95**
	78	72*	112**	99
	80	85	125	112**
	87	91	126	120
	110	99	137	135
	120	118	138	142
	131	142	140	146
	145*	145*	151**	151**
Total	38	38	38	38

Note. * indicates that the items that chosen as anchor items for both anchor tests in Form A; ** indicates that the items chosen as anchor items for both anchor test in Form B.

APPENDIX F

Raw Score Statistics for Tests and Anchors in the SL and TL Examinee Groups

Table 38

Raw Score Statistics for Tests and Anchors for Anchor Test One

				Source Language (SL)				Target Language (TL)			
Group	Form	Section	No. of Anchor Items	N	Test Mean(SD)	Anchor Mean(SD)	Cor.[1]	N	Test Mean(SD)	Anchor Mean(SD)	Cor.
Korean	A	V1	23	875	77.32(11.92)	15.69(3.9)	.878	71	76.93(9.33)	15.94(4.15)	.894
		NV1	13		29.55(8.40)	7.51(2.86)	.878		35.27(7.17)	8.82(2.78)	.862
	B	V2	23	1422	81.28(11.68)	15.83(3.88)	.868	123	77.99(11.16)	15.81(3.94)	.857
		NV2	13		29.39(7.57)	6.88(2.89)	.863		31.38(6.82)	7.85(2.68)	.873
Spanish	A	V1	24	1454	75.67(11.67)	17.22(3.40)	.899	62	71.11(11.76)	15.84(4.44)	.908
		NV1	13		33.20(8.13)	8.63(2.80)	.884		15.84(4.43)	8.61(2.74)	.896
	B	V2	24	1441	75.75(11.49	17.53(3.96)	.885	114	70.14(12.23)	16.09(4.50)	.926
		NV2	13		33.22(8.11)	8.38(2.91)	.885		30.74(8.43)	7.92(3.05)	.916
Chinese	A	V1	24	1677	66.38(10.52)	16.80(4.11)	.882	130	63.6(11.79)	15.69(4.62)	.926
		NV1	14		27.48(7.48)	8.72(3.08)	.866		27.58(8.80)	8.75(3.43)	.892
	B	V2	14	1463	62.83(15.37)	16.99(5.56)	.912	116	63.87(12.46)	16.79(4.96)	.879
		NV2	14		37.10(8.33)	9.74(3.02)	.895		38.11(5.80)	9.64(2.77)	.826

Note. [1] Pearson correlation between test raw score and anchor raw score

Table 39

Raw Score Statistics for Tests and Anchors for Anchor Test Two

				Source Language (SL)				Target Language (TL)			
Group	Form	Section	No. of Anchor Items	N	Test Mean(SD)	Anchor Mean(SD)	Cor.[1]	N	Test Mean(SD)	Anchor Mean(SD)	Cor.
Korean	A	V1	23	875	77.32(11.92)	17.21(3.01)	.849	71	76.93(9.33)	17.07(2.46)	.783
		NV1	13		29.55(8.40)	7.14(2.46)	.850		35.27(7.17)	8.31(2.24)	.892
	B	V2	23	1422	81.28(11.68)	16.07(2.76)	.838	123	77.99(11.16)	15.41(2.81)	.805
		NV2	13		29.39(7.57)	7.45(2.09)	.826		31.38(6.82)	7.98(1.87)	.764
Spanish	A	V1	24	1454	75.67(11.67)	18.17(3.10)	.862	62	71.11(11.76)	17.31(3.25)	.877
		NV1	13		33.20(8.13)	7.92(2.44)	.835		31.39(8.02)	7.50(2.67)	.837
	B	V2	24	1441	75.75(11.49)	18.12(2.80)	.827	114	70.14(12.23)	17.07(2.97)	.823
		NV2	13		33.22(8.11)	8.50(2.38)	.839		30.74(8.43)	7.95(2.32)	.827
Chinese	A	V1	24	1677	66.38(10.52)	12.99(2.41)	.738	130	63.60(11.79)	12.47(2.86)	.826
		NV1	14		27.48(7.48)	7.51(2.52)	.835		27.58(8.80)	7.57(2.87)	.851
	B	V2	24	1463	62.83(15.37)	14.56(3.65)	.881	116	63.87(12.46)	14.80(2.80)	.819
		NV2	14		37.10(8.33)	9.52(2.74)	.874		38.11(5.80)	9.81(2.28)	.712

Note. [1] Pearson correlation between test raw score and anchor raw score

APPENDIX G

Statistics for Levine Linear Equating

and Mean-Sigma Equating

Table 40

Korean Language: Levine Equating and Mean-Sigma Equating for the Verbal Section

Raw Scores	Anchor Test One				Anchor Test Two			
	Levine Equating		Mean-Sigma Equating		Levine Equating		Mean-Sigma Equating	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
.00	-24.16	5.38	5.52	-14.03	5.69	-3.87	5.51	-14.03
1.00	-22.82	6.35	6.50	-12.83	6.62	-2.82	6.49	-12.83
2.00	-21.48	7.33	7.48	-11.64	7.54	-1.77	7.47	-11.63
3.00	-20.14	8.30	8.46	-10.44	8.46	-.72	8.45	-10.43
4.00	-18.80	9.28	9.44	-9.24	9.39	.33	9.43	-9.23
5.00	-17.47	10.25	10.42	-8.05	10.31	1.38	10.41	-8.03
6.00	-16.13	11.23	11.40	-6.85	11.23	2.43	11.39	-6.83
7.00	-14.79	12.20	12.38	-5.66	12.15	3.48	12.37	-5.63
8.00	-13.45	13.18	13.36	-4.46	13.08	4.53	13.35	-4.43
9.00	-12.11	14.15	14.34	-3.26	14.00	5.58	14.33	-3.23
10.00	-10.78	15.12	15.32	-2.07	14.92	6.63	15.31	-2.03
11.00	-9.44	16.10	16.30	-.87	15.84	7.68	16.29	-.83
12.00	-8.10	17.07	17.28	.32	16.77	8.73	17.27	.37
13.00	-6.76	18.05	18.26	1.52	17.69	9.78	18.25	1.57
14.00	-5.43	19.02	19.23	2.72	18.61	10.83	19.23	2.77
15.00	-4.09	20.00	20.21	3.91	19.53	11.88	20.21	3.97
16.00	-2.75	20.97	21.19	5.11	20.46	12.93	21.19	5.17
17.00	-1.41	21.95	22.17	6.31	21.38	13.98	22.17	6.37
18.00	-.07	22.92	23.15	7.50	22.30	15.03	23.15	7.57
19.00	1.26	23.90	24.13	8.70	23.23	16.08	24.13	8.77
20.00	2.60	24.87	25.11	9.89	24.15	17.13	25.11	9.97
21.00	3.94	25.84	26.09	11.09	25.07	18.18	26.09	11.17
22.00	5.28	26.82	27.07	12.29	25.99	19.23	27.07	12.37
23.00	6.61	27.79	28.05	13.48	26.92	20.28	28.05	13.57

24.00	7.95	28.77	29.03	14.68	27.84	21.32	29.03	14.77
25.00	9.29	29.74	30.01	15.87	28.76	22.37	30.01	15.97
26.00	10.63	30.72	30.99	17.07	29.68	23.42	30.99	17.17
27.00	11.97	31.69	31.97	18.27	30.61	24.47	31.97	18.37
28.00	13.30	32.67	32.95	19.46	31.53	25.52	32.95	19.57
29.00	14.64	33.64	33.93	20.66	32.45	26.57	33.93	20.77
30.00	15.98	34.62	34.91	21.86	33.38	27.62	34.91	21.97
31.00	17.32	35.59	35.89	23.05	34.30	28.67	35.89	23.17
32.00	18.66	36.57	36.87	24.25	35.22	29.72	36.87	24.37
33.00	19.99	37.54	37.85	25.44	36.14	30.77	37.85	25.57
34.00	21.33	38.51	38.83	26.64	37.07	31.82	38.83	26.77
35.00	22.67	39.49	39.81	27.84	37.99	32.87	39.81	27.97
36.00	24.01	40.46	40.79	29.03	38.91	33.92	40.79	29.17
37.00	25.34	41.44	41.77	30.23	39.83	34.97	41.77	30.37
38.00	26.68	42.41	42.75	31.42	40.76	36.02	42.75	31.57
39.00	28.02	43.39	43.73	32.62	41.68	37.07	43.73	32.77
40.00	29.36	44.36	44.71	33.82	42.60	38.12	44.71	33.97
41.00	30.70	45.34	45.69	35.01	43.53	39.17	45.69	35.17
42.00	32.03	46.31	46.67	36.21	44.45	40.22	46.67	36.37
43.00	33.37	47.29	47.65	37.40	45.37	41.27	47.65	37.57
44.00	34.71	48.26	48.63	38.60	46.29	42.32	48.63	38.77
45.00	36.05	49.23	49.61	39.80	47.22	43.37	49.61	39.97
46.00	37.39	50.21	50.59	40.99	48.14	44.42	50.59	41.17
47.00	38.72	51.18	51.57	42.19	49.06	45.47	51.57	42.37
48.00	40.06	52.16	52.55	43.39	49.98	46.52	52.55	43.57
49.00	41.40	53.13	53.53	44.58	50.91	47.57	53.53	44.77
50.00	42.74	54.11	54.51	45.78	51.83	48.62	54.51	45.97
51.00	44.07	55.08	55.49	46.97	52.75	49.67	55.49	47.17
52.00	45.41	56.06	56.47	48.17	53.67	50.72	56.47	48.37
53.00	46.75	57.03	57.45	49.37	54.60	51.77	57.45	49.57

54.00	48.09	58.01	58.43	50.56	55.52	52.82	58.43	50.77
55.00	49.43	58.98	59.41	51.76	56.44	53.86	59.41	51.97
56.00	50.76	59.96	60.39	52.95	57.37	54.91	60.39	53.17
57.00	52.10	60.93	61.37	54.15	58.29	55.96	61.37	54.37
58.00	53.44	61.90	62.35	55.35	59.21	57.01	62.35	55.57
59.00	54.78	62.88	63.33	56.54	60.13	58.06	63.33	56.77
60.00	56.11	63.85	64.31	57.74	61.06	59.11	64.31	57.97
61.00	57.45	64.83	65.29	58.94	61.98	60.16	65.29	59.17
62.00	58.79	65.80	66.27	60.13	62.90	61.21	66.27	60.37
63.00	60.13	66.78	67.25	61.33	63.82	62.26	67.25	61.57
64.00	61.47	67.75	68.23	62.52	64.75	63.31	68.23	62.77
65.00	62.80	68.73	69.21	63.72	65.67	64.36	69.21	63.97
66.00	64.14	69.70	70.19	64.92	66.59	65.41	70.19	65.17
67.00	65.48	70.68	71.17	66.11	67.52	66.46	71.17	66.37
68.00	66.82	71.65	72.15	67.31	68.44	67.51	72.15	67.57
69.00	68.16	72.62	73.13	68.50	69.36	68.56	73.13	68.77
70.00	69.49	73.60	74.11	69.70	70.28	69.61	74.11	69.97
71.00	70.83	74.57	75.09	70.90	71.21	70.66	75.09	71.17
72.00	72.17	75.55	76.07	72.09	72.13	71.71	76.07	72.37
73.00	73.51	76.52	77.05	73.29	73.05	72.76	77.05	73.57
74.00	74.84	77.50	78.03	74.49	73.97	73.81	78.03	74.77
75.00	76.18	78.47	79.01	75.68	74.90	74.86	79.01	75.97
76.00	77.52	79.45	79.99	76.88	75.82	75.91	79.99	77.17
77.00	78.86	80.42	80.97	78.07	76.74	76.96	80.97	78.37
78.00	80.20	81.40	81.95	79.27	77.67	78.01	81.95	79.57
79.00	81.53	82.37	82.93	80.47	78.59	79.06	82.93	80.77
80.00	82.87	83.35	83.91	81.66	79.51	80.11	83.91	81.97
81.00	84.21	84.32	84.89	82.86	80.43	81.16	84.89	83.17
82.00	85.55	85.29	85.87	84.05	81.36	82.21	85.87	84.37
83.00	86.88	86.27	86.85	85.25	82.28	83.26	86.85	85.57

84.00	88.22	87.24	87.83	86.45	83.20	84.31	87.83	86.77
85.00	89.56	88.22	88.81	87.64	84.12	85.36	88.81	87.97
86.00	90.90	89.19	89.79	88.84	85.05	86.41	89.79	89.17
87.00	92.24	90.17	90.77	90.04	85.97	87.45	90.77	90.37
88.00	93.57	91.14	91.74	91.23	86.89	88.50	91.75	91.57
89.00	94.91	92.12	92.72	92.43	87.81	89.55	92.73	92.77
90.00	96.25	93.09	93.70	93.62	88.74	90.60	93.71	93.97
91.00	97.59	94.07	94.68	94.82	89.66	91.65	94.69	95.17
92.00	98.93	95.04	95.66	96.02	90.58	92.70	95.67	96.37
93.00	100.26	96.01	96.64	97.21	91.51	93.75	96.65	97.57
94.00	101.60	96.99	97.62	98.41	92.43	94.80	97.63	98.77
95.00	102.94	97.96	98.60	99.60	93.35	95.85	98.61	99.97
96.00	104.28	98.94	99.58	100.80	94.27	96.90	99.59	101.17
97.00	105.61	99.91	100.56	102.00	95.20	97.95	100.57	102.37
98.00	106.95	100.89	101.54	103.19	96.12	99.00	101.55	103.57
99.00	108.29	101.86	102.52	104.39	97.04	100.05	102.53	104.77
100.00	109.63	102.84	103.50	105.58	97.96	101.10	103.51	105.97
101.00	110.97	103.81	104.48	106.78	98.89	102.15	104.49	107.17
102.00	112.30	104.79	105.46	107.98	99.81	103.20	105.47	108.37
103.00	113.64	105.76	106.44	109.17	100.73	104.25	106.45	109.57
104.00	114.98	106.74	107.42	110.37	101.66	105.30	107.43	110.77
105.00	116.32	107.71	108.40	111.57	102.58	106.35	108.41	111.97
106.00	117.65	108.68	109.38	112.76	103.50	107.40	109.39	113.17
107.00	118.99	109.66	110.36	113.96	104.42	108.45	110.37	114.37

Table 41

Korean Language: Levine Equating and Mean-Sigma Equating for Non-Verbal Section

Raw Scores	Anchor Test One				Anchor Test Two			
	Levine Equating		Mean-Sigma Equating		Levine Equating		Mean-Sigma Equating	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
.00	-5.78	2.11	2.72	-2.17	-1.61	1.62	2.72	-2.17
1.00	-4.65	3.07	3.63	-1.22	-.60	2.58	3.62	-1.22
2.00	-3.53	4.03	4.53	-.27	.42	3.54	4.52	-.27
3.00	-2.40	5.00	5.43	.69	1.44	4.50	5.42	.68
4.00	-1.27	5.96	6.33	1.64	2.45	5.45	6.32	1.63
5.00	-.15	6.93	7.24	2.59	3.47	6.41	7.22	2.58
6.00	.98	7.89	8.14	3.54	4.48	7.37	8.12	3.53
7.00	2.10	8.86	9.04	4.49	5.50	8.33	9.02	4.48
8.00	3.23	9.82	9.94	5.44	6.52	9.28	9.92	5.43
9.00	4.36	10.78	10.85	6.39	7.53	10.24	10.82	6.38
10.00	5.48	11.75	11.75	7.34	8.55	11.20	11.72	7.33
11.00	6.61	12.71	12.65	8.29	9.57	12.16	12.62	8.28
12.00	7.73	13.68	13.55	9.25	10.58	13.12	13.52	9.23
13.00	8.86	14.64	14.46	10.20	11.60	14.07	14.42	10.18
14.00	9.99	15.61	15.36	11.15	12.61	15.03	15.32	11.13
15.00	11.11	16.57	16.26	12.10	13.63	15.99	16.22	12.08
16.00	12.24	17.53	17.16	13.05	14.65	16.95	17.12	13.03
17.00	13.37	18.50	18.07	14.00	15.66	17.91	18.02	13.98
18.00	14.49	19.46	18.97	14.95	16.68	18.86	18.92	14.93
19.00	15.62	20.43	19.87	15.90	17.70	19.82	19.82	15.88
20.00	16.74	21.39	20.77	16.86	18.71	20.78	20.72	16.83
21.00	17.87	22.36	21.67	17.81	19.73	21.74	21.62	17.78
22.00	19.00	23.32	22.58	18.76	20.74	22.69	22.52	18.73
23.00	20.12	24.28	23.48	19.71	21.76	23.65	23.42	19.68

24.00	21.25	25.25	24.38	20.66	22.78	24.61	24.32	20.63
25.00	22.37	26.21	25.28	21.61	23.79	25.57	25.22	21.58
26.00	23.50	27.18	26.19	22.56	24.81	26.53	26.12	22.53
27.00	24.63	28.14	27.09	23.51	25.83	27.48	27.02	23.48
28.00	25.75	29.11	27.99	24.46	26.84	28.44	27.92	24.43
29.00	26.88	30.07	28.89	25.42	27.86	29.40	28.82	25.38
30.00	28.01	31.03	29.80	26.37	28.87	30.36	29.72	26.33
31.00	29.13	32.00	30.70	27.32	29.89	31.31	30.62	27.28
32.00	30.26	32.96	31.60	28.27	30.91	32.27	31.52	28.23
33.00	31.38	33.93	32.50	29.22	31.92	33.23	32.42	29.18
34.00	32.51	34.89	33.41	30.17	32.94	34.19	33.32	30.13
35.00	33.64	35.86	34.31	31.12	33.96	35.15	34.22	31.08
36.00	34.76	36.82	35.21	32.07	34.97	36.10	35.12	32.03
37.00	35.89	37.78	36.11	33.03	35.99	37.06	36.02	32.98
38.00	37.01	38.75	37.02	33.98	37.00	38.02	36.92	33.93
39.00	38.14	39.71	37.92	34.93	38.02	38.98	37.82	34.88
40.00	39.27	40.68	38.82	35.88	39.04	39.94	38.72	35.83
41.00	40.39	41.64	39.72	36.83	40.05	40.89	39.62	36.78
42.00	41.52	42.61	40.62	37.78	41.07	41.85	40.52	37.73
43.00	42.64	43.57	41.53	38.73	42.09	42.81	41.42	38.68
44.00	43.77	44.53	42.43	39.68	43.10	43.77	42.32	39.63
45.00	44.90	45.50	43.33	40.64	44.12	44.72	43.22	40.58
46.00	46.02	46.46	44.23	41.59	45.13	45.68	44.12	41.53
47.00	47.15	47.43	45.14	42.54	46.15	46.64	45.02	42.48
48.00	48.28	48.39	46.04	43.49	47.17	47.60	45.92	43.43
49.00	49.40	49.36	46.94	44.44	48.18	48.56	46.82	44.38
50.00	50.53	50.32	47.84	45.39	49.20	49.51	47.72	45.33
51.00	51.65	51.28	48.75	46.34	50.22	50.47	48.62	46.28
52.00	52.78	52.25	49.65	47.29	51.23	51.43	49.52	47.23
53.00	53.91	53.21	50.55	48.24	52.25	52.39	50.42	48.18

Table 42

Spanish Language: Levine Equating and Mean-Sigma Equating for the Verbal Section

Raw Scores	Anchor Test One				Anchor Test Two			
	Levine Equating		Mean-Sigma Equating		Levine Equating		Mean-Sigma Equating	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
.00	-8.96	-6.22	1.25	-3.80	-2.91	-1.05	1.25	-3.80
1.00	-7.84	-5.12	2.23	-2.76	-1.86	-.03	2.23	-2.76
2.00	-6.71	-4.02	3.21	-1.72	-.81	.99	3.21	-1.72
3.00	-5.58	-2.91	4.19	-.68	.24	2.01	4.19	-.68
4.00	-4.46	-1.81	5.17	.36	1.30	3.03	5.17	.36
5.00	-3.33	-.71	6.15	1.40	2.35	4.05	6.15	1.40
6.00	-2.20	.39	7.13	2.44	3.40	5.07	7.13	2.44
7.00	-1.08	1.49	8.11	3.48	4.45	6.09	8.11	3.48
8.00	.05	2.59	9.09	4.52	5.50	7.11	9.09	4.52
9.00	1.18	3.69	10.07	5.56	6.55	8.13	10.07	5.56
10.00	2.30	4.79	11.05	6.60	7.61	9.15	11.05	6.60
11.00	3.43	5.89	12.03	7.64	8.66	10.17	12.03	7.64
12.00	4.56	7.00	13.01	8.68	9.71	11.19	13.01	8.68
13.00	5.68	8.10	13.99	9.72	10.76	12.21	13.99	9.72
14.00	6.81	9.20	14.97	10.76	11.81	13.23	14.97	10.76
15.00	7.93	10.30	15.95	11.80	12.86	14.25	15.95	11.80
16.00	9.06	11.40	16.93	12.84	13.92	15.27	16.93	12.84
17.00	10.19	12.50	17.91	13.88	14.97	16.29	17.91	13.88
18.00	11.31	13.60	18.89	14.92	16.02	17.31	18.89	14.92
19.00	12.44	14.70	19.87	15.96	17.07	18.33	19.87	15.96
20.00	13.57	15.81	20.85	17.00	18.12	19.36	20.85	17.00
21.00	14.69	16.91	21.83	18.04	19.17	20.38	21.83	18.04
22.00	15.82	18.01	22.81	19.08	20.23	21.40	22.81	19.08
23.00	16.95	19.11	23.79	20.12	21.28	22.42	23.79	20.12

24.00	18.07	20.21	24.77	21.16	22.33	23.44	24.77	21.16
25.00	19.20	21.31	25.75	22.20	23.38	24.46	25.75	22.20
26.00	20.32	22.41	26.73	23.24	24.43	25.48	26.73	23.24
27.00	21.45	23.51	27.71	24.28	25.48	26.50	27.71	24.28
28.00	22.58	24.62	28.69	25.32	26.54	27.52	28.69	25.32
29.00	23.70	25.72	29.67	26.36	27.59	28.54	29.67	26.36
30.00	24.83	26.82	30.65	27.40	28.64	29.56	30.65	27.40
31.00	25.96	27.92	31.63	28.44	29.69	30.58	31.63	28.44
32.00	27.08	29.02	32.61	29.48	30.74	31.60	32.61	29.48
33.00	28.21	30.12	33.59	30.52	31.80	32.62	33.59	30.52
34.00	29.34	31.22	34.57	31.56	32.85	33.64	34.57	31.56
35.00	30.46	32.32	35.55	32.60	33.90	34.66	35.55	32.60
36.00	31.59	33.43	36.53	33.64	34.95	35.68	36.53	33.64
37.00	32.72	34.53	37.51	34.68	36.00	36.70	37.51	34.68
38.00	33.84	35.63	38.49	35.72	37.05	37.72	38.49	35.72
39.00	34.97	36.73	39.47	36.76	38.11	38.74	39.47	36.76
40.00	36.09	37.83	40.45	37.80	39.16	39.76	40.45	37.80
41.00	37.22	38.93	41.43	38.84	40.21	40.78	41.43	38.84
42.00	38.35	40.03	42.41	39.88	41.26	41.81	42.41	39.88
43.00	39.47	41.13	43.39	40.92	42.31	42.83	43.39	40.92
44.00	40.60	42.24	44.37	41.96	43.36	43.85	44.37	41.96
45.00	41.73	43.34	45.35	43.00	44.42	44.87	45.35	43.00
46.00	42.85	44.44	46.33	44.04	45.47	45.89	46.33	44.04
47.00	43.98	45.54	47.31	45.08	46.52	46.91	47.31	45.08
48.00	45.11	46.64	48.29	46.12	47.57	47.93	48.29	46.12
49.00	46.23	47.74	49.27	47.16	48.62	48.95	49.27	47.16
50.00	47.36	48.84	50.25	48.20	49.67	49.97	50.25	48.20
51.00	48.49	49.94	51.23	49.24	50.73	50.99	51.23	49.24
52.00	49.61	51.05	52.21	50.28	51.78	52.01	52.21	50.28
53.00	50.74	52.15	53.19	51.32	52.83	53.03	53.19	51.32

54.00	51.86	53.25	54.17	52.36	53.88	54.05	54.17	52.36
55.00	52.99	54.35	55.15	53.40	54.93	55.07	55.15	53.40
56.00	54.12	55.45	56.13	54.44	55.98	56.09	56.13	54.44
57.00	55.24	56.55	57.11	55.48	57.04	57.11	57.11	55.48
58.00	56.37	57.65	58.09	56.52	58.09	58.13	58.09	56.52
59.00	57.50	58.75	59.07	57.56	59.14	59.15	59.07	57.56
60.00	58.62	59.86	60.05	58.60	60.19	60.17	60.05	58.60
61.00	59.75	60.96	61.03	59.64	61.24	61.19	61.03	59.64
62.00	60.88	62.06	62.01	60.68	62.30	62.21	62.01	60.68
63.00	62.00	63.16	62.99	61.72	63.35	63.23	62.99	61.72
64.00	63.13	64.26	63.97	62.76	64.40	64.26	63.97	62.76
65.00	64.26	65.36	64.95	63.80	65.45	65.28	64.95	63.80
66.00	65.38	66.46	65.93	64.84	66.50	66.30	65.93	64.84
67.00	66.51	67.56	66.91	65.88	67.55	67.32	66.91	65.88
68.00	67.63	68.67	67.89	66.92	68.61	68.34	67.89	66.92
69.00	68.76	69.77	68.87	67.96	69.66	69.36	68.87	67.96
70.00	69.89	70.87	69.85	69.00	70.71	70.38	69.85	69.00
71.00	71.01	71.97	70.83	70.04	71.76	71.40	70.83	70.04
72.00	72.14	73.07	71.81	71.08	72.81	72.42	71.81	71.08
73.00	73.27	74.17	72.79	72.12	73.86	73.44	72.79	72.12
74.00	74.39	75.27	73.77	73.16	74.92	74.46	73.77	73.16
75.00	75.52	76.37	74.75	74.20	75.97	75.48	74.75	74.20
76.00	76.65	77.48	75.73	75.24	77.02	76.50	75.73	75.24
77.00	77.77	78.58	76.71	76.28	78.07	77.52	76.71	76.28
78.00	78.90	79.68	77.69	77.32	79.12	78.54	77.69	77.32
79.00	80.03	80.78	78.67	78.36	80.17	79.56	78.67	78.36
80.00	81.15	81.88	79.65	79.40	81.23	80.58	79.65	79.40
81.00	82.28	82.98	80.63	80.44	82.28	81.60	80.63	80.44
82.00	83.40	84.08	81.61	81.48	83.33	82.62	81.61	81.48
83.00	84.53	85.18	82.59	82.52	84.38	83.64	82.59	82.52

84.00	85.66	86.28	83.57	83.56	85.43	84.66	83.57	83.56
85.00	86.78	87.39	84.55	84.60	86.48	85.68	84.55	84.60
86.00	87.91	88.49	85.53	85.64	87.54	86.70	85.53	85.64
87.00	89.04	89.59	86.51	86.68	88.59	87.73	86.51	86.68
88.00	90.16	90.69	87.49	87.72	89.64	88.75	87.49	87.72
89.00	91.29	91.79	88.47	88.76	90.69	89.77	88.47	88.76
90.00	92.42	92.89	89.45	89.80	91.74	90.79	89.45	89.80
91.00	93.54	93.99	90.43	90.84	92.79	91.81	90.43	90.84
92.00	94.67	95.09	91.41	91.88	93.85	92.83	91.41	91.88
93.00	95.79	96.20	92.39	92.92	94.90	93.85	92.39	92.92
94.00	96.92	97.30	93.37	93.96	95.95	94.87	93.37	93.96
95.00	98.05	98.40	94.35	95.00	97.00	95.89	94.35	95.00
96.00	99.17	99.50	95.33	96.04	98.05	96.91	95.33	96.04
97.00	100.30	100.60	96.31	97.08	99.11	97.93	96.31	97.08
98.00	101.43	101.70	97.29	98.12	100.16	98.95	97.29	98.12
99.00	102.55	102.80	98.27	99.16	101.21	99.97	98.27	99.16
100.00	103.68	103.90	99.25	100.20	102.26	100.99	99.25	100.20
101.00	104.81	105.01	100.23	101.24	103.31	102.01	100.23	101.24
102.00	105.93	106.11	101.21	102.28	104.36	103.03	101.21	102.28
103.00	107.06	107.21	102.19	103.32	105.42	104.05	102.19	103.32
104.00	108.19	108.31	103.17	104.36	106.47	105.07	103.17	104.36
105.00	109.31	109.41	104.15	105.40	107.52	106.09	104.15	105.40
106.00	110.44	110.51	105.13	106.44	108.57	107.11	105.13	106.44
107.00	111.56	111.61	106.11	107.48	109.62	108.13	106.11	107.48

Table 43

Spanish Language: Levine Equating and Mean-Sigma Equating for Non-Verbal Section

Raw Scores	Anchor Test One				Anchor Test Two			
	Levine Equating		Mean-Sigma Equating		Levine Equating		Mean-Sigma Equating	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
.00	2.19	.40	.10	-2.25	-4.40	2.63	.10	-2.25
1.00	3.17	1.42	1.10	-1.20	-3.26	3.55	1.10	-1.20
2.00	4.16	2.44	2.10	-.15	-2.11	4.47	2.10	-.15
3.00	5.15	3.46	3.09	.90	-.97	5.40	3.09	.90
4.00	6.13	4.48	4.09	1.95	.18	6.32	4.09	1.95
5.00	7.12	5.50	5.09	3.00	1.32	7.24	5.09	3.00
6.00	8.11	6.53	6.09	4.05	2.47	8.16	6.09	4.05
7.00	9.09	7.55	7.08	5.10	3.61	9.08	7.08	5.10
8.00	10.08	8.57	8.08	6.15	4.76	10.01	8.08	6.15
9.00	11.07	9.59	9.08	7.20	5.90	10.93	9.08	7.20
10.00	12.05	10.61	10.08	8.25	7.05	11.85	10.08	8.25
11.00	13.04	11.63	11.07	9.30	8.19	12.77	11.07	9.30
12.00	14.03	12.65	12.07	10.35	9.34	13.69	12.07	10.35
13.00	15.01	13.67	13.07	11.40	10.48	14.61	13.07	11.40
14.00	16.00	14.69	14.07	12.45	11.63	15.54	14.07	12.45
15.00	16.99	15.72	15.06	13.50	12.77	16.46	15.06	13.50
16.00	17.97	16.74	16.06	14.55	13.92	17.38	16.06	14.55
17.00	18.96	17.76	17.06	15.60	15.06	18.30	17.06	15.60
18.00	19.95	18.78	18.06	16.65	16.20	19.22	18.06	16.65
19.00	20.94	19.80	19.05	17.70	17.35	20.15	19.05	17.70
20.00	21.92	20.82	20.05	18.75	18.49	21.07	20.05	18.75
21.00	22.91	21.84	21.05	19.80	19.64	21.99	21.05	19.80
22.00	23.90	22.86	22.05	20.85	20.78	22.91	22.05	20.85
23.00	24.88	23.88	23.04	21.90	21.93	23.83	23.04	21.90

24.00	25.87	24.91	24.04	22.95	23.07	24.75	24.04	22.95
25.00	26.86	25.93	25.04	24.00	24.22	25.68	25.04	24.00
26.00	27.84	26.95	26.04	25.05	25.36	26.60	26.04	25.05
27.00	28.83	27.97	27.03	26.10	26.51	27.52	27.03	26.10
28.00	29.82	28.99	28.03	27.15	27.65	28.44	28.03	27.15
29.00	30.80	30.01	29.03	28.20	28.80	29.36	29.03	28.20
30.00	31.79	31.03	30.03	29.25	29.94	30.29	30.03	29.25
31.00	32.78	32.05	31.02	30.30	31.09	31.21	31.02	30.30
32.00	33.76	33.07	32.02	31.35	32.23	32.13	32.02	31.35
33.00	34.75	34.10	33.02	32.40	33.38	33.05	33.02	32.40
34.00	35.74	35.12	34.02	33.45	34.52	33.97	34.02	33.45
35.00	36.72	36.14	35.01	34.50	35.67	34.89	35.01	34.50
36.00	37.71	37.16	36.01	35.55	36.81	35.82	36.01	35.55
37.00	38.70	38.18	37.01	36.60	37.96	36.74	37.01	36.60
38.00	39.68	39.20	38.01	37.65	39.10	37.66	38.01	37.65
39.00	40.67	40.22	39.00	38.70	40.25	38.58	39.00	38.70
40.00	41.66	41.24	40.00	39.75	41.39	39.50	40.00	39.75
41.00	42.65	42.26	41.00	40.80	42.54	40.43	41.00	40.80
42.00	43.63	43.28	42.00	41.85	43.68	41.35	42.00	41.85
43.00	44.62	44.31	42.99	42.90	44.82	42.27	42.99	42.90
44.00	45.61	45.33	43.99	43.95	45.97	43.19	43.99	43.95
45.00	46.59	46.35	44.99	45.00	47.11	44.11	44.99	45.00
46.00	47.58	47.37	45.99	46.05	48.26	45.03	45.99	46.05
47.00	48.57	48.39	46.98	47.10	49.40	45.96	46.98	47.10
48.00	49.55	49.41	47.98	48.15	50.55	46.88	47.98	48.15
49.00	50.54	50.43	48.98	49.20	51.69	47.80	48.98	49.20
50.00	51.53	51.45	49.98	50.25	52.84	48.72	49.98	50.25
51.00	52.51	52.47	50.97	51.30	53.98	49.64	50.97	51.30
52.00	53.50	53.50	51.97	52.35	55.13	50.57	51.97	52.35
53.00	54.49	54.52	52.97	53.40	56.27	51.49	52.97	53.40

Table 44

Chinese Language: Levine Equating and Mean-Sigma Equating for the Verbal Section

Raw Scores	Anchor Test One				Anchor Test Two			
	Levine Equating		Mean-Sigma Equating		Levine Equating		Mean-Sigma Equating	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
.00	-2.46	-6.15	-34.06	-3.34	-11.51	6.81	-34.06	-3.34
1.00	-1.43	-5.08	-32.60	-2.28	-10.33	7.70	-32.60	-2.28
2.00	-.40	-4.01	-31.14	-1.22	-9.16	8.60	-31.14	-1.22
3.00	.63	-2.94	-29.68	-.16	-7.98	9.49	-29.68	-.16
4.00	1.66	-1.87	-28.22	.90	-6.81	10.38	-28.22	.90
5.00	2.69	-.80	-26.76	1.96	-5.63	11.28	-26.76	1.96
6.00	3.73	.28	-25.30	3.02	-4.45	12.17	-25.30	3.02
7.00	4.76	1.35	-23.84	4.08	-3.28	13.06	-23.84	4.08
8.00	5.79	2.42	-22.38	5.14	-2.10	13.96	-22.38	5.14
9.00	6.82	3.49	-20.92	6.20	-.93	14.85	-20.92	6.20
10.00	7.85	4.56	-19.46	7.26	.25	15.74	-19.46	7.26
11.00	8.88	5.63	-18.00	8.32	1.43	16.64	-18.00	8.32
12.00	9.91	6.70	-16.54	9.38	2.60	17.53	-16.54	9.38
13.00	10.95	7.77	-15.08	10.44	3.78	18.42	-15.08	10.44
14.00	11.98	8.84	-13.62	11.50	4.95	19.32	-13.62	11.50
15.00	13.01	9.91	-12.16	12.56	6.13	20.21	-12.16	12.56
16.00	14.04	10.98	-10.70	13.62	7.31	21.10	-10.70	13.62
17.00	15.07	12.05	-9.24	14.68	8.48	22.00	-9.24	14.68
18.00	16.10	13.13	-7.78	15.74	9.66	22.89	-7.78	15.74
19.00	17.13	14.20	-6.32	16.80	10.83	23.78	-6.32	16.80
20.00	18.17	15.27	-4.86	17.86	12.01	24.68	-4.86	17.86
21.00	19.20	16.34	-3.40	18.92	13.19	25.57	-3.40	18.92
22.00	20.23	17.41	-1.94	19.98	14.36	26.46	-1.94	19.98
23.00	21.26	18.48	-.48	21.04	15.54	27.36	-.48	21.04

24.00	22.29	19.55	.98	22.10	16.71	28.25	.98	22.10
25.00	23.32	20.62	2.44	23.16	17.89	29.14	2.44	23.16
26.00	24.36	21.69	3.90	24.22	19.07	30.04	3.90	24.22
27.00	25.39	22.76	5.36	25.28	20.24	30.93	5.36	25.28
28.00	26.42	23.83	6.82	26.34	21.42	31.82	6.82	26.34
29.00	27.45	24.90	8.28	27.40	22.59	32.72	8.28	27.40
30.00	28.48	25.98	9.74	28.46	23.77	33.61	9.74	28.46
31.00	29.51	27.05	11.20	29.52	24.95	34.50	11.20	29.52
32.00	30.54	28.12	12.66	30.58	26.12	35.40	12.66	30.58
33.00	31.58	29.19	14.12	31.64	27.30	36.29	14.12	31.64
34.00	32.61	30.26	15.58	32.70	28.48	37.18	15.58	32.70
35.00	33.64	31.33	17.04	33.76	29.65	38.08	17.04	33.76
36.00	34.67	32.40	18.50	34.82	30.83	38.97	18.50	34.82
37.00	35.70	33.47	19.96	35.88	32.00	39.86	19.96	35.88
38.00	36.73	34.54	21.42	36.94	33.18	40.76	21.42	36.94
39.00	37.77	35.61	22.88	38.00	34.36	41.65	22.88	38.00
40.00	38.80	36.68	24.34	39.06	35.53	42.54	24.34	39.06
41.00	39.83	37.75	25.80	40.12	36.71	43.44	25.80	40.12
42.00	40.86	38.83	27.26	41.18	37.88	44.33	27.26	41.18
43.00	41.89	39.90	28.72	42.24	39.06	45.22	28.72	42.24
44.00	42.92	40.97	30.18	43.30	40.24	46.12	30.18	43.30
45.00	43.95	42.04	31.64	44.36	41.41	47.01	31.64	44.36
46.00	44.99	43.11	33.10	45.42	42.59	47.90	33.10	45.42
47.00	46.02	44.18	34.56	46.48	43.76	48.80	34.56	46.48
48.00	47.05	45.25	36.02	47.54	44.94	49.69	36.02	47.54
49.00	48.08	46.32	37.48	48.60	46.12	50.58	37.48	48.60
50.00	49.11	47.39	38.94	49.66	47.29	51.48	38.94	49.66
51.00	50.14	48.46	40.40	50.72	48.47	52.37	40.40	50.72
52.00	51.18	49.53	41.86	51.78	49.64	53.26	41.86	51.78
53.00	52.21	50.60	43.32	52.84	50.82	54.16	43.32	52.84

54.00	53.24	51.68	44.78	53.90	52.00	55.05	44.78	53.90
55.00	54.27	52.75	46.24	54.96	53.17	55.94	46.24	54.96
56.00	55.30	53.82	47.70	56.02	54.35	56.84	47.70	56.02
57.00	56.33	54.89	49.16	57.08	55.52	57.73	49.16	57.08
58.00	57.36	55.96	50.62	58.14	56.70	58.62	50.62	58.14
59.00	58.40	57.03	52.08	59.20	57.88	59.52	52.08	59.20
60.00	59.43	58.10	53.54	60.26	59.05	60.41	53.54	60.26
61.00	60.46	59.17	55.00	61.32	60.23	61.30	55.00	61.32
62.00	61.49	60.24	56.46	62.38	61.40	62.20	56.46	62.38
63.00	62.52	61.31	57.92	63.44	62.58	63.09	57.92	63.44
64.00	63.55	62.38	59.38	64.50	63.76	63.98	59.38	64.50
65.00	64.59	63.45	60.84	65.56	64.93	64.88	60.84	65.56
66.00	65.62	64.53	62.30	66.62	66.11	65.77	62.30	66.62
67.00	66.65	65.60	63.76	67.68	67.28	66.66	63.76	67.68
68.00	67.68	66.67	65.22	68.74	68.46	67.56	65.22	68.74
69.00	68.71	67.74	66.68	69.80	69.64	68.45	66.68	69.80
70.00	69.74	68.81	68.14	70.86	70.81	69.34	68.14	70.86
71.00	70.77	69.88	69.60	71.92	71.99	70.24	69.60	71.92
72.00	71.81	70.95	71.06	72.98	73.16	71.13	71.06	72.98
73.00	72.84	72.02	72.52	74.04	74.34	72.02	72.52	74.04
74.00	73.87	73.09	73.98	75.10	75.52	72.92	73.98	75.10
75.00	74.90	74.16	75.44	76.16	76.69	73.81	75.44	76.16
76.00	75.93	75.23	76.90	77.22	77.87	74.70	76.90	77.22
77.00	76.96	76.31	78.36	78.28	79.04	75.60	78.36	78.28
78.00	78.00	77.38	79.82	79.34	80.22	76.49	79.82	79.34
79.00	79.03	78.45	81.28	80.40	81.40	77.38	81.28	80.40
80.00	80.06	79.52	82.74	81.46	82.57	78.28	82.74	81.46
81.00	81.09	80.59	84.20	82.52	83.75	79.17	84.20	82.52
82.00	82.12	81.66	85.66	83.58	84.92	80.06	85.66	83.58
83.00	83.15	82.73	87.12	84.64	86.10	80.96	87.12	84.64

84.00	84.18	83.80	88.58	85.70	87.28	81.85	88.58	85.70
85.00	85.22	84.87	90.04	86.76	88.45	82.74	90.04	86.76
86.00	86.25	85.94	91.50	87.82	89.63	83.64	91.50	87.82
87.00	87.28	87.01	92.96	88.88	90.80	84.53	92.96	88.88
88.00	88.31	88.08	94.42	89.94	91.98	85.42	94.42	89.94
89.00	89.34	89.16	95.88	91.00	93.16	86.32	95.88	91.00
90.00	90.37	90.23	97.34	92.06	94.33	87.21	97.34	92.06
91.00	91.41	91.30	98.80	93.12	95.51	88.10	98.80	93.12
92.00	92.44	92.37	100.26	94.18	96.68	89.00	100.26	94.18
93.00	93.47	93.44	101.72	95.24	97.86	89.89	101.72	95.24
94.00	94.50	94.51	103.18	96.30	99.04	90.78	103.18	96.30
95.00	95.53	95.58	104.64	97.36	100.21	91.67	104.64	97.36
96.00	96.56	96.65	106.10	98.42	101.39	92.57	106.10	98.42
97.00	97.59	97.72	107.56	99.48	102.56	93.46	107.56	99.48
98.00	98.63	98.79	109.02	100.54	103.74	94.35	109.02	100.54
99.00	99.66	99.86	110.48	101.60	104.92	95.25	110.48	101.60
100.00	100.69	100.93	111.94	102.66	106.09	96.14	111.94	102.66
101.00	101.72	102.01	113.40	103.72	107.27	97.03	113.40	103.72
102.00	102.75	103.08	114.86	104.78	108.44	97.93	114.86	104.78
103.00	103.78	104.15	116.32	105.84	109.62	98.82	116.32	105.84
104.00	104.82	105.22	117.78	106.90	110.80	99.71	117.78	106.90
105.00	105.85	106.29	119.24	107.96	111.97	100.61	119.24	107.96
106.00	106.88	107.36	120.70	109.02	113.15	101.51	120.70	109.02
107.00	107.91	108.43	122.16	110.08	114.33	102.41	122.16	110.08

Table 45

Chinese Language: Levine Equating and Mean-Sigma Equating for Non-Verbal Section

Raw Scores	Anchor Test One				Anchor Test Two			
	Levine Equating		Mean-Sigma Equating		Levine Equating		Mean-Sigma Equating	
	First Link	Second Link	First Link	Second Link	First Link	Second Link	First Link	Second Link
.00	.65	-12.25	6.50	19.93	-.27	-4.25	6.50	19.93
1.00	1.62	-10.97	7.61	20.59	.74	-3.14	7.61	20.59
2.00	2.60	-9.68	8.72	21.25	1.76	-2.03	8.72	21.25
3.00	3.58	-8.39	9.83	21.91	2.77	-.91	9.83	21.91
4.00	4.55	-7.11	10.94	22.57	3.78	.20	10.94	22.57
5.00	5.53	-5.82	12.05	23.23	4.80	1.31	12.05	23.23
6.00	6.50	-4.53	13.16	23.89	5.81	2.42	13.16	23.89
7.00	7.48	-3.25	14.27	24.55	6.82	3.53	14.27	24.55
8.00	8.46	-1.96	15.38	25.21	7.84	4.64	15.38	25.21
9.00	9.43	-.67	16.49	25.87	8.85	5.75	16.49	25.87
10.00	10.41	.61	17.60	26.53	9.86	6.87	17.60	26.53
11.00	11.38	1.90	18.71	27.19	10.88	7.98	18.71	27.19
12.00	12.36	3.19	19.82	27.85	11.89	9.09	19.82	27.85
13.00	13.34	4.47	20.93	28.51	12.91	10.20	20.93	28.51
14.00	14.31	5.76	22.04	29.17	13.92	11.31	22.04	29.17
15.00	15.29	7.05	23.15	29.83	14.93	12.42	23.15	29.83
16.00	16.26	8.33	24.26	30.49	15.95	13.53	24.26	30.49
17.00	17.24	9.62	25.37	31.15	16.96	14.65	25.37	31.15
18.00	18.22	10.91	26.48	31.81	17.97	15.76	26.48	31.81
19.00	19.19	12.19	27.59	32.47	18.99	16.87	27.59	32.47
20.00	20.17	13.48	28.70	33.13	20.00	17.98	28.70	33.13
21.00	21.14	14.77	29.81	33.79	21.01	19.09	29.81	33.79
22.00	22.12	16.06	30.92	34.45	22.03	20.20	30.92	34.45
23.00	23.10	17.34	32.03	35.11	23.04	21.31	32.03	35.11

24.00	24.07	18.63	33.14	35.77	24.05	22.43	33.14	35.77
25.00	25.05	19.92	34.25	36.43	25.07	23.54	34.25	36.43
26.00	26.02	21.20	35.36	37.09	26.08	24.65	35.36	37.09
27.00	27.00	22.49	36.47	37.75	27.09	25.76	36.47	37.75
28.00	27.97	23.78	37.58	38.41	28.11	26.87	37.58	38.41
29.00	28.95	25.06	38.69	39.07	29.12	27.98	38.69	39.07
30.00	29.93	26.35	39.80	39.73	30.13	29.09	39.80	39.73
31.00	30.90	27.64	40.91	40.39	31.15	30.21	40.91	40.39
32.00	31.88	28.92	42.02	41.05	32.16	31.32	42.02	41.05
33.00	32.85	30.21	43.13	41.71	33.17	32.43	43.13	41.71
34.00	33.83	31.50	44.24	42.37	34.19	33.54	44.24	42.37
35.00	34.81	32.78	45.35	43.03	35.20	34.65	45.35	43.03
36.00	35.78	34.07	46.46	43.69	36.21	35.76	46.46	43.69
37.00	36.76	35.36	47.57	44.35	37.23	36.87	47.57	44.35
38.00	37.73	36.64	48.68	45.01	38.24	37.99	48.68	45.01
39.00	38.71	37.93	49.79	45.67	39.25	39.10	49.79	45.67
40.00	39.69	39.22	50.90	46.33	40.27	40.21	50.90	46.33
41.00	40.66	40.50	52.01	46.99	41.28	41.32	52.01	46.99
42.00	41.64	41.79	53.12	47.65	42.30	42.43	53.12	47.65
43.00	42.61	43.08	54.23	48.31	43.31	43.54	54.23	48.31
44.00	43.59	44.36	55.34	48.97	44.32	44.65	55.34	48.97
45.00	44.57	45.65	56.45	49.63	45.34	45.77	56.45	49.63
46.00	45.54	46.94	57.56	50.29	46.35	46.88	57.56	50.29
47.00	46.52	48.22	58.67	50.95	47.36	47.99	58.67	50.95
48.00	47.49	49.51	59.78	51.61	48.38	49.10	59.78	51.61
49.00	48.47	50.80	60.89	52.27	49.39	50.21	60.89	52.27
50.00	49.45	52.08	62.00	52.93	50.40	51.32	62.00	52.93
51.00	50.42	53.37	63.11	53.59	51.42	52.43	63.11	53.59
52.00	51.40	54.66	64.22	54.25	52.43	53.54	64.22	54.25
53.00	52.37	55.94	65.33	54.91	53.44	54.66	65.33	54.91

APPENDIX H

Statistics for Double Linking Equating

Table 46

Korean Language: Double Linking Results for the Verbal Section of Two Anchor Tests

	Anchor Test One		Anchor Test Two	
Raw Scores	First Chain	Second Chain	First Chain	Second Chain
.00	-18.17	-8.23	11.09	-18.60
1.00	-16.86	-7.07	11.99	-17.34
2.00	-15.55	-5.91	12.89	-16.08
3.00	-14.24	-4.75	13.79	-14.82
4.00	-12.93	-3.59	14.69	-13.56
5.00	-11.62	-2.43	15.59	-12.30
6.00	-10.31	-1.27	16.49	-11.04
7.00	-9.00	-.11	17.39	-9.78
8.00	-7.69	1.05	18.29	-8.52
9.00	-6.38	2.21	19.19	-7.26
10.00	-5.07	3.37	20.09	-6.00
11.00	-3.76	4.53	20.99	-4.74
12.00	-2.45	5.69	21.89	-3.48
13.00	-1.14	6.85	22.79	-2.22
14.00	.17	8.01	23.69	-.96
15.00	1.48	9.17	24.59	.30
16.00	2.79	10.33	25.49	1.56
17.00	4.10	11.49	26.39	2.82
18.00	5.41	12.65	27.29	4.08
19.00	6.72	13.81	28.19	5.34
20.00	8.03	14.97	29.09	6.60
21.00	9.34	16.13	29.99	7.86
22.00	10.65	17.29	30.89	9.12
23.00	11.96	18.45	31.79	10.38
24.00	13.27	19.61	32.69	11.64

25.00	14.58	20.77	33.59	12.90
26.00	15.89	21.93	34.49	14.16
27.00	17.20	23.09	35.39	15.42
28.00	18.51	24.25	36.29	16.68
29.00	19.82	25.41	37.19	17.94
30.00	21.13	26.57	38.09	19.20
31.00	22.44	27.73	38.99	20.46
32.00	23.75	28.89	39.89	21.72
33.00	25.06	30.05	40.79	22.98
34.00	26.37	31.21	41.69	24.24
35.00	27.68	32.37	42.59	25.50
36.00	28.99	33.53	43.49	26.76
37.00	30.30	34.69	44.39	28.02
38.00	31.61	35.85	45.29	29.28
39.00	32.92	37.01	46.19	30.54
40.00	34.23	38.17	47.09	31.80
41.00	35.54	39.33	47.99	33.06
42.00	36.85	40.49	48.89	34.32
43.00	38.16	41.65	49.79	35.58
44.00	39.47	42.81	50.69	36.84
45.00	40.78	43.97	51.59	38.10
46.00	42.09	45.13	52.49	39.36
47.00	43.40	46.29	53.39	40.62
48.00	44.71	47.45	54.29	41.88
49.00	46.02	48.61	55.19	43.14
50.00	47.33	49.77	56.09	44.40
51.00	48.64	50.93	56.99	45.66
52.00	49.95	52.09	57.89	46.92
53.00	51.26	53.25	58.79	48.18
54.00	52.57	54.41	59.69	49.44

55.00	53.88	55.57	60.59	50.70
56.00	55.19	56.73	61.49	51.96
57.00	56.50	57.89	62.39	53.22
58.00	57.81	59.05	63.29	54.48
59.00	59.12	60.21	64.19	55.74
60.00	60.43	61.37	65.09	57.00
61.00	61.74	62.53	65.99	58.26
62.00	63.05	63.69	66.89	59.52
63.00	64.36	64.85	67.79	60.78
64.00	65.67	66.01	68.69	62.04
65.00	66.98	67.17	69.59	63.30
66.00	68.29	68.33	70.49	64.56
67.00	69.60	69.49	71.39	65.82
68.00	70.91	70.65	72.29	67.08
69.00	72.22	71.81	73.19	68.34
70.00	73.53	72.97	74.09	69.60
71.00	74.84	74.13	74.99	70.86
72.00	76.15	75.29	75.89	72.12
73.00	77.46	76.45	76.79	73.38
74.00	78.77	77.61	77.69	74.64
75.00	80.08	78.77	78.59	75.90
76.00	81.39	79.93	79.49	77.16
77.00	82.70	81.09	80.39	78.42
78.00	84.01	82.25	81.29	79.68
79.00	85.32	83.41	82.19	80.94
80.00	86.63	84.57	83.09	82.20
81.00	87.94	85.73	83.99	83.46
82.00	89.25	86.89	84.89	84.72
83.00	90.56	88.05	85.79	85.98
84.00	91.87	89.21	86.69	87.24

85.00	93.18	90.37	87.59	88.50
86.00	94.49	91.53	88.49	89.76
87.00	95.80	92.69	89.39	91.02
88.00	97.11	93.85	90.29	92.28
89.00	98.42	95.01	91.19	93.54
90.00	99.73	96.17	92.09	94.80
91.00	101.04	97.33	92.99	96.06
92.00	102.35	98.49	93.89	97.32
93.00	103.66	99.65	94.79	98.58
94.00	104.97	100.81	95.69	99.84
95.00	106.28	101.97	96.59	101.10
96.00	107.59	103.13	97.49	102.36
97.00	108.90	104.29	98.39	103.62
98.00	110.21	105.45	99.29	104.88
99.00	111.52	106.61	100.19	106.14
100.00	112.83	107.77	101.09	107.40
101.00	114.14	108.93	101.99	108.66
102.00	115.45	110.09	102.89	109.92
103.00	116.76	111.25	103.79	111.18
104.00	118.07	112.41	104.69	112.44
105.00	119.38	113.57	105.59	113.70
106.00	120.69	114.73	106.49	114.96
107.00	122.00	115.89	107.39	116.22

Table 47

Korean Language: Double Linking Results for the Non-Verbal Section

	Anchor Test One		Anchor Test Two	
Raw Scores	First Chain	Second Chain	First Chain	Second Chain
.00	-2.48	.02	1.27	-.46
1.00	-1.47	.93	2.19	.45
2.00	-.46	1.84	3.11	1.36
3.00	.55	2.75	4.03	2.27
4.00	1.56	3.66	4.95	3.18
5.00	2.57	4.57	5.87	4.09
6.00	3.58	5.48	6.79	5.00
7.00	4.59	6.39	7.71	5.91
8.00	5.60	7.30	8.63	6.82
9.00	6.61	8.21	9.55	7.73
10.00	7.62	9.12	10.47	8.64
11.00	8.63	10.03	11.39	9.55
12.00	9.64	10.94	12.31	10.46
13.00	10.65	11.85	13.23	11.37
14.00	11.66	12.76	14.15	12.28
15.00	12.67	13.67	15.07	13.19
16.00	13.68	14.58	15.99	14.10
17.00	14.69	15.49	16.91	15.01
18.00	15.70	16.40	17.83	15.92
19.00	16.71	17.31	18.75	16.83
20.00	17.72	18.22	19.67	17.74
21.00	18.73	19.13	20.59	18.65
22.00	19.74	20.04	21.51	19.56
23.00	20.75	20.95	22.43	20.47
24.00	21.76	21.86	23.35	21.38

25.00	22.77	22.77	24.27	22.29
26.00	23.78	23.68	25.19	23.20
27.00	24.79	24.59	26.11	24.11
28.00	25.80	25.50	27.03	25.02
29.00	26.81	26.41	27.95	25.93
30.00	27.82	27.32	28.87	26.84
31.00	28.83	28.23	29.79	27.75
32.00	29.84	29.14	30.71	28.66
33.00	30.85	30.05	31.63	29.57
34.00	31.86	30.96	32.55	30.48
35.00	32.87	31.87	33.47	31.39
36.00	33.88	32.78	34.39	32.30
37.00	34.89	33.69	35.31	33.21
38.00	35.90	34.60	36.23	34.12
39.00	36.91	35.51	37.15	35.03
40.00	37.92	36.42	38.07	35.94
41.00	38.93	37.33	38.99	36.85
42.00	39.94	38.24	39.91	37.76
43.00	40.95	39.15	40.83	38.67
44.00	41.96	40.06	41.75	39.58
45.00	42.97	40.97	42.67	40.49
46.00	43.98	41.88	43.59	41.40
47.00	44.99	42.79	44.51	42.31
48.00	46.00	43.70	45.43	43.22
49.00	47.01	44.61	46.35	44.13
50.00	48.02	45.52	47.27	45.04
51.00	49.03	46.43	48.19	45.95
52.00	50.04	47.34	49.11	46.86
53.00	51.05	48.25	50.03	47.77

Table 48

Spanish Language: Double Linking Results for the Verbal Section of Two Anchor Tests

	Anchor Test One		Anchor Test Two	
Raw Scores	First Chain	Second Chain	First Chain	Second Chain
.00	-7.53	-10.40	-1.60	-4.93
1.00	-6.42	-9.26	-.57	-3.87
2.00	-5.31	-8.12	.46	-2.81
3.00	-4.20	-6.98	1.49	-1.75
4.00	-3.09	-5.84	2.52	-.69
5.00	-1.98	-4.70	3.55	.37
6.00	-.87	-3.56	4.58	1.43
7.00	.24	-2.42	5.61	2.49
8.00	1.35	-1.28	6.64	3.55
9.00	2.46	-.14	7.67	4.61
10.00	3.57	1.00	8.70	5.67
11.00	4.68	2.14	9.73	6.73
12.00	5.79	3.28	10.76	7.79
13.00	6.90	4.42	11.79	8.85
14.00	8.01	5.56	12.82	9.91
15.00	9.12	6.70	13.85	10.97
16.00	10.23	7.84	14.88	12.03
17.00	11.34	8.98	15.91	13.09
18.00	12.45	10.12	16.94	14.15
19.00	13.56	11.26	17.97	15.21
20.00	14.67	12.40	19.00	16.27
21.00	15.78	13.54	20.03	17.33
22.00	16.89	14.68	21.06	18.39
23.00	18.00	15.82	22.09	19.45
24.00	19.11	16.96	23.12	20.51

25.00	20.22	18.10	24.15	21.57
26.00	21.33	19.24	25.18	22.63
27.00	22.44	20.38	26.21	23.69
28.00	23.55	21.52	27.24	24.75
29.00	24.66	22.66	28.27	25.81
30.00	25.77	23.80	29.30	26.87
31.00	26.88	24.94	30.33	27.93
32.00	27.99	26.08	31.36	28.99
33.00	29.10	27.22	32.39	30.05
34.00	30.21	28.36	33.42	31.11
35.00	31.32	29.50	34.45	32.17
36.00	32.43	30.64	35.48	33.23
37.00	33.54	31.78	36.51	34.29
38.00	34.65	32.92	37.54	35.35
39.00	35.76	34.06	38.57	36.41
40.00	36.87	35.20	39.60	37.47
41.00	37.98	36.34	40.63	38.53
42.00	39.09	37.48	41.66	39.59
43.00	40.20	38.62	42.69	40.65
44.00	41.31	39.76	43.72	41.71
45.00	42.42	40.90	44.75	42.77
46.00	43.53	42.04	45.78	43.83
47.00	44.64	43.18	46.81	44.89
48.00	45.75	44.32	47.84	45.95
49.00	46.86	45.46	48.87	47.01
50.00	47.97	46.60	49.90	48.07
51.00	49.08	47.74	50.93	49.13
52.00	50.19	48.88	51.96	50.19
53.00	51.30	50.02	52.99	51.25
54.00	52.41	51.16	54.02	52.31

55.00	53.52	52.30	55.05	53.37
56.00	54.63	53.44	56.08	54.43
57.00	55.74	54.58	57.11	55.49
58.00	56.85	55.72	58.14	56.55
59.00	57.96	56.86	59.17	57.61
60.00	59.07	58.00	60.20	58.67
61.00	60.18	59.14	61.23	59.73
62.00	61.29	60.28	62.26	60.79
63.00	62.40	61.42	63.29	61.85
64.00	63.51	62.56	64.32	62.91
65.00	64.62	63.70	65.35	63.97
66.00	65.73	64.84	66.38	65.03
67.00	66.84	65.98	67.41	66.09
68.00	67.95	67.12	68.44	67.15
69.00	69.06	68.26	69.47	68.21
70.00	70.17	69.40	70.50	69.27
71.00	71.28	70.54	71.53	70.33
72.00	72.39	71.68	72.56	71.39
73.00	73.50	72.82	73.59	72.45
74.00	74.61	73.96	74.62	73.51
75.00	75.72	75.10	75.65	74.57
76.00	76.83	76.24	76.68	75.63
77.00	77.94	77.38	77.71	76.69
78.00	79.05	78.52	78.74	77.75
79.00	80.16	79.66	79.77	78.81
80.00	81.27	80.80	80.80	79.87
81.00	82.38	81.94	81.83	80.93
82.00	83.49	83.08	82.86	81.99
83.00	84.60	84.22	83.89	83.05
84.00	85.71	85.36	84.92	84.11

85.00	86.82	86.50	85.95	85.17
86.00	87.93	87.64	86.98	86.23
87.00	89.04	88.78	88.01	87.29
88.00	90.15	89.92	89.04	88.35
89.00	91.26	91.06	90.07	89.41
90.00	92.37	92.20	91.10	90.47
91.00	93.48	93.34	92.13	91.53
92.00	94.59	94.48	93.16	92.59
93.00	95.70	95.62	94.19	93.65
94.00	96.81	96.76	95.22	94.71
95.00	97.92	97.90	96.25	95.77
96.00	99.03	99.04	97.28	96.83
97.00	100.14	100.18	98.31	97.89
98.00	101.25	101.32	99.34	98.95
99.00	102.36	102.46	100.37	100.01
100.00	103.47	103.60	101.40	101.07
101.00	104.58	104.74	102.43	102.13
102.00	105.69	105.88	103.46	103.19
103.00	106.80	107.02	104.49	104.25
104.00	107.91	108.16	105.52	105.31
105.00	109.02	109.30	106.55	106.37
106.00	110.13	110.44	107.58	107.43
107.00	111.24	111.58	108.61	108.49

Table 49

Spanish Language: Double Linking Results for the Non-Verbal Section

	Anchor Test One		Anchor Test Two	
Raw Scores	First Chain	Second Chain	First Chain	Second Chain
.00	2.29	-1.90	-4.30	.56
1.00	3.28	-.83	-3.15	1.53
2.00	4.27	.24	-2.00	2.50
3.00	5.26	1.31	-.85	3.47
4.00	6.25	2.38	.30	4.44
5.00	7.24	3.45	1.45	5.41
6.00	8.23	4.52	2.60	6.38
7.00	9.22	5.59	3.75	7.35
8.00	10.21	6.66	4.90	8.32
9.00	11.20	7.73	6.05	9.29
10.00	12.19	8.80	7.20	10.26
11.00	13.18	9.87	8.35	11.23
12.00	14.17	10.94	9.50	12.20
13.00	15.16	12.01	10.65	13.17
14.00	16.15	13.08	11.80	14.14
15.00	17.14	14.15	12.95	15.11
16.00	18.13	15.22	14.10	16.08
17.00	19.12	16.29	15.25	17.05
18.00	20.11	17.36	16.40	18.02
19.00	21.10	18.43	17.55	18.99
20.00	22.09	19.50	18.70	19.96
21.00	23.08	20.57	19.85	20.93
22.00	24.07	21.64	21.00	21.90
23.00	25.06	22.71	22.15	22.87
24.00	26.05	23.78	23.30	23.84

25.00	27.04	24.85	24.45	24.81
26.00	28.03	25.92	25.60	25.78
27.00	29.02	26.99	26.75	26.75
28.00	30.01	28.06	27.90	27.72
29.00	31.00	29.13	29.05	28.69
30.00	31.99	30.20	30.20	29.66
31.00	32.98	31.27	31.35	30.63
32.00	33.97	32.34	32.50	31.60
33.00	34.96	33.41	33.65	32.57
34.00	35.95	34.48	34.80	33.54
35.00	36.94	35.55	35.95	34.51
36.00	37.93	36.62	37.10	35.48
37.00	38.92	37.69	38.25	36.45
38.00	39.91	38.76	39.40	37.42
39.00	40.90	39.83	40.55	38.39
40.00	41.89	40.90	41.70	39.36
41.00	42.88	41.97	42.85	40.33
42.00	43.87	43.04	44.00	41.30
43.00	44.86	44.11	45.15	42.27
44.00	45.85	45.18	46.30	43.24
45.00	46.84	46.25	47.45	44.21
46.00	47.83	47.32	48.60	45.18
47.00	48.82	48.39	49.75	46.15
48.00	49.81	49.46	50.90	47.12
49.00	50.80	50.53	52.05	48.09
50.00	51.79	51.60	53.20	49.06
51.00	52.78	52.67	54.35	50.03
52.00	53.77	53.74	55.50	51.00
53.00	54.76	54.81	56.65	51.97

Table 50

Chinese Language: Double Linking Results for the Verbal Section of Two Anchor Tests

	Anchor Test One		Anchor Test Two	
Raw Scores	First Chain	Second Chain	First Chain	Second Chain
.00	-37.65	-9.72	-50.86	3.84
1.00	-36.15	-8.59	-49.14	4.78
2.00	-34.65	-7.46	-47.42	5.72
3.00	-33.15	-6.33	-45.70	6.66
4.00	-31.65	-5.20	-43.98	7.60
5.00	-30.15	-4.07	-42.26	8.54
6.00	-28.65	-2.94	-40.54	9.48
7.00	-27.15	-1.81	-38.82	10.42
8.00	-25.65	-.68	-37.10	11.36
9.00	-24.15	.45	-35.38	12.30
10.00	-22.65	1.58	-33.66	13.24
11.00	-21.15	2.71	-31.94	14.18
12.00	-19.65	3.84	-30.22	15.12
13.00	-18.15	4.97	-28.50	16.06
14.00	-16.65	6.10	-26.78	17.00
15.00	-15.15	7.23	-25.06	17.94
16.00	-13.65	8.36	-23.34	18.88
17.00	-12.15	9.49	-21.62	19.82
18.00	-10.65	10.62	-19.90	20.76
19.00	-9.15	11.75	-18.18	21.70
20.00	-7.65	12.88	-16.46	22.64
21.00	-6.15	14.01	-14.74	23.58
22.00	-4.65	15.14	-13.02	24.52
23.00	-3.15	16.27	-11.30	25.46
24.00	-1.65	17.40	-9.58	26.40

25.00	-.15	18.53	-7.86	27.34
26.00	1.35	19.66	-6.14	28.28
27.00	2.85	20.79	-4.42	29.22
28.00	4.35	21.92	-2.70	30.16
29.00	5.85	23.05	-.98	31.10
30.00	7.35	24.18	.74	32.04
31.00	8.85	25.31	2.46	32.98
32.00	10.35	26.44	4.18	33.92
33.00	11.85	27.57	5.90	34.86
34.00	13.35	28.70	7.62	35.80
35.00	14.85	29.83	9.34	36.74
36.00	16.35	30.96	11.06	37.68
37.00	17.85	32.09	12.78	38.62
38.00	19.35	33.22	14.50	39.56
39.00	20.85	34.35	16.22	40.50
40.00	22.35	35.48	17.94	41.44
41.00	23.85	36.61	19.66	42.38
42.00	25.35	37.74	21.38	43.32
43.00	26.85	38.87	23.10	44.26
44.00	28.35	40.00	24.82	45.20
45.00	29.85	41.13	26.54	46.14
46.00	31.35	42.26	28.26	47.08
47.00	32.85	43.39	29.98	48.02
48.00	34.35	44.52	31.70	48.96
49.00	35.85	45.65	33.42	49.90
50.00	37.35	46.78	35.14	50.84
51.00	38.85	47.91	36.86	51.78
52.00	40.35	49.04	38.58	52.72
53.00	41.85	50.17	40.30	53.66
54.00	43.35	51.30	42.02	54.60

55.00	44.85	52.43	43.74	55.54
56.00	46.35	53.56	45.46	56.48
57.00	47.85	54.69	47.18	57.42
58.00	49.35	55.82	48.90	58.36
59.00	50.85	56.95	50.62	59.30
60.00	52.35	58.08	52.34	60.24
61.00	53.85	59.21	54.06	61.18
62.00	55.35	60.34	55.78	62.12
63.00	56.85	61.47	57.50	63.06
64.00	58.35	62.60	59.22	64.00
65.00	59.85	63.73	60.94	64.94
66.00	61.35	64.86	62.66	65.88
67.00	62.85	65.99	64.38	66.82
68.00	64.35	67.12	66.10	67.76
69.00	65.85	68.25	67.82	68.70
70.00	67.35	69.38	69.54	69.64
71.00	68.85	70.51	71.26	70.58
72.00	70.35	71.64	72.98	71.52
73.00	71.85	72.77	74.70	72.46
74.00	73.35	73.90	76.42	73.40
75.00	74.85	75.03	78.14	74.34
76.00	76.35	76.16	79.86	75.28
77.00	77.85	77.29	81.58	76.22
78.00	79.35	78.42	83.30	77.16
79.00	80.85	79.55	85.02	78.10
80.00	82.35	80.68	86.74	79.04
81.00	83.85	81.81	88.46	79.98
82.00	85.35	82.94	90.18	80.92
83.00	86.85	84.07	91.90	81.86
84.00	88.35	85.20	93.62	82.80

85.00	89.85	86.33	95.34	83.74
86.00	91.35	87.46	97.06	84.68
87.00	92.85	88.59	98.78	85.62
88.00	94.35	89.72	100.50	86.56
89.00	95.85	90.85	102.22	87.50
90.00	97.35	91.98	103.94	88.44
91.00	98.85	93.11	105.66	89.38
92.00	100.35	94.24	107.38	90.32
93.00	101.85	95.37	109.10	91.26
94.00	103.35	96.50	110.82	92.20
95.00	104.85	97.63	112.54	93.14
96.00	106.35	98.76	114.26	94.08
97.00	107.85	99.89	115.98	95.02
98.00	109.35	101.02	117.70	95.96
99.00	110.85	102.15	119.42	96.90
100.00	112.35	103.28	121.14	97.84
101.00	113.85	104.41	122.86	98.78
102.00	115.35	105.54	124.58	99.72
103.00	116.85	106.67	126.30	100.66
104.00	118.35	107.80	128.02	101.60
105.00	119.85	108.93	129.74	102.54
106.00	121.35	110.06	131.46	103.48
107.00	122.85	111.19	133.18	104.42

Table 51

Chinese Language: Double Linking Results for the Non-Verbal Section

	Anchor Test One		Anchor Test Two	
Raw Scores	First Chain	Second Chain	First Chain	Second Chain
.00	7.22	12.30	6.20	17.87
1.00	8.31	13.18	7.32	18.60
2.00	9.40	14.06	8.44	19.33
3.00	10.49	14.94	9.56	20.06
4.00	11.58	15.82	10.68	20.79
5.00	12.67	16.70	11.80	21.52
6.00	13.76	17.58	12.92	22.25
7.00	14.85	18.46	14.04	22.98
8.00	15.94	19.34	15.16	23.71
9.00	17.03	20.22	16.28	24.44
10.00	18.12	21.10	17.40	25.17
11.00	19.21	21.98	18.52	25.90
12.00	20.30	22.86	19.64	26.63
13.00	21.39	23.74	20.76	27.36
14.00	22.48	24.62	21.88	28.09
15.00	23.57	25.50	23.00	28.82
16.00	24.66	26.38	24.12	29.55
17.00	25.75	27.26	25.24	30.28
18.00	26.84	28.14	26.36	31.01
19.00	27.93	29.02	27.48	31.74
20.00	29.02	29.90	28.60	32.47
21.00	30.11	30.78	29.72	33.20
22.00	31.20	31.66	30.84	33.93
23.00	32.29	32.54	31.96	34.66
24.00	33.38	33.42	33.08	35.39

25.00	34.47	34.30	34.20	36.12
26.00	35.56	35.18	35.32	36.85
27.00	36.65	36.06	36.44	37.58
28.00	37.74	36.94	37.56	38.31
29.00	38.83	37.82	38.68	39.04
30.00	39.92	38.70	39.80	39.77
31.00	41.01	39.58	40.92	40.50
32.00	42.10	40.46	42.04	41.23
33.00	43.19	41.34	43.16	41.96
34.00	44.28	42.22	44.28	42.69
35.00	45.37	43.10	45.40	43.42
36.00	46.46	43.98	46.52	44.15
37.00	47.55	44.86	47.64	44.88
38.00	48.64	45.74	48.76	45.61
39.00	49.73	46.62	49.88	46.34
40.00	50.82	47.50	51.00	47.07
41.00	51.91	48.38	52.12	47.80
42.00	53.00	49.26	53.24	48.53
43.00	54.09	50.14	54.36	49.26
44.00	55.18	51.02	55.48	49.99
45.00	56.27	51.90	56.60	50.72
46.00	57.36	52.78	57.72	51.45
47.00	58.45	53.66	58.84	52.18
48.00	59.54	54.54	59.96	52.91
49.00	60.63	55.42	61.08	53.64
50.00	61.72	56.30	62.20	54.37
51.00	62.81	57.18	63.32	55.10
52.00	63.90	58.06	64.44	55.83
53.00	64.99	58.94	65.56	56.56

APPENDIX I

Differences between the Two Functions for Two Anchor Tests

Table 52

All Language Group: Differences between the Two Functions for the Verbal Section of Two Anchor Tests

	Korean Language		Spanish Language		Chinese Language	
Raw Score	Anchor One	Anchor Two	Anchor One	Anchor Two	Anchor One	Anchor Two
.00	-9.94	29.69	2.87	3.33	-27.93	-54.70
1.00	-9.79	29.33	2.84	3.30	-27.56	-53.92
2.00	-9.64	28.97	2.81	3.27	-27.19	-53.14
3.00	-9.49	28.61	2.78	3.24	-26.82	-52.36
4.00	-9.34	28.25	2.75	3.21	-26.45	-51.58
5.00	-9.19	27.89	2.72	3.18	-26.08	-50.80
6.00	-9.04	27.53	2.69	3.15	-25.71	-50.02
7.00	-8.89	27.17	2.66	3.12	-25.34	-49.24
8.00	-8.74	26.81	2.63	3.09	-24.97	-48.46
9.00	-8.59	26.45	2.60	3.06	-24.60	-47.68
10.00	-8.44	26.09	2.57	3.03	-24.23	-46.90
11.00	-8.29	25.73	2.54	3.00	-23.86	-46.12
12.00	-8.14	25.37	2.51	2.97	-23.49	-45.34
13.00	-7.99	25.01	2.48	2.94	-23.12	-44.56
14.00	-7.84	24.65	2.45	2.91	-22.75	-43.78
15.00	-7.69	24.29	2.42	2.88	-22.38	-43.00
16.00	-7.54	23.93	2.39	2.85	-22.01	-42.22
17.00	-7.39	23.57	2.36	2.82	-21.64	-41.44
18.00	-7.24	23.21	2.33	2.79	-21.27	-40.66
19.00	-7.09	22.85	2.30	2.76	-20.90	-39.88
20.00	-6.94	22.49	2.27	2.73	-20.53	-39.10
21.00	-6.79	22.13	2.24	2.70	-20.16	-38.32
22.00	-6.64	21.77	2.21	2.67	-19.79	-37.54

23.00	-6.49	21.41	2.18	2.64	-19.42	-36.76
24.00	-6.34	21.05	2.15	2.61	-19.05	-35.98
25.00	-6.19	20.69	2.12	2.58	-18.68	-35.20
26.00	-6.04	20.33	2.09	2.55	-18.31	-34.42
27.00	-5.89	19.97	2.06	2.52	-17.94	-33.64
28.00	-5.74	19.61	2.03	2.49	-17.57	-32.86
29.00	-5.59	19.25	2.00	2.46	-17.20	-32.08
30.00	-5.44	18.89	1.97	2.43	-16.83	-31.30
31.00	-5.29	18.53	1.94	2.40	-16.46	-30.52
32.00	-5.14	18.17	1.91	2.37	-16.09	-29.74
33.00	-4.99	17.81	1.88	2.34	-15.72	-28.96
34.00	-4.84	17.45	1.85	2.31	-15.35	-28.18
35.00	-4.69	17.09	1.82	2.28	-14.98	-27.40
36.00	-4.54	16.73	1.79	2.25	-14.61	-26.62
37.00	-4.39	16.37	1.76	2.22	-14.24	-25.84
38.00	-4.24	16.01	1.73	2.19	-13.87	-25.06
39.00	-4.09	15.65	1.70	2.16	-13.50	-24.28
40.00	-3.94	15.29	1.67	2.13	-13.13	-23.50
41.00	-3.79	14.93	1.64	2.10	-12.76	-22.72
42.00	-3.64	14.57	1.61	2.07	-12.39	-21.94
43.00	-3.49	14.21	1.58	2.04	-12.02	-21.16
44.00	-3.34	13.85	1.55	2.01	-11.65	-20.38
45.00	-3.19	13.49	1.52	1.98	-11.28	-19.60
46.00	-3.04	13.13	1.49	1.95	-10.91	-18.82
47.00	-2.89	12.77	1.46	1.92	-10.54	-18.04
48.00	-2.74	12.41	1.43	1.89	-10.17	-17.26
49.00	-2.59	12.05	1.40	1.86	-9.80	-16.48
50.00	-2.44	11.69	1.37	1.83	-9.43	-15.70
51.00	-2.29	11.33	1.34	1.80	-9.06	-14.92
52.00	-2.14	10.97	1.31	1.77	-8.69	-14.14

53.00	-1.99	10.61	1.28	1.74	-8.32	-13.36
54.00	-1.84	10.25	1.25	1.71	-7.95	-12.58
55.00	-1.69	9.89	1.22	1.68	-7.58	-11.80
56.00	-1.54	9.53	1.19	1.65	-7.21	-11.02
57.00	-1.39	9.17	1.16	1.62	-6.84	-10.24
58.00	-1.24	8.81	1.13	1.59	-6.47	-9.46
59.00	-1.09	8.45	1.10	1.56	-6.10	-8.68
60.00	-.94	8.09	1.07	1.53	-5.73	-7.90
61.00	-.79	7.73	1.04	1.50	-5.36	-7.12
62.00	-.64	7.37	1.01	1.47	-4.99	-6.34
63.00	-.49	7.01	.98	1.44	-4.62	-5.56
64.00	-.34	6.65	.95	1.41	-4.25	-4.78
65.00	-.19	6.29	.92	1.38	-3.88	-4.00
66.00	-.04	5.93	.89	1.35	-3.51	-3.22
67.00	.11	5.57	.86	1.32	-3.14	-2.44
68.00	.26	5.21	.83	1.29	-2.77	-1.66
69.00	.41	4.85	.80	1.26	-2.40	-.88
70.00	.56	4.49	.77	1.23	-2.03	-.10
71.00	.71	4.13	.74	1.20	-1.66	.68
72.00	.86	3.77	.71	1.17	-1.29	1.46
73.00	1.01	3.41	.68	1.14	-.92	2.24
74.00	1.16	3.05	.65	1.11	-.55	3.02
75.00	1.31	2.69	.62	1.08	-.18	3.80
76.00	1.46	2.33	.59	1.05	.19	4.58
77.00	1.61	1.97	.56	1.02	.56	5.36
78.00	1.76	1.61	.53	.99	.93	6.14
79.00	1.91	1.25	.50	.96	1.30	6.92
80.00	2.06	.89	.47	.93	1.67	7.70
81.00	2.21	.53	.44	.90	2.04	8.48
82.00	2.36	.17	.41	.87	2.41	9.26

83.00	2.51	-.19	.38	.84	2.78	10.04
84.00	2.66	-.55	.35	.81	3.15	10.82
85.00	2.81	-.91	.32	.78	3.52	11.60
86.00	2.96	-1.27	.29	.75	3.89	12.38
87.00	3.11	-1.63	.26	.72	4.26	13.16
88.00	3.26	-1.99	.23	.69	4.63	13.94
89.00	3.41	-2.35	.20	.66	5.00	14.72
90.00	3.56	-2.71	.17	.63	5.37	15.50
91.00	3.71	-3.07	.14	.60	5.74	16.28
92.00	3.86	-3.43	.11	.57	6.11	17.06
93.00	4.01	-3.79	.08	.54	6.48	17.84
94.00	4.16	-4.15	.05	.51	6.85	18.62
95.00	4.31	-4.51	.02	.48	7.22	19.40
96.00	4.46	-4.87	-.01	.45	7.59	20.18
97.00	4.61	-5.23	-.04	.42	7.96	20.96
98.00	4.76	-5.59	-.07	.39	8.33	21.74
99.00	4.91	-5.95	-.10	.36	8.70	22.52
100.00	5.06	-6.31	-.13	.33	9.07	23.30
101.00	5.21	-6.67	-.16	.30	9.44	24.08
102.00	5.36	-7.03	-.19	.27	9.81	24.86
103.00	5.51	-7.39	-.22	.24	10.18	25.64
104.00	5.66	-7.75	-.25	.21	10.55	26.42
105.00	5.81	-8.11	-.28	.18	10.92	27.20
106.00	5.96	-8.47	-.31	.15	11.29	27.98
107.00	6.11	-8.83	-.34	.12	11.66	28.76

Table 53

All Language Group: Differences between the Two Functions for the Non-Verbal Section

	Korean Language		Spanish Language		Chinese Language	
Raw Scores	Anchor One	Anchor Two	Anchor One	Anchor Two	Anchor One	Anchor Two
.00	-2.50	1.73	4.19	-4.86	-5.08	-11.67
1.00	-2.40	1.74	4.11	-4.68	-4.87	-11.28
2.00	-2.30	1.75	4.03	-4.50	-4.66	-10.89
3.00	-2.20	1.76	3.95	-4.32	-4.45	-10.50
4.00	-2.10	1.77	3.87	-4.14	-4.24	-10.11
5.00	-2.00	1.78	3.79	-3.96	-4.03	-9.72
6.00	-1.90	1.79	3.71	-3.78	-3.82	-9.33
7.00	-1.80	1.80	3.63	-3.60	-3.61	-8.94
8.00	-1.70	1.81	3.55	-3.42	-3.40	-8.55
9.00	-1.60	1.82	3.47	-3.24	-3.19	-8.16
10.00	-1.50	1.83	3.39	-3.06	-2.98	-7.77
11.00	-1.40	1.84	3.31	-2.88	-2.77	-7.38
12.00	-1.30	1.85	3.23	-2.70	-2.56	-6.99
13.00	-1.20	1.86	3.15	-2.52	-2.35	-6.60
14.00	-1.10	1.87	3.07	-2.34	-2.14	-6.21
15.00	-1.00	1.88	2.99	-2.16	-1.93	-5.82
16.00	-.90	1.89	2.91	-1.98	-1.72	-5.43
17.00	-.80	1.90	2.83	-1.80	-1.51	-5.04
18.00	-.70	1.91	2.75	-1.62	-1.30	-4.65
19.00	-.60	1.92	2.67	-1.44	-1.09	-4.26
20.00	-.50	1.93	2.59	-1.26	-.88	-3.87
21.00	-.40	1.94	2.51	-1.08	-.67	-3.48
22.00	-.30	1.95	2.43	-.90	-.46	-3.09
23.00	-.20	1.96	2.35	-.72	-.25	-2.70
24.00	-.10	1.97	2.27	-.54	-.04	-2.31

25.00	.00	1.98	2.19	-.36	.17	-1.92
26.00	.10	1.99	2.11	-.18	.38	-1.53
27.00	.20	2.00	2.03	.00	.59	-1.14
28.00	.30	2.01	1.95	.18	.80	-.75
29.00	.40	2.02	1.87	.36	1.01	-.36
30.00	.50	2.03	1.79	.54	1.22	.03
31.00	.60	2.04	1.71	.72	1.43	.42
32.00	.70	2.05	1.63	.90	1.64	.81
33.00	.80	2.06	1.55	1.08	1.85	1.20
34.00	.90	2.07	1.47	1.26	2.06	1.59
35.00	1.00	2.08	1.39	1.44	2.27	1.98
36.00	1.10	2.09	1.31	1.62	2.48	2.37
37.00	1.20	2.10	1.23	1.80	2.69	2.76
38.00	1.30	2.11	1.15	1.98	2.90	3.15
39.00	1.40	2.12	1.07	2.16	3.11	3.54
40.00	1.50	2.13	.99	2.34	3.32	3.93
41.00	1.60	2.14	.91	2.52	3.53	4.32
42.00	1.70	2.15	.83	2.70	3.74	4.71
43.00	1.80	2.16	.75	2.88	3.95	5.10
44.00	1.90	2.17	.67	3.06	4.16	5.49
45.00	2.00	2.18	.59	3.24	4.37	5.88
46.00	2.10	2.19	.51	3.42	4.58	6.27
47.00	2.20	2.20	.43	3.60	4.79	6.66
48.00	2.30	2.21	.35	3.78	5.00	7.05
49.00	2.40	2.22	.27	3.96	5.21	7.44
50.00	2.50	2.23	.19	4.14	5.42	7.83
51.00	2.60	2.24	.11	4.32	5.63	8.22
52.00	2.70	2.25	.03	4.50	5.84	8.61
53.00	2.80	2.26	-.05	4.68	6.05	9.00

APPENDIX J

Graphs for Double Linking

Figure 69

Korean Language: Equating Functions of Two Equating Chains for the Verbal Section of Anchor Test One

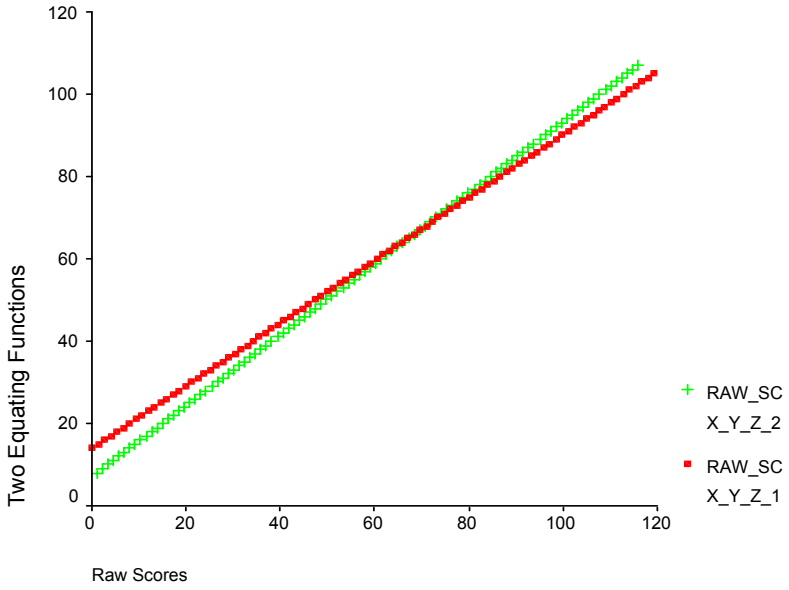


Figure 70

Korean Language: the Differences between the Two Functions for Anchor Test One

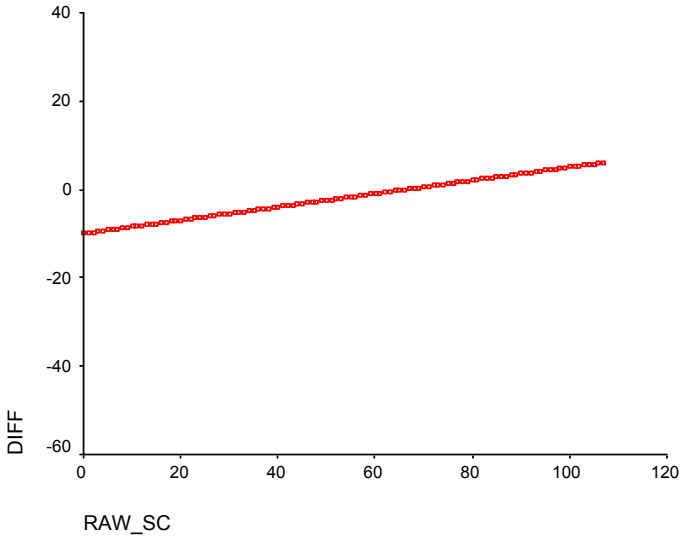


Figure 71

Korean Language: Equating Functions of the Two Equating Chains for the Verbal Section of Anchor Test Two

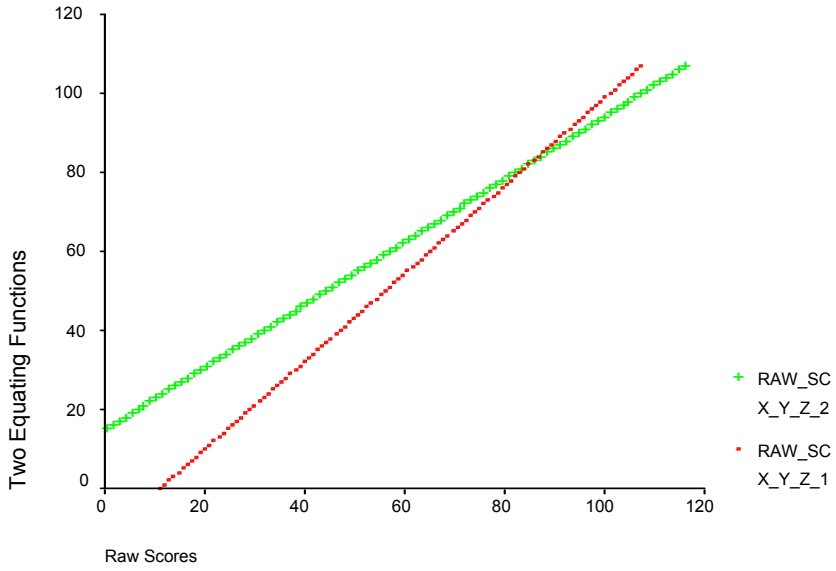


Figure 72

Korean Language: the Differences between the Two Functions for Anchor Test Two

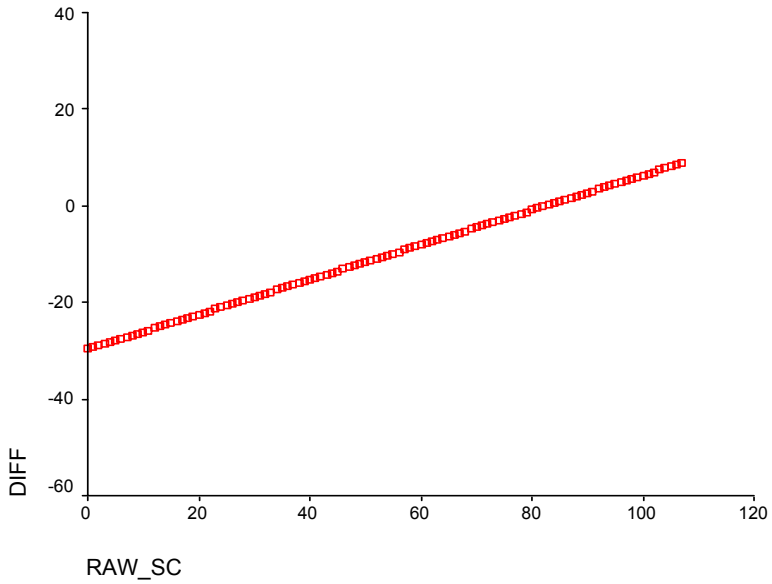


Figure 73

Korean Language: Equating Functions of the Two Equating Chains for the Non-Verbal Section of Anchor Test One

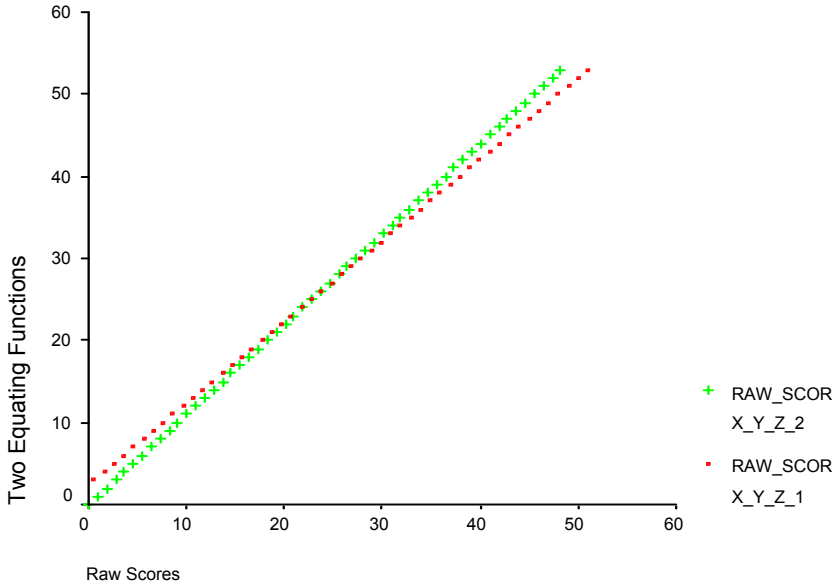


Figure 74

Korean Language: the Differences between the Two Functions for Anchor Test One

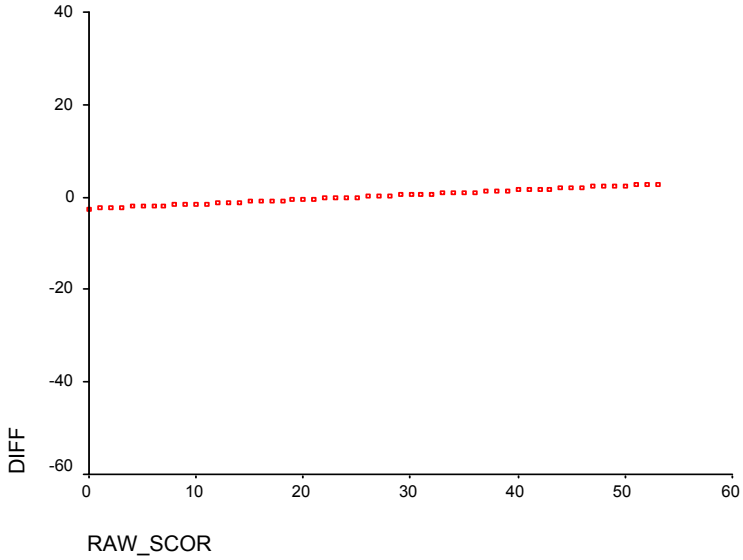


Figure 75

Korean Language: Equating Functions of the Two Equating Chains for the Non-Verbal Section of Anchor Test Two

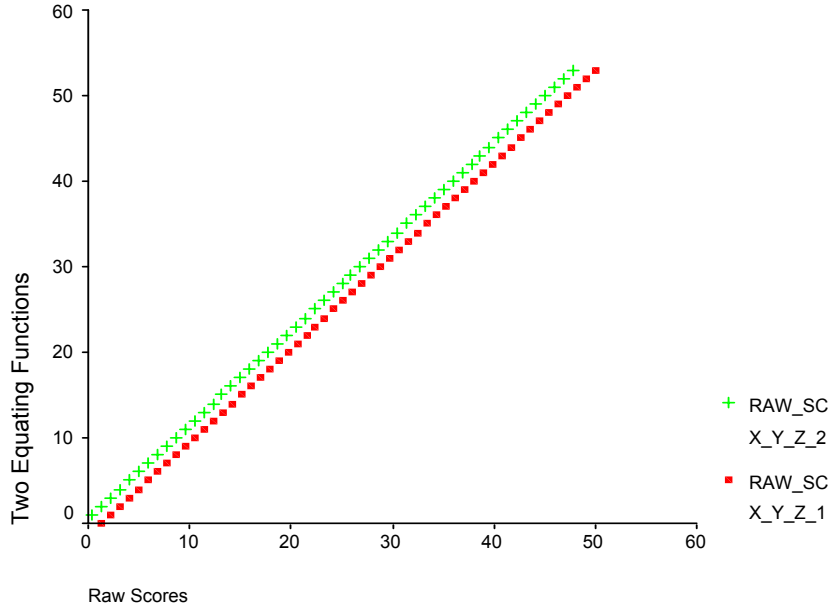


Figure 76

Korean Language: the Differences between the Two Functions for Anchor Test Two

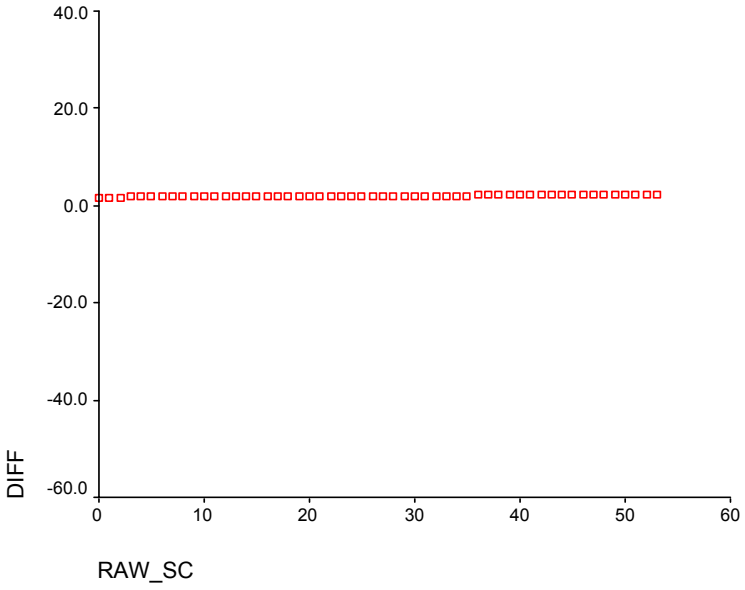


Figure 77

Spanish Language: Equating Functions of the Two Equating Chains for the Verbal Section of Anchor Test One

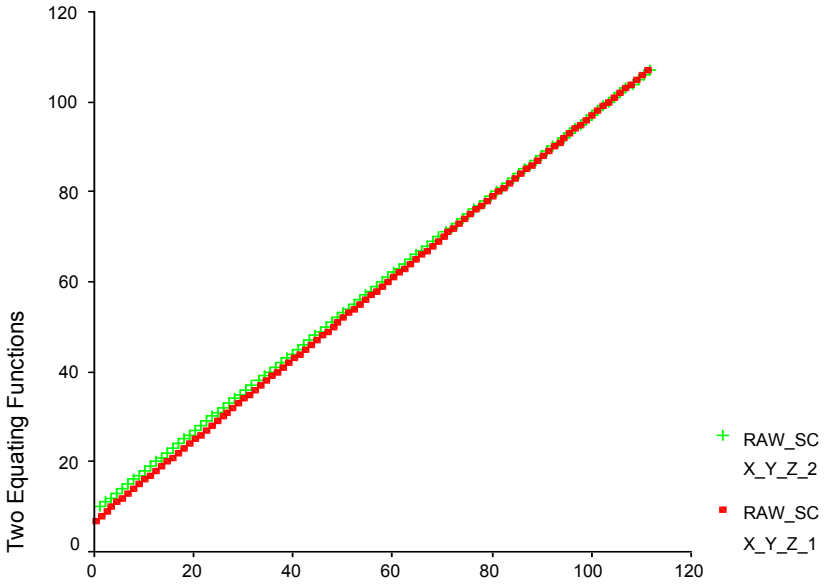


Figure 78

Spanish Language: the Differences between Two Functions for Anchor Test One

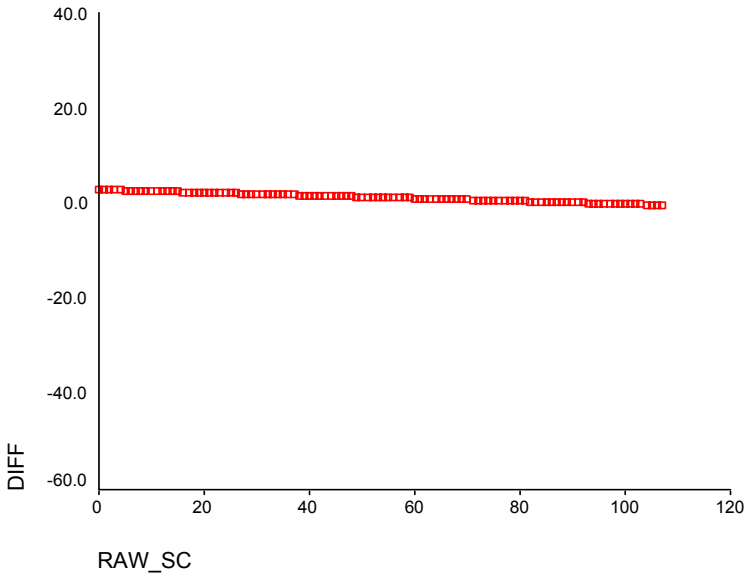


Figure 79

Spanish Language: Equating Functions of the Two Equating Chains for the Verbal Section of Anchor Test Two

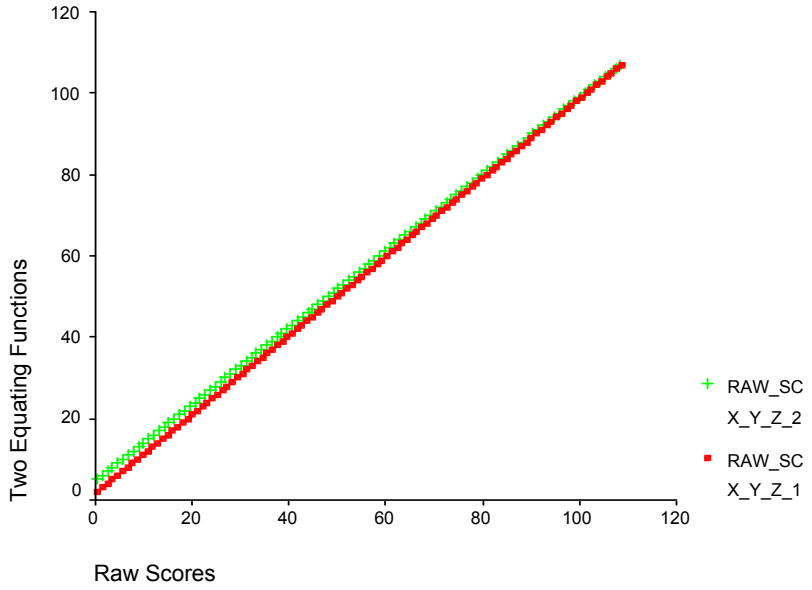


Figure 80

Spanish Language: the Differences between the Two Functions for Anchor Test Two

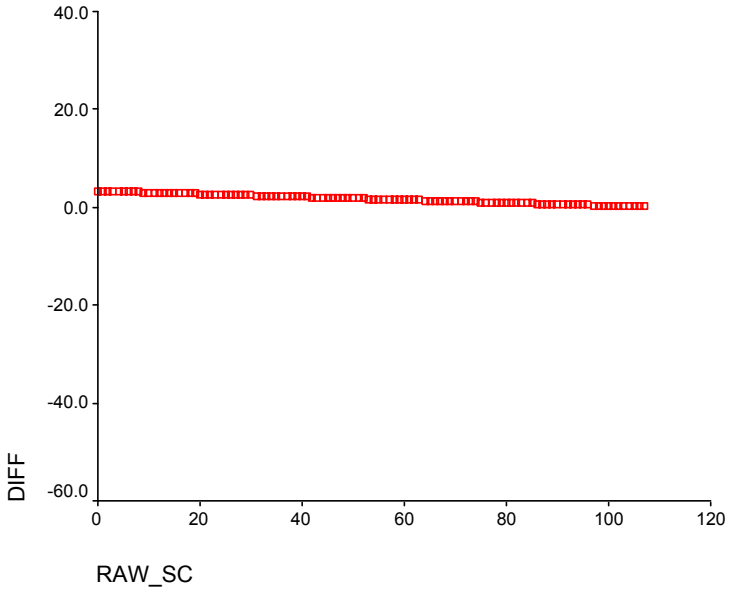


Figure 81

Spanish Language: Equating Functions of the Two Equating Chains for the Non-Verbal Section of Anchor Test One

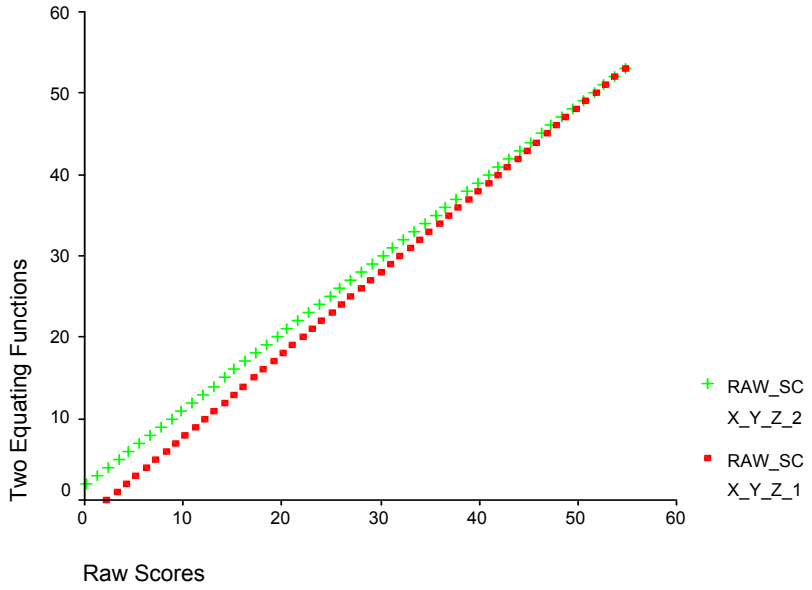


Figure 82

Spanish Language: the Differences between the Two Functions for Anchor Test One

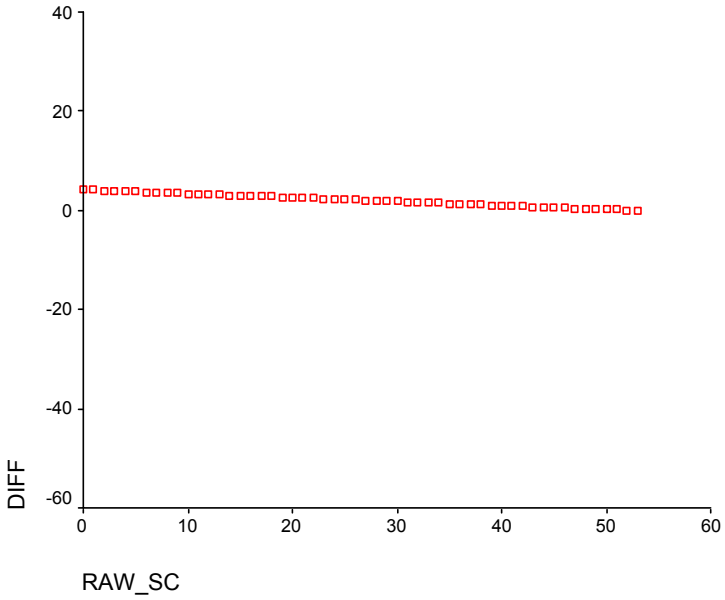


Figure 83

Spanish Language: Equating Functions of the Two Equating Chains for the Non-Verbal Section of Anchor Test Two

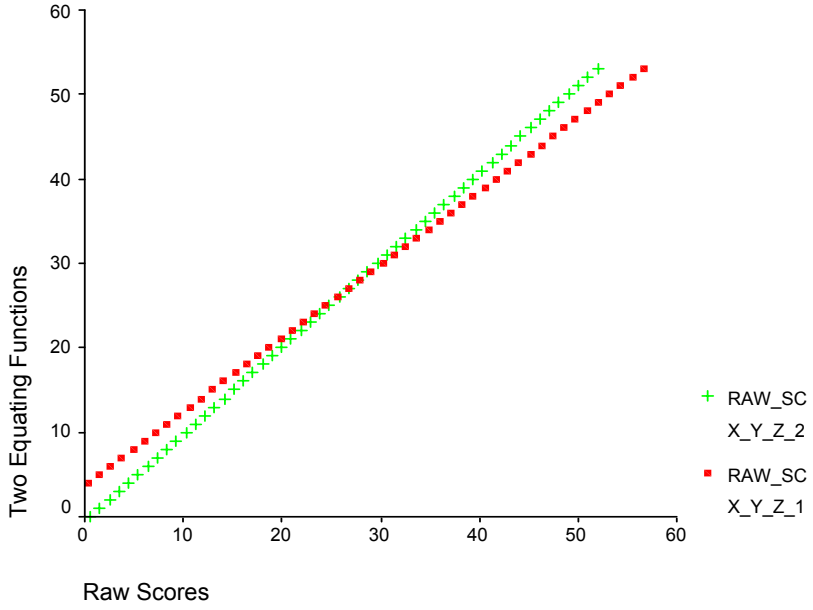


Figure 84

Spanish Language: the Differences between the Two Functions for Anchor Test Two

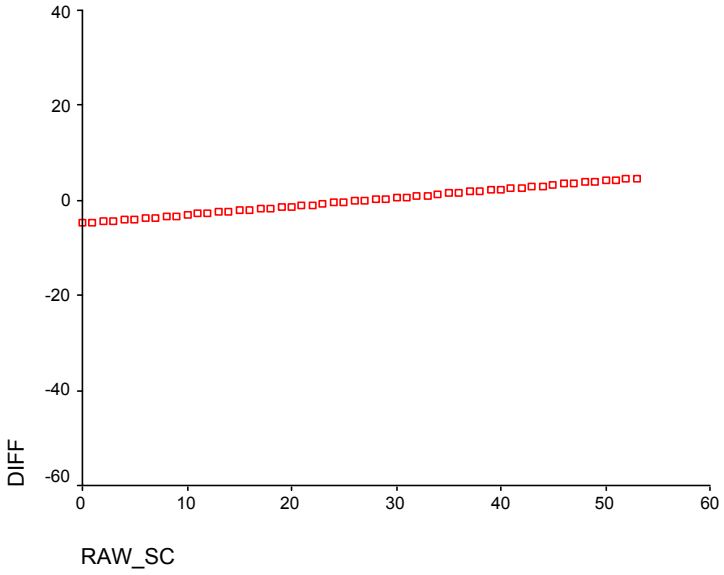


Figure 85

Chinese Language: Equating Functions of the Two Equating Chains for the Verbal Section of Anchor Test One

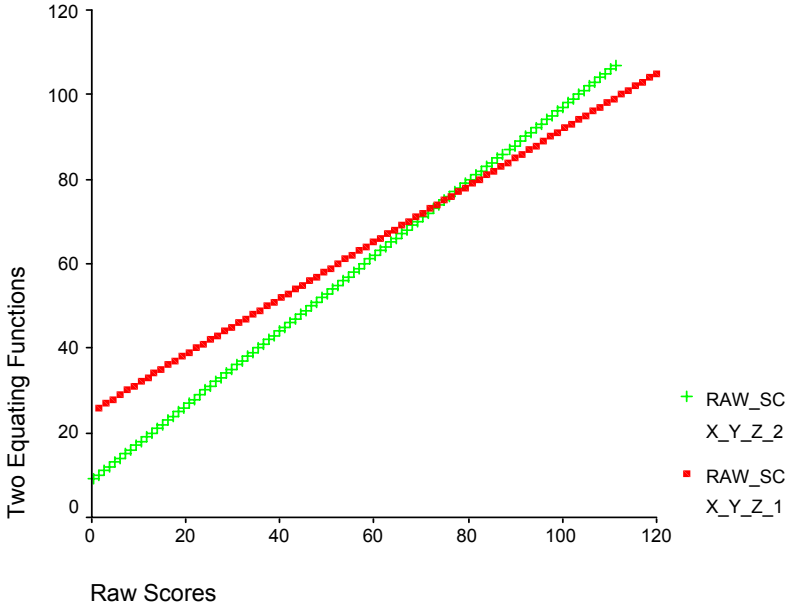


Figure 86

Chinese Language: the Differences between the Two Functions for Anchor Test One

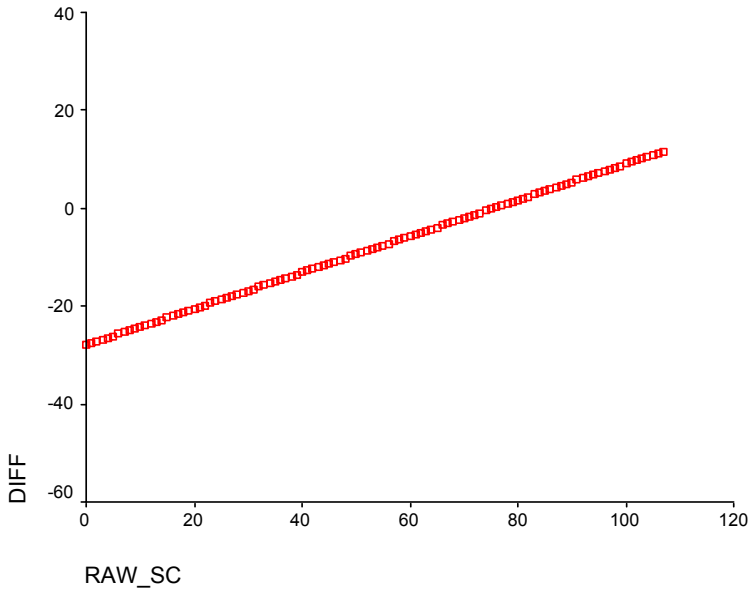


Figure 87

Chinese Language: Equating Functions of the Two Equating Chains for the Verbal Section of Anchor Test Two

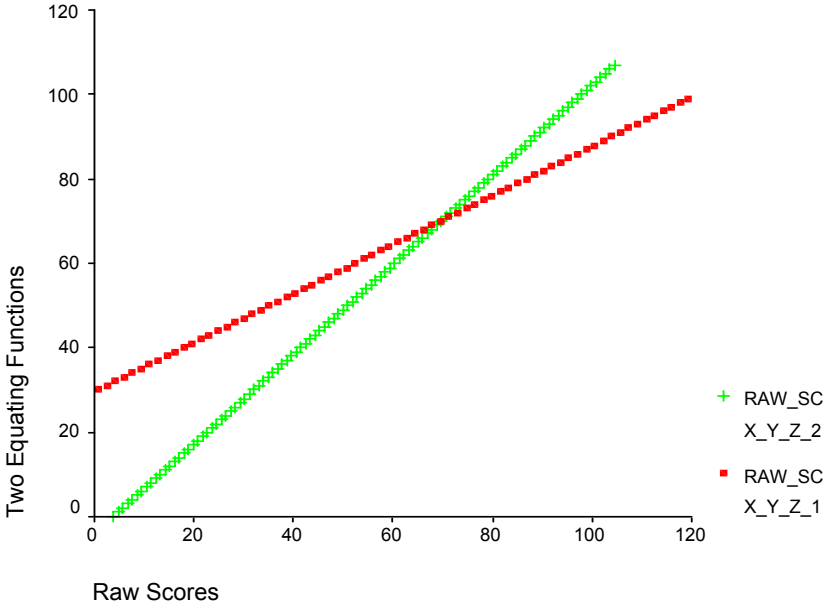


Figure 88

Chinese Language: the Differences between the Two Functions for Anchor Test Two

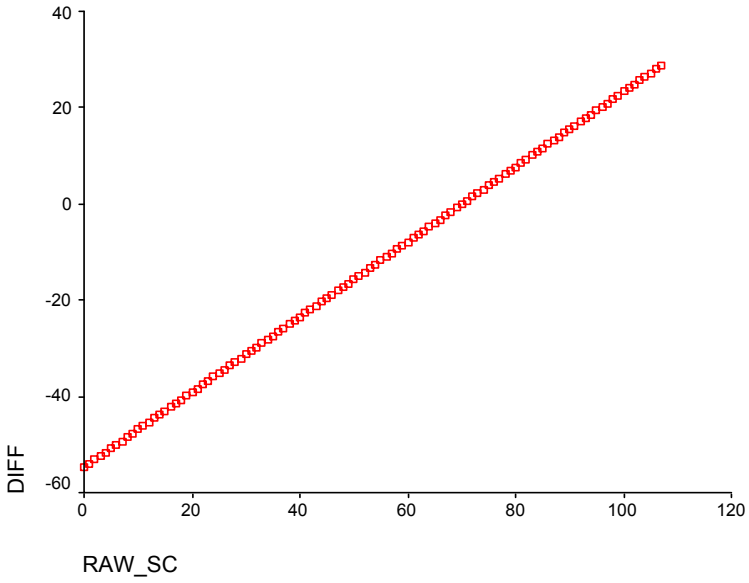


Figure 89

Chinese Language: Equating Functions of the Two Equating Chains for the Non-Verbal Section of Anchor Test One

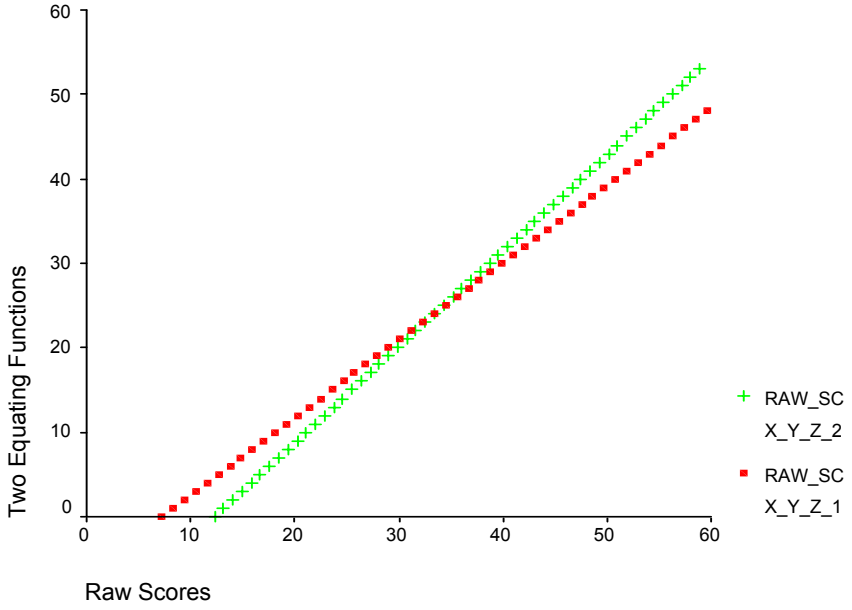


Figure 90

Chinese Language: the Differences between the Two Functions for Anchor Test One

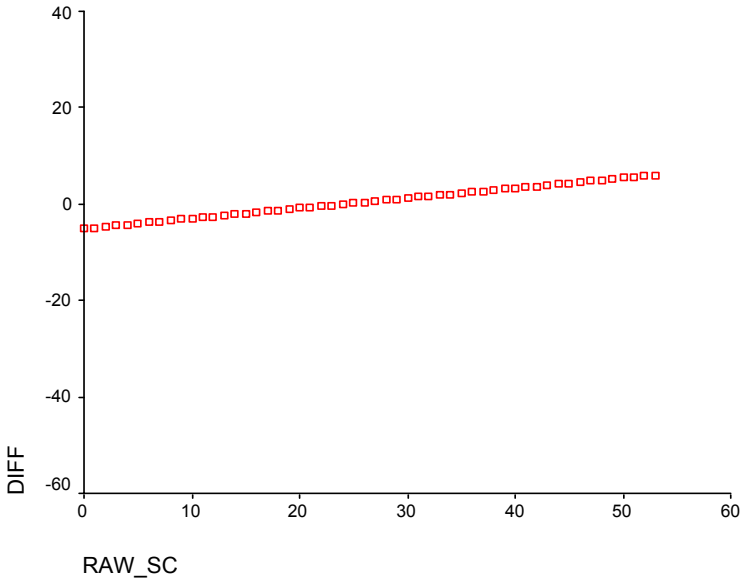


Figure 91

Chinese Language: Equating Functions of Two Equating Chains for Non-Verbal of Anchor Test Two

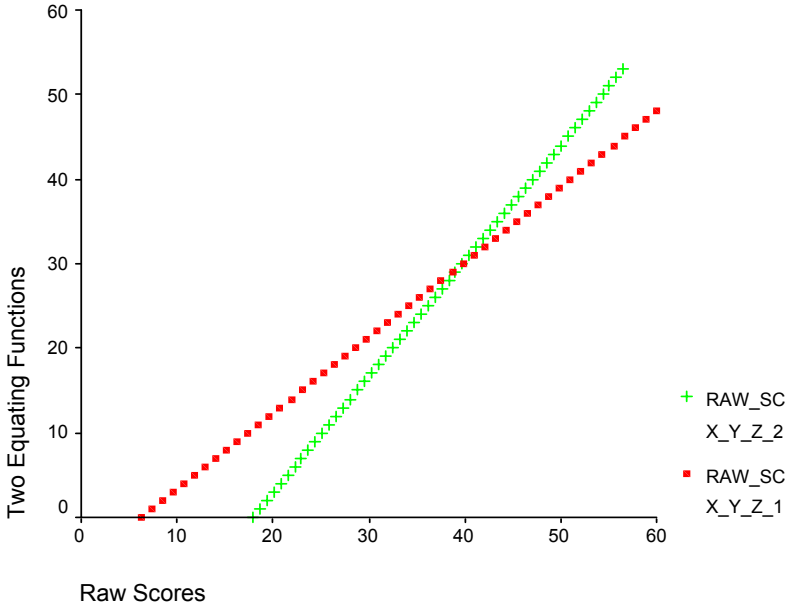
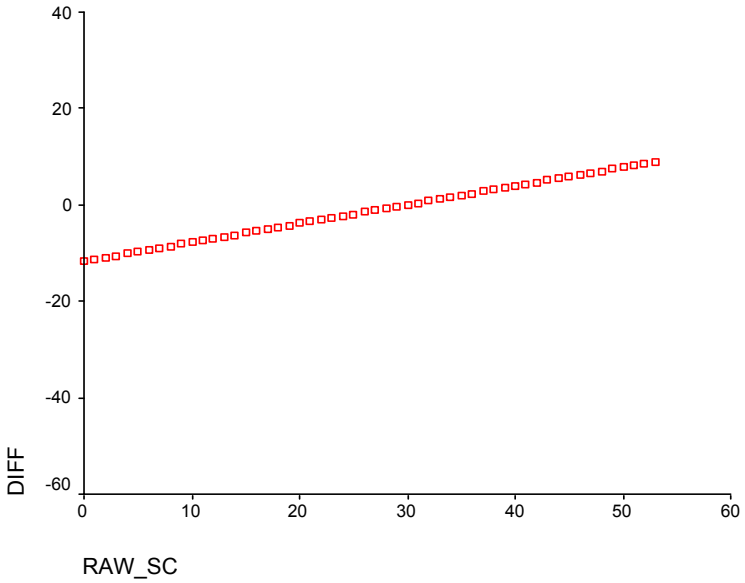


Figure 92

Chinese Language: Differences between the above Two Functions



APPENDIX K

Statistics for Standard Errors of Equating

Table 54

Korean Language: SEE for the Verbal Section of Two Anchor Tests

	Anchor Test One			Anchor Test Two		
Raw Scores	First Chain	Second Chain	Average SEE	First Chain	Second Chain	Average SEE
.00	7.90	16.08	11.99	11.93	13.95	12.94
1.00	7.80	15.88	11.84	11.78	13.78	12.78
2.00	7.70	15.67	11.69	11.63	13.61	12.62
3.00	7.60	15.47	11.54	11.48	13.43	12.45
4.00	7.50	15.27	11.39	11.32	13.26	12.29
5.00	7.40	15.07	11.23	11.17	13.08	12.13
6.00	7.30	14.87	11.08	11.02	12.91	11.97
7.00	7.20	14.66	10.93	10.87	12.74	11.80
8.00	7.10	14.46	10.78	10.72	12.56	11.64
9.00	7.00	14.26	10.63	10.57	12.39	11.48
10.00	6.90	14.06	10.48	10.42	12.21	11.31
11.00	6.81	13.86	10.33	10.26	12.04	11.15
12.00	6.71	13.66	10.18	10.11	11.87	10.99
13.00	6.61	13.46	10.03	9.96	11.69	10.83
14.00	6.51	13.26	9.88	9.81	11.52	10.67
15.00	6.41	13.05	9.73	9.66	11.35	10.50
16.00	6.31	12.85	9.58	9.51	11.18	10.34
17.00	6.21	12.65	9.43	9.36	11.00	10.18
18.00	6.11	12.45	9.28	9.21	10.83	10.02
19.00	6.01	12.25	9.13	9.06	10.66	9.86
20.00	5.91	12.05	8.98	8.91	10.49	9.70
21.00	5.81	11.85	8.83	8.75	10.31	9.53
22.00	5.71	11.65	8.68	8.60	10.14	9.37
23.00	5.61	11.45	8.53	8.45	9.97	9.21

24.00	5.52	11.25	8.38	8.30	9.80	9.05
25.00	5.42	11.05	8.23	8.15	9.63	8.89
26.00	5.32	10.85	8.08	8.00	9.45	8.73
27.00	5.22	10.65	7.94	7.85	9.28	8.57
28.00	5.12	10.45	7.79	7.70	9.11	8.41
29.00	5.02	10.25	7.64	7.55	8.94	8.25
30.00	4.92	10.06	7.49	7.40	8.77	8.09
31.00	4.82	9.86	7.34	7.25	8.60	7.93
32.00	4.73	9.66	7.19	7.10	8.43	7.77
33.00	4.63	9.46	7.04	6.95	8.26	7.61
34.00	4.53	9.26	6.90	6.80	8.09	7.45
35.00	4.43	9.07	6.75	6.65	7.92	7.29
36.00	4.33	8.87	6.60	6.51	7.76	7.13
37.00	4.23	8.67	6.45	6.36	7.59	6.97
38.00	4.14	8.48	6.31	6.21	7.42	6.81
39.00	4.04	8.28	6.16	6.06	7.25	6.66
40.00	3.94	8.08	6.01	5.91	7.09	6.50
41.00	3.84	7.89	5.87	5.76	6.92	6.34
42.00	3.75	7.69	5.72	5.62	6.75	6.18
43.00	3.65	7.50	5.57	5.47	6.59	6.03
44.00	3.55	7.31	5.43	5.32	6.42	5.87
45.00	3.45	7.11	5.28	5.17	6.26	5.72
46.00	3.36	6.92	5.14	5.03	6.10	5.56
47.00	3.26	6.73	4.99	4.88	5.93	5.41
48.00	3.16	6.54	4.85	4.73	5.77	5.25
49.00	3.07	6.35	4.71	4.59	5.61	5.10
50.00	2.97	6.16	4.57	4.44	5.45	4.95
51.00	2.88	5.97	4.42	4.30	5.29	4.80
52.00	2.78	5.78	4.28	4.16	5.14	4.65
53.00	2.69	5.60	4.14	4.01	4.98	4.50

54.00	2.59	5.41	4.00	3.87	4.82	4.35
55.00	2.50	5.23	3.86	3.73	4.67	4.20
56.00	2.41	5.05	3.73	3.59	4.52	4.05
57.00	2.31	4.87	3.59	3.45	4.37	3.91
58.00	2.22	4.69	3.45	3.31	4.22	3.77
59.00	2.13	4.51	3.32	3.17	4.08	3.63
60.00	2.04	4.34	3.19	3.03	3.94	3.49
61.00	1.95	4.17	3.06	2.90	3.80	3.35
62.00	1.86	4.00	2.93	2.77	3.66	3.21
63.00	1.77	3.83	2.80	2.64	3.53	3.08
64.00	1.69	3.67	2.68	2.51	3.40	2.95
65.00	1.60	3.52	2.56	2.38	3.28	2.83
66.00	1.52	3.37	2.44	2.26	3.16	2.71
67.00	1.44	3.22	2.33	2.14	3.05	2.59
68.00	1.37	3.08	2.22	2.02	2.94	2.48
69.00	1.29	2.95	2.12	1.91	2.84	2.37
70.00	1.22	2.82	2.02	1.80	2.75	2.28
71.00	1.16	2.71	1.93	1.71	2.66	2.19
72.00	1.10	2.61	1.85	1.62	2.59	2.10
73.00	1.04	2.51	1.78	1.54	2.53	2.03
74.00	1.00	2.43	1.72	1.47	2.47	1.97
75.00	.96	2.37	1.66	1.42	2.43	1.93
76.00	.93	2.32	1.62	1.38	2.41	1.89
77.00	.91	2.29	1.60	1.36	2.39	1.88
78.00	.90	2.27	1.59	1.36	2.39	1.87
79.00	.91	2.28	1.59	1.38	2.40	1.89
80.00	.92	2.30	1.61	1.42	2.42	1.92
81.00	.94	2.35	1.65	1.47	2.46	1.96
82.00	.98	2.41	1.69	1.54	2.50	2.02
83.00	1.02	2.49	1.75	1.62	2.56	2.09

84.00	1.07	2.58	1.83	1.71	2.63	2.17
85.00	1.13	2.68	1.91	1.81	2.71	2.26
86.00	1.19	2.80	2.00	1.92	2.80	2.36
87.00	1.26	2.93	2.09	2.03	2.90	2.46
88.00	1.33	3.06	2.20	2.14	3.00	2.57
89.00	1.41	3.21	2.31	2.26	3.11	2.69
90.00	1.49	3.35	2.42	2.39	3.22	2.81
91.00	1.57	3.51	2.54	2.52	3.35	2.93
92.00	1.65	3.67	2.66	2.64	3.47	3.06
93.00	1.73	3.83	2.78	2.78	3.60	3.19
94.00	1.82	4.00	2.91	2.91	3.74	3.32
95.00	1.91	4.17	3.04	3.04	3.87	3.46
96.00	2.00	4.34	3.17	3.18	4.01	3.60
97.00	2.09	4.52	3.30	3.32	4.16	3.74
98.00	2.18	4.70	3.44	3.46	4.30	3.88
99.00	2.27	4.88	3.57	3.60	4.45	4.03
100.00	2.36	5.06	3.71	3.74	4.60	4.17
101.00	2.46	5.24	3.85	3.88	4.76	4.32
102.00	2.55	5.43	3.99	4.02	4.91	4.47
103.00	2.64	5.61	4.13	4.17	5.06	4.62
104.00	2.74	5.80	4.27	4.31	5.22	4.77
105.00	2.83	5.99	4.41	4.46	5.38	4.92
106.00	2.93	6.18	4.55	4.60	5.54	5.07
107.00	3.02	6.37	4.70	4.75	5.70	5.22

Table 55

Korean Language: SEE for the Non-Verbal Section of Two Anchor Tests

	Anchor Test One			Anchor Test Two		
Raw Scores	First Chain	Second Chain	Average SEE	First Chain	Second Chain	Average SEE
.00	3.94	5.33	4.64	3.92	6.07	4.99
1.00	3.84	5.19	4.51	3.82	5.90	4.86
2.00	3.73	5.04	4.39	3.71	5.73	4.72
3.00	3.63	4.90	4.26	3.61	5.57	4.59
4.00	3.53	4.75	4.14	3.50	5.40	4.45
5.00	3.42	4.61	4.02	3.40	5.23	4.32
6.00	3.32	4.47	3.89	3.29	5.07	4.18
7.00	3.21	4.32	3.77	3.19	4.90	4.05
8.00	3.11	4.18	3.65	3.08	4.74	3.91
9.00	3.01	4.04	3.52	2.98	4.58	3.78
10.00	2.91	3.90	3.40	2.88	4.41	3.64
11.00	2.81	3.76	3.28	2.77	4.25	3.51
12.00	2.70	3.62	3.16	2.67	4.09	3.38
13.00	2.60	3.48	3.04	2.57	3.93	3.25
14.00	2.50	3.34	2.92	2.47	3.77	3.12
15.00	2.40	3.20	2.80	2.37	3.61	2.99
16.00	2.31	3.07	2.69	2.27	3.45	2.86
17.00	2.21	2.93	2.57	2.17	3.30	2.73
18.00	2.11	2.80	2.46	2.07	3.14	2.61
19.00	2.02	2.67	2.34	1.98	2.99	2.48
20.00	1.92	2.54	2.23	1.88	2.84	2.36
21.00	1.83	2.41	2.12	1.79	2.70	2.24
22.00	1.74	2.29	2.02	1.70	2.55	2.12
23.00	1.65	2.17	1.91	1.61	2.41	2.01

24.00	1.57	2.05	1.81	1.52	2.28	1.90
25.00	1.49	1.94	1.71	1.44	2.15	1.79
26.00	1.41	1.84	1.62	1.35	2.02	1.69
27.00	1.33	1.74	1.53	1.28	1.91	1.59
28.00	1.26	1.65	1.45	1.20	1.80	1.50
29.00	1.19	1.57	1.38	1.13	1.70	1.42
30.00	1.13	1.50	1.31	1.07	1.62	1.35
31.00	1.08	1.44	1.26	1.01	1.56	1.29
32.00	1.03	1.40	1.21	.96	1.51	1.24
33.00	.99	1.38	1.18	.92	1.48	1.20
34.00	.95	1.37	1.16	.89	1.48	1.18
35.00	.93	1.38	1.15	.87	1.49	1.18
36.00	.92	1.40	1.16	.87	1.52	1.19
37.00	.93	1.44	1.18	.88	1.57	1.22
38.00	.94	1.49	1.22	.90	1.64	1.27
39.00	.97	1.56	1.27	.94	1.72	1.33
40.00	1.02	1.63	1.32	.99	1.81	1.40
41.00	1.07	1.72	1.39	1.05	1.92	1.48
42.00	1.13	1.81	1.47	1.12	2.03	1.57
43.00	1.20	1.91	1.56	1.19	2.15	1.67
44.00	1.27	2.02	1.65	1.27	2.27	1.77
45.00	1.35	2.13	1.74	1.36	2.41	1.88
46.00	1.44	2.25	1.84	1.45	2.54	1.99
47.00	1.53	2.37	1.95	1.54	2.68	2.11
48.00	1.62	2.49	2.05	1.63	2.83	2.23
49.00	1.71	2.62	2.16	1.73	2.97	2.35
50.00	1.80	2.75	2.27	1.82	3.12	2.47
51.00	1.90	2.88	2.39	1.92	3.28	2.60
52.00	2.00	3.01	2.50	2.02	3.43	2.73
53.00	2.09	3.14	2.62	2.12	3.58	2.85

Table 56

Spanish Language: SEE for the Verbal Section of Two Anchor Tests

	Anchor Test One			Anchor Test Two		
Raw Scores	First Chain	Second Chain	Average SEE	First Chain	Second Chain	Average SEE
.00	7.55	11.63	9.59	8.47	12.97	10.72
1.00	7.45	11.48	9.46	8.36	12.80	10.58
2.00	7.35	11.32	9.34	8.25	12.63	10.44
3.00	7.25	11.17	9.21	8.13	12.45	10.29
4.00	7.15	11.01	9.08	8.02	12.28	10.15
5.00	7.06	10.85	8.95	7.91	12.11	10.01
6.00	6.96	10.70	8.83	7.80	11.94	9.87
7.00	6.86	10.54	8.70	7.69	11.77	9.73
8.00	6.76	10.39	8.57	7.57	11.60	9.59
9.00	6.66	10.23	8.45	7.46	11.43	9.44
10.00	6.57	10.08	8.32	7.35	11.26	9.30
11.00	6.47	9.92	8.19	7.24	11.09	9.16
12.00	6.37	9.77	8.07	7.13	10.92	9.02
13.00	6.27	9.61	7.94	7.02	10.74	8.88
14.00	6.18	9.46	7.82	6.91	10.57	8.74
15.00	6.08	9.30	7.69	6.79	10.40	8.60
16.00	5.98	9.15	7.56	6.68	10.24	8.46
17.00	5.88	8.99	7.44	6.57	10.07	8.32
18.00	5.79	8.84	7.31	6.46	9.90	8.18
19.00	5.69	8.69	7.19	6.35	9.73	8.04
20.00	5.59	8.53	7.06	6.24	9.56	7.90
21.00	5.49	8.38	6.94	6.13	9.39	7.76
22.00	5.40	8.23	6.81	6.02	9.22	7.62
23.00	5.30	8.08	6.69	5.91	9.05	7.48

24.00	5.20	7.92	6.56	5.80	8.89	7.34
25.00	5.11	7.77	6.44	5.69	8.72	7.20
26.00	5.01	7.62	6.31	5.58	8.55	7.06
27.00	4.91	7.47	6.19	5.47	8.38	6.93
28.00	4.82	7.32	6.07	5.36	8.22	6.79
29.00	4.72	7.17	5.94	5.25	8.05	6.65
30.00	4.62	7.02	5.82	5.14	7.89	6.51
31.00	4.53	6.87	5.70	5.03	7.72	6.37
32.00	4.43	6.72	5.58	4.92	7.56	6.24
33.00	4.34	6.57	5.45	4.81	7.39	6.10
34.00	4.24	6.42	5.33	4.70	7.23	5.96
35.00	4.15	6.27	5.21	4.59	7.07	5.83
36.00	4.05	6.13	5.09	4.48	6.90	5.69
37.00	3.96	5.98	4.97	4.37	6.74	5.56
38.00	3.86	5.83	4.85	4.26	6.58	5.42
39.00	3.77	5.69	4.73	4.16	6.42	5.29
40.00	3.67	5.54	4.61	4.05	6.26	5.16
41.00	3.58	5.40	4.49	3.94	6.10	5.02
42.00	3.49	5.26	4.37	3.84	5.95	4.89
43.00	3.39	5.12	4.25	3.73	5.79	4.76
44.00	3.30	4.98	4.14	3.62	5.63	4.63
45.00	3.21	4.84	4.02	3.52	5.48	4.50
46.00	3.11	4.70	3.91	3.41	5.33	4.37
47.00	3.02	4.56	3.79	3.31	5.18	4.24
48.00	2.93	4.43	3.68	3.20	5.03	4.11
49.00	2.84	4.29	3.57	3.10	4.88	3.99
50.00	2.75	4.16	3.46	3.00	4.73	3.86
51.00	2.66	4.03	3.35	2.90	4.59	3.74
52.00	2.57	3.91	3.24	2.79	4.45	3.62
53.00	2.49	3.78	3.13	2.69	4.31	3.50

54.00	2.40	3.66	3.03	2.60	4.17	3.38
55.00	2.31	3.54	2.93	2.50	4.04	3.27
56.00	2.23	3.43	2.83	2.40	3.91	3.15
57.00	2.14	3.31	2.73	2.31	3.78	3.04
58.00	2.06	3.21	2.63	2.21	3.66	2.93
59.00	1.98	3.10	2.54	2.12	3.54	2.83
60.00	1.90	3.00	2.45	2.03	3.42	2.73
61.00	1.82	2.91	2.37	1.95	3.32	2.63
62.00	1.75	2.82	2.29	1.86	3.21	2.54
63.00	1.68	2.74	2.21	1.78	3.12	2.45
64.00	1.61	2.67	2.14	1.71	3.03	2.37
65.00	1.54	2.60	2.07	1.63	2.95	2.29
66.00	1.48	2.54	2.01	1.57	2.87	2.22
67.00	1.42	2.49	1.96	1.50	2.81	2.16
68.00	1.37	2.45	1.91	1.45	2.75	2.10
69.00	1.32	2.42	1.87	1.40	2.71	2.05
70.00	1.27	2.41	1.84	1.36	2.67	2.02
71.00	1.24	2.40	1.82	1.33	2.65	1.99
72.00	1.21	2.40	1.81	1.31	2.64	1.97
73.00	1.19	2.41	1.80	1.30	2.64	1.97
74.00	1.18	2.44	1.81	1.30	2.65	1.97
75.00	1.18	2.47	1.83	1.31	2.67	1.99
76.00	1.18	2.52	1.85	1.33	2.71	2.02
77.00	1.20	2.57	1.89	1.35	2.75	2.05
78.00	1.22	2.64	1.93	1.39	2.81	2.10
79.00	1.25	2.71	1.98	1.44	2.88	2.16
80.00	1.29	2.79	2.04	1.49	2.95	2.22
81.00	1.33	2.87	2.10	1.55	3.04	2.29
82.00	1.38	2.96	2.17	1.61	3.13	2.37
83.00	1.44	3.06	2.25	1.68	3.23	2.46

84.00	1.50	3.16	2.33	1.76	3.33	2.55
85.00	1.56	3.27	2.41	1.84	3.45	2.64
86.00	1.63	3.38	2.50	1.92	3.56	2.74
87.00	1.70	3.49	2.59	2.00	3.69	2.84
88.00	1.77	3.61	2.69	2.09	3.81	2.95
89.00	1.84	3.73	2.79	2.18	3.94	3.06
90.00	1.92	3.86	2.89	2.27	4.07	3.17
91.00	2.00	3.98	2.99	2.36	4.21	3.29
92.00	2.08	4.11	3.10	2.46	4.35	3.40
93.00	2.16	4.24	3.20	2.55	4.49	3.52
94.00	2.25	4.37	3.31	2.65	4.63	3.64
95.00	2.33	4.51	3.42	2.75	4.78	3.77
96.00	2.42	4.64	3.53	2.85	4.93	3.89
97.00	2.50	4.78	3.64	2.95	5.08	4.02
98.00	2.59	4.92	3.76	3.05	5.23	4.14
99.00	2.68	5.06	3.87	3.16	5.38	4.27
100.00	2.77	5.20	3.99	3.26	5.54	4.40
101.00	2.86	5.34	4.10	3.36	5.69	4.53
102.00	2.95	5.49	4.22	3.47	5.85	4.66
103.00	3.04	5.63	4.34	3.57	6.01	4.79
104.00	3.13	5.78	4.45	3.68	6.17	4.92
105.00	3.23	5.92	4.57	3.78	6.32	5.05
106.00	3.32	6.07	4.69	3.89	6.48	5.19
107.00	3.41	6.21	4.81	4.00	6.65	5.32

Table 57

Spanish Language: SEE for the Non-Verbal Section of Two Anchor Tests

	Anchor Test One			Anchor Test Two		
Raw Scores	First Chain	Second Chain	Average SEE	First Chain	Second Chain	Average SEE
.00	3.83	5.20	4.51	3.69	5.67	4.68
1.00	3.72	5.05	4.39	3.59	5.51	4.55
2.00	3.61	4.91	4.26	3.48	5.35	4.42
3.00	3.50	4.76	4.13	3.38	5.18	4.28
4.00	3.39	4.61	4.00	3.28	5.02	4.15
5.00	3.28	4.47	3.88	3.17	4.86	4.02
6.00	3.18	4.33	3.75	3.07	4.70	3.89
7.00	3.07	4.18	3.63	2.96	4.55	3.75
8.00	2.96	4.04	3.50	2.86	4.39	3.62
9.00	2.85	3.90	3.38	2.76	4.23	3.50
10.00	2.74	3.76	3.25	2.66	4.08	3.37
11.00	2.64	3.62	3.13	2.56	3.92	3.24
12.00	2.53	3.48	3.01	2.46	3.77	3.11
13.00	2.42	3.35	2.89	2.36	3.62	2.99
14.00	2.32	3.21	2.77	2.26	3.47	2.87
15.00	2.21	3.08	2.65	2.16	3.32	2.74
16.00	2.11	2.95	2.53	2.07	3.18	2.62
17.00	2.01	2.82	2.42	1.97	3.04	2.51
18.00	1.91	2.70	2.30	1.88	2.90	2.39
19.00	1.80	2.58	2.19	1.79	2.77	2.28
20.00	1.70	2.46	2.08	1.70	2.64	2.17
21.00	1.61	2.35	1.98	1.62	2.52	2.07
22.00	1.51	2.24	1.88	1.54	2.40	1.97
23.00	1.42	2.14	1.78	1.46	2.29	1.87

24.00	1.33	2.04	1.68	1.38	2.18	1.78
25.00	1.24	1.96	1.60	1.31	2.09	1.70
26.00	1.15	1.88	1.51	1.25	2.00	1.63
27.00	1.07	1.81	1.44	1.19	1.93	1.56
28.00	1.00	1.75	1.37	1.15	1.87	1.51
29.00	.94	1.70	1.32	1.11	1.83	1.47
30.00	.88	1.67	1.27	1.08	1.80	1.44
31.00	.84	1.65	1.24	1.06	1.79	1.42
32.00	.81	1.64	1.22	1.05	1.79	1.42
33.00	.79	1.65	1.22	1.06	1.81	1.43
34.00	.80	1.68	1.24	1.07	1.85	1.46
35.00	.82	1.72	1.27	1.10	1.90	1.50
36.00	.85	1.77	1.31	1.14	1.97	1.55
37.00	.90	1.83	1.37	1.18	2.05	1.61
38.00	.96	1.91	1.43	1.24	2.14	1.69
39.00	1.03	1.99	1.51	1.30	2.23	1.77
40.00	1.11	2.08	1.59	1.36	2.34	1.85
41.00	1.19	2.18	1.68	1.44	2.46	1.95
42.00	1.27	2.29	1.78	1.51	2.58	2.05
43.00	1.36	2.40	1.88	1.59	2.70	2.15
44.00	1.46	2.51	1.98	1.68	2.84	2.26
45.00	1.55	2.63	2.09	1.76	2.97	2.37
46.00	1.65	2.76	2.20	1.85	3.11	2.48
47.00	1.75	2.88	2.32	1.94	3.25	2.60
48.00	1.85	3.01	2.43	2.04	3.40	2.72
49.00	1.95	3.14	2.55	2.13	3.54	2.84
50.00	2.05	3.28	2.66	2.23	3.69	2.96
51.00	2.16	3.41	2.78	2.33	3.85	3.09
52.00	2.26	3.55	2.90	2.42	4.00	3.21
53.00	2.36	3.69	3.03	2.52	4.15	3.34

Table 58

Chinese Language: SEE for the Verbal Section of Two Anchor Tests

	Anchor Test One			Anchor Test Two		
Raw Scores	First Chain	Second Chain	Average SEE	First Chain	Second Chain	Average SEE
.00	5.18	10.59	7.89	8.61	10.82	7.44
1.00	5.10	10.44	7.77	8.49	10.66	7.33
2.00	5.03	10.28	7.65	8.36	10.50	7.22
3.00	4.95	10.12	7.54	8.24	10.35	7.11
4.00	4.88	9.97	7.42	8.11	10.19	7.00
5.00	4.80	9.81	7.31	7.98	10.03	6.90
6.00	4.73	9.66	7.19	7.86	9.87	6.79
7.00	4.65	9.50	7.08	7.73	9.71	6.68
8.00	4.57	9.35	6.96	7.61	9.55	6.57
9.00	4.50	9.19	6.84	7.48	9.39	6.46
10.00	4.42	9.04	6.73	7.36	9.24	6.35
11.00	4.35	8.88	6.61	7.23	9.08	6.24
12.00	4.27	8.73	6.50	7.11	8.92	6.14
13.00	4.20	8.57	6.39	6.98	8.76	6.03
14.00	4.12	8.42	6.27	6.86	8.61	5.92
15.00	4.05	8.26	6.16	6.73	8.45	5.81
16.00	3.97	8.11	6.04	6.61	8.29	5.71
17.00	3.90	7.96	5.93	6.48	8.14	5.60
18.00	3.82	7.80	5.81	6.36	7.98	5.49
19.00	3.75	7.65	5.70	6.23	7.83	5.39
20.00	3.68	7.50	5.59	6.11	7.67	5.28
21.00	3.60	7.35	5.47	5.98	7.51	5.17
22.00	3.53	7.19	5.36	5.86	7.36	5.07

23.00	3.45	7.04	5.25	5.74	7.21	4.96
24.00	3.38	6.89	5.14	5.61	7.05	4.85
25.00	3.31	6.74	5.02	5.49	6.90	4.75
26.00	3.23	6.59	4.91	5.37	6.75	4.64
27.00	3.16	6.44	4.80	5.24	6.59	4.54
28.00	3.09	6.29	4.69	5.12	6.44	4.43
29.00	3.01	6.14	4.58	5.00	6.29	4.33
30.00	2.94	5.99	4.47	4.88	6.14	4.22
31.00	2.87	5.85	4.36	4.75	5.99	4.12
32.00	2.80	5.70	4.25	4.63	5.84	4.02
33.00	2.72	5.55	4.14	4.51	5.69	3.92
34.00	2.65	5.41	4.03	4.39	5.54	3.81
35.00	2.58	5.26	3.92	4.27	5.39	3.71
36.00	2.51	5.12	3.82	4.15	5.25	3.61
37.00	2.44	4.98	3.71	4.03	5.10	3.51
38.00	2.37	4.84	3.60	3.91	4.96	3.41
39.00	2.30	4.70	3.50	3.79	4.82	3.31
40.00	2.23	4.56	3.39	3.67	4.67	3.21
41.00	2.16	4.42	3.29	3.55	4.53	3.11
42.00	2.09	4.29	3.19	3.44	4.40	3.02
43.00	2.03	4.15	3.09	3.32	4.26	2.92
44.00	1.96	4.02	2.99	3.20	4.12	2.83
45.00	1.89	3.89	2.89	3.09	3.99	2.74
46.00	1.83	3.76	2.80	2.98	3.86	2.64
47.00	1.77	3.64	2.70	2.86	3.73	2.55
48.00	1.70	3.51	2.61	2.75	3.61	2.46
49.00	1.64	3.39	2.52	2.64	3.48	2.38
50.00	1.58	3.28	2.43	2.54	3.37	2.29
51.00	1.52	3.16	2.34	2.43	3.25	2.21
52.00	1.47	3.06	2.26	2.33	3.14	2.13

53.00	1.41	2.95	2.18	2.22	3.03	2.05
54.00	1.36	2.85	2.11	2.13	2.93	1.98
55.00	1.31	2.76	2.03	2.03	2.83	1.91
56.00	1.26	2.67	1.97	1.94	2.74	1.84
57.00	1.22	2.59	1.90	1.85	2.66	1.78
58.00	1.17	2.52	1.85	1.77	2.59	1.72
59.00	1.14	2.45	1.80	1.69	2.52	1.67
60.00	1.10	2.40	1.75	1.62	2.46	1.62
61.00	1.08	2.35	1.71	1.56	2.41	1.58
62.00	1.05	2.31	1.68	1.50	2.37	1.55
63.00	1.03	2.29	1.66	1.45	2.34	1.52
64.00	1.02	2.27	1.65	1.42	2.32	1.50
65.00	1.01	2.27	1.64	1.39	2.32	1.49
66.00	1.01	2.28	1.65	1.38	2.32	1.49
67.00	1.02	2.30	1.66	1.38	2.34	1.50
68.00	1.03	2.33	1.68	1.39	2.37	1.52
69.00	1.05	2.38	1.71	1.41	2.41	1.54
70.00	1.07	2.43	1.75	1.45	2.47	1.58
71.00	1.10	2.50	1.80	1.49	2.53	1.62
72.00	1.13	2.57	1.85	1.55	2.60	1.66
73.00	1.16	2.65	1.91	1.61	2.68	1.72
74.00	1.20	2.74	1.97	1.68	2.77	1.78
75.00	1.25	2.83	2.04	1.76	2.86	1.84
76.00	1.29	2.93	2.11	1.84	2.96	1.91
77.00	1.34	3.03	2.19	1.93	3.07	1.98
78.00	1.40	3.14	2.27	2.02	3.18	2.06
79.00	1.45	3.26	2.35	2.12	3.29	2.14
80.00	1.51	3.38	2.44	2.22	3.41	2.22
81.00	1.56	3.50	2.53	2.32	3.53	2.31
82.00	1.62	3.62	2.62	2.42	3.66	2.39

83.00	1.68	3.75	2.72	2.53	3.79	2.48
84.00	1.75	3.88	2.81	2.63	3.92	2.57
85.00	1.81	4.01	2.91	2.74	4.05	2.66
86.00	1.87	4.14	3.01	2.85	4.19	2.76
87.00	1.94	4.28	3.11	2.97	4.32	2.85
88.00	2.01	4.41	3.21	3.08	4.46	2.95
89.00	2.07	4.55	3.31	3.19	4.60	3.04
90.00	2.14	4.69	3.42	3.31	4.75	3.14
91.00	2.21	4.83	3.52	3.43	4.89	3.24
92.00	2.28	4.97	3.63	3.54	5.03	3.34
93.00	2.35	5.12	3.73	3.66	5.18	3.44
94.00	2.42	5.26	3.84	3.78	5.33	3.54
95.00	2.49	5.41	3.95	3.90	5.47	3.64
96.00	2.56	5.55	4.05	4.02	5.62	3.74
97.00	2.63	5.70	4.16	4.14	5.77	3.85
98.00	2.70	5.85	4.27	4.26	5.92	3.95
99.00	2.77	5.99	4.38	4.38	6.07	4.05
100.00	2.84	6.14	4.49	4.50	6.22	4.16
101.00	2.92	6.29	4.60	4.62	6.37	4.26
102.00	2.99	6.44	4.71	4.74	6.53	4.37
103.00	3.06	6.59	4.83	4.87	6.68	4.47
104.00	3.13	6.74	4.94	4.99	6.83	4.58
105.00	3.21	6.89	5.05	5.11	6.99	4.68
106.00	3.28	7.05	5.17	5.23	7.14	4.79
107.00	3.35	7.20	5.28	5.36	7.29	4.89

Table 59

Chinese Language: SEE for the Non-Verbal Section of Two Anchor Tests

	Anchor Test One			Anchor Test Two		
Raw Scores	First Chain	Second Chain	Average SEE	First Chain	Second Chain	Average SEE
.00	2.29	5.76	4.02	2.22	5.24	3.73
1.00	2.21	5.60	3.91	2.15	5.10	3.62
2.00	2.14	5.45	3.79	2.08	4.96	3.52
3.00	2.07	5.30	3.68	2.01	4.82	3.42
4.00	1.99	5.15	3.57	1.94	4.68	3.31
5.00	1.92	4.99	3.46	1.87	4.55	3.21
6.00	1.85	4.84	3.35	1.80	4.41	3.11
7.00	1.78	4.69	3.24	1.74	4.27	3.00
8.00	1.70	4.54	3.12	1.67	4.14	2.90
9.00	1.63	4.40	3.02	1.60	4.00	2.80
10.00	1.56	4.25	2.91	1.54	3.87	2.70
11.00	1.49	4.10	2.80	1.47	3.73	2.60
12.00	1.43	3.96	2.69	1.41	3.60	2.51
13.00	1.36	3.81	2.59	1.35	3.47	2.41
14.00	1.29	3.67	2.48	1.29	3.34	2.32
15.00	1.23	3.53	2.38	1.23	3.21	2.22
16.00	1.16	3.39	2.28	1.17	3.09	2.13
17.00	1.10	3.25	2.18	1.12	2.97	2.04
18.00	1.04	3.12	2.08	1.07	2.84	1.96
19.00	.99	2.98	1.99	1.02	2.73	1.87
20.00	.93	2.85	1.89	.97	2.61	1.79
21.00	.88	2.73	1.81	.93	2.50	1.71
22.00	.84	2.61	1.72	.89	2.39	1.64
23.00	.80	2.49	1.64	.86	2.29	1.57

24.00	.76	2.37	1.57	.83	2.19	1.51
25.00	.73	2.26	1.50	.81	2.09	1.45
26.00	.71	2.16	1.43	.79	2.00	1.40
27.00	.69	2.06	1.38	.78	1.92	1.35
28.00	.69	1.96	1.33	.78	1.84	1.31
29.00	.69	1.88	1.28	.79	1.77	1.28
30.00	.71	1.79	1.25	.80	1.71	1.25
31.00	.73	1.72	1.22	.82	1.65	1.23
32.00	.76	1.65	1.20	.85	1.60	1.22
33.00	.79	1.59	1.19	.88	1.56	1.22
34.00	.83	1.53	1.18	.91	1.52	1.22
35.00	.88	1.49	1.18	.96	1.50	1.23
36.00	.93	1.45	1.19	1.00	1.49	1.24
37.00	.98	1.43	1.21	1.05	1.49	1.27
38.00	1.04	1.43	1.23	1.10	1.50	1.30
39.00	1.10	1.45	1.27	1.15	1.53	1.34
40.00	1.16	1.49	1.32	1.21	1.58	1.39
41.00	1.23	1.55	1.39	1.27	1.64	1.45
42.00	1.29	1.64	1.46	1.33	1.72	1.52
43.00	1.36	1.74	1.55	1.39	1.80	1.60
44.00	1.42	1.86	1.64	1.45	1.90	1.68
45.00	1.49	1.98	1.74	1.52	2.01	1.76
46.00	1.56	2.11	1.83	1.58	2.12	1.85
47.00	1.63	2.24	1.94	1.65	2.23	1.94
48.00	1.70	2.38	2.04	1.71	2.36	2.03
49.00	1.77	2.52	2.15	1.78	2.48	2.13
50.00	1.85	2.66	2.25	1.85	2.61	2.23
51.00	1.92	2.81	2.36	1.92	2.74	2.33
52.00	1.99	2.95	2.47	1.99	2.87	2.43
53.00	2.06	3.10	2.58	2.06	3.00	2.53

APPENDIX L

Abstract

ABSTRACT

The Effect of Different Anchor Selection Approaches on the Accuracy of Test Equating for Test Adaptation

Hua Gao

Educational Research and Evaluation

Director of Dissertation: George Johanson, Ed.D. Professor

The focus of this study was to evaluate the effect of different approaches of anchor test construction on the accuracy of equating for test adaptation. The term "equating" in cross-lingual studies refers to a statistical procedure that adjusts test scores from the source language (SL) version of the test and the target language (TL) version of the test using a set of common translated items of the same examination so that scores can be interpreted interchangeably. In each test, the verbal section and the non-verbal section of the test were investigated. The Levine Linear equating method and Mean-Sigma equating method were utilized with an anchor item design and an equivalent group design, respectively. The double linking method and the standard errors of equating method were used to evaluate the accuracy of the equating for different anchor tests. The average difference between the two anchor tests for the verbal and non-verbal sections of the test over three target language groups reflected the degree of overall instability that existed in the cross-lingual equating process. These differences were associated with real and systematic variance that underlies the cross-lingual equating process. Scoring outcomes of an actual certification examination with a sample of nearly 9,000 examinees taking both SL and TL versions of the test data set were utilized for this research study.

Findings indicated that the differences between the double linking chains for each anchor test were greater for the verbal section than the non-verbal section of the test. The results of the double linking method supported the notion that different choices for anchor items can result in different equatings and using items with the more stable parameters was a better choice than using items with less DIF. The results of MSEE did not show large differences between the parameter and the DIF methods of anchor item selection. However, the MSEE differences were in the same direction as the double linking method differences. That is, the parameter method was superior to the DIF method using both criteria.