This is to certify that the
dissertation entitled

A COMPARISON BETWEEN THE VERTICAL SCALING
OF TESTS SENSITIVE TO MULTIPLE DIMENSIONS
USING COMMON-ITEM AND COMMON-GROUP DESIGNS

presented by

Jing Yu

has been accepted towards fulfillment
of the requirements for the

Ph.D degree in Measurement and Quantitative Methods

Mark D. Reckase
Major Professor's Signature

3/28/07
Date

MSU is an affirmative-action, equal-opportunity employer

A COMPARISON BETWEEN THE VERTICAL SCALING OF TESTS SENSITIVE TO MULTIPLE DIMENSIONS USING COMMON-ITEM AND COMMON-GROUP DESIGNS

By

Jing Yu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COUNSELING, EDUCATIONAL PSYCHOLOGY AND SPECIAL EDUCATION

2007

Abstract

A COMPARISON BETWEEN THE VERTICAL SCALING
OF TESTS SENSITIVE TO MULTIPLE DIMENSIONS
USING COMMON-ITEM AND COMMON-GROUP DESIGNS

by

Jing Yu

Three methods of item response theory (IRT) linking—common-item, common-group and a combination of common-item and common-group (referred to as common-common) linking designs were compared using real testing data from an English as second language (ESL) exam program. The methods were considered as "vertical scaling" instead of "equating" because, first, the test was designed to examine three different traits of English ability; multidimensional IRT and factor analysis on testing data confirms that the test was multidimensional. Second, the two test forms are not at the same difficulty level, the averaged difficulty parameters were different by about 0.5, 1.0 or 1.5 standard units, thus the linking was considered vertical. The effects of test length and averaged difficulty level differences were also analyzed. For practical reasons, the anchor test used in the common-item linking design could not represent all the dimensions of the test forms.

The original data contained dichotomous responses from about 30,000 individuals on 130 items. For the evaluation of each linking design, a sub-sample of cases and responses were selected. The linking designs were evaluated by calculating the standard error of equating and by comparing the examinees' scores and item parameters before vs. after equating. Results of the analyses indicate that common-group and common-common linking designs can serve as adequate alternatives to the well-recognized common-item design. Longer test forms work better for item parameter estimation and have smaller standard errors of equating. When the ability of the group does not match the difficulty level of the assigned

form, the common-item design has a slightly smaller standard error of equating than the common-group and common-common designs.

ACKNOWLEDGMENTS

First, I want to give all the glory and honor to my Lord, Jesus Christ. He hears prayers, and has led me through each step of my PhD study.

My special appreciation goes to my academic advisor, Dr. Mark D. Reckase, I enjoyed being his advisee in the past four years. As one of the most knowledgeable and respected expert in our field today, not only his knowledge and skills in psychometrics, but also his humility, efficiency and patience had a great impact on me.

Dr. Amy D. Yamashiro has been a wonderful supervisor in my part time job for about three years. Through her support, I got access to the data for this dissertation study. I appreciate her being an excellent mentor who leads me into the field of language testing.

I appreciate Dr. Kimberly S. Maier and Dr. Yeow Meng Thum in making time from their busy schedules to attend the proposal and defense meetings for my dissertation. Their valuable comments on the design and analyses of the dissertation study are highly appreciated.

Finally and most importantly, I want to thank my dear husband, Jin Wang, for his care and support during the years. I also thank my parents for building up my love for the truth and my analytical way of thinking—characteristics that are important in my academic life.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1. Introduction

I. Equating or Scaling

IRT (item response theory) models gain their flexibility by making strong statistical assumptions, which likely do not hold precisely in real testing situations. For this reason, studying the robustness of the models to violations of the assumptions, as well as studying the fit of the IRT model, is a crucial aspect of IRT applications.
(Kolen and Brennan, 2004, p156)

Equating the scores of two test forms has been in practice for at least half a century. In the early stage of the development of score equating, the similarity of forms in terms of structure or reliability was emphasized. Such requirements are reflected in the early works in educational measurements of Lord (1980), Angoff (1971), and Lord and Novick (1968). Kolen and Brennan (2004) summarized five desirable properties of equating relationships between the forms or between the equated scores. The properties are: symmetry of equating transformations; same specifications between the two test forms; equity property, which holds when examinees with a given true score have the same distribution of converted scores on Form X as they would on Form Y; observed score equating property in observed score equating; and group invariance property that means that the same equating relationship can be found using different groups of examinees. If all requirements are met, the forms are strictly parallel and need not to be equated. What has never been clear is the degree to which the requirements require fulfillment so that equating can be performed.

In recent years, as the demand for measuring achievement has increased, psychometricians have been challenged to put scores from test forms of differing content, structure or difficulty level on the same score scale. To differentiate traditional equating

methods from the more recently developed ones that link two forms obviously differing in structure or difficulty, three names have been applied to the statistical process. *Equating* refers to the traditional approach wherein the five properties are relatively met. The process is called *vertical scaling* when two forms differ in difficulty levels while the test structures are believed to be similar. And *vertical scaling* is most often used to assess growth such as in math achievement between different grades. *Linking* usually refers to the statistical process for putting scores from tests that are different in both difficulty and in content on the same scale. A typical example is to identify the equivalent ACT score for an SAT I (V+M) composite score (Kolen and Brennan, 2004). In this dissertation, the two forms studied are of the same test specification but are at different difficulty levels, thus the equating method should be categorized as *vertical scaling*. However, in this dissertation, the process is sometimes called *equating* to for convenience.

As often repeated in the history of science, practical issues usually challenge theories and technologies to improve. So is the case for test equating. When IRT was first developed, it required all items in a test measure the same trait or ability, or the same combination of traits/abilities. However in practice, it is well acknowledged that multiple skills are required to determine the correct answers for many test items. Multidimensional item response theory (MIRT) was developed to quantitatively analyze test structure. It has also been applied to equate tests that do not meet the unidimensionality assumption required by other procedures. However, MIRT models are more sophisticated and difficult to apply in practice. IRT models are robust and can tolerate multidimensionality to a certain degree, but it is necessary to consider in what situations, and to what degree, the simplicity of unidimensional IRT will reasonably model the data (Goldstein and Wood, 1989; Wang, 1985).

In recent years, especially after the NCLB (No Child Left Behind) implementation, accurate and accountable methods have been required to measure growth in achievement.

Measurement instruments that are different in specifications need to be put on the same scale to enable accurate growth evaluation. Recently developed IRT software, BILOG-MG and ICL, have integrated vertical equating features so that two forms of different difficulty levels can be put on the same scale. These developments make it possible to link forms that are different from each other in terms of test structure, examinee population, test difficulty etc. However, the validity of the scoring method and its accountability of measuring student achievement and growth still remain to be evaluated.

II. Purposes and Research Questions

The testing data studied in this dissertation characterizes the above mentioned challenges of IRT equating: the data have multidimensional structure, and difficulty differences exist between the forms—the equating is vertical by nature. Less studied equating designs are applied to explore their feasibility, and to find optimum solutions for today's testing practice.

Among the equating designs that are compared in this dissertation, common-item equating is most often seen in equating literature. In today's well-recognized textbook of equating by Kolen and Brennan (2004), common-group and common/common equating designs are not even mentioned. Common-item equating links the two test forms by items that appear in both forms. Equating methods seldom utilized—common-group equating and common-group/common-item combined equating—will undergo a detailed examination in this dissertation; the comparison between the common-item and common-group equating designs in vertical scaling will be discussed. Common-group equating here refers to an equating design that has three groups of examinees who take different test form(s). Group 1 takes test form 1, Group 2 takes test form 2 and Group 3 takes both forms (form 1 and form 2 share no common items). The data collected from all three groups will be used to concurrently calibrate the test item parameters and examinees' ability scores with the

unidimensional IRT three parameter logistic (3-PL) model using maximum likelihood estimation.

Another relatively obscure equating method—common-group/common-item equating, is also applied in this study. This equating design combines the common-group and common-item design. A detailed description of this method is found in chapter 3. All equating was done as multiple group concurrent estimation of item parameters with the unidimensional 3-PL IRT model, using maximum marginal likelihood (MML) estimation. Therefore, the research questions will be answered by comparing the results of the three equating designs are:

1. What is the difference in item parameter calibration or ability score calculation between the three equating designs?
2. Which design is more advantageous at different test lengths: 36, 48 and 60 items?
3. Which design is more advantageous when the average difficulty difference between the exams is 0.5 unit, 1 unit and 1.5 unit?

Standard errors of equating are calculated in evaluating the quality of equating designs; the item parameters and examinees' ability scores obtained through equating are compared with those of their real values (the "real values" will be defined in chapter 3). IRT model fit and practical issues of equating design are also discussed.

Chapter 2. Literature Review

I. IRT and IRT Vertical Scaling Designs

A. The Strength and Limitations of IRT in Test Equating

Before the IRT models were widely used in testing practice, several equating designs based on true score theory were developed and applied. Kolen and Brennan (2004) thoroughly describe these equivalent group equating, non-equivalent group equating, linear equating, equipercentile equating, and other methods. Compared to equating methods based on classical testing theories, IRT models are more advantageous in that they model examinee responses at the item level instead of the total score level. IRT models are now widely used in almost all aspects of psychometrics such as item banking, scoring, differential item functioning (DIF) analysis, adaptive testing etc. Increasingly more powerful computer software for IRT models have been developed for the expanding IRT applications. Due to the simplicity of the IRT models and the availability of software, progressively more equating or scaling are now performed with IRT.

Despite its strengths, IRT makes strong statistical assumptions, which are hard to meet precisely in real testing situations. The two major assumptions are local independence and unidimensionality; the two assumptions are related. Local independence means that the answer to one question is not related in any way to the answer(s) of other question(s). Lord (1980) stated it as Lazarsfeld's assumption of local independence, which is described as: "if we know the examinee's ability, any knowledge of his success or failure on other items will add nothing to the determination (of θ), if it did add something, then performance on the items in question would depend in part on some trait other than θ....." Lord (1980) further

described this assumption in a mathematical statement "that the probability of success on all items is equal to the product of the separate probabilities of success" (p19). This assumption cannot hold when testlets (like a reading test where items are grouped according to reading passages) are included in an exam. By unidimensionality is meant that all the test items test the same type of knowledge/ability or the same combinations of knowledge/abilities; put in the context of Lord (1980), a single θ is measured by the test.

When multidimensionality exists, more complicated IRT models are needed to accurately express the mathematical relationships between θ and item response pattern. Reckase (1997, p271) stated that "The number of skill dimensions needed to model the item scores from a sample of individuals for a set of test tasks is dependent upon both the number of skill dimensions and level on those dimensions exhibited by the examinees, and the number of cognitive dimensions to which the test tasks are sensitive." Unidimensionality almost certainly does not hold for data from most achievement test. Fortunately, IRT models are robust to a certain degree against assumption violations, which means that, although sometimes the model does not perfectly fit, the estimation based on it is still accurate enough to make educational decisions.

A combination of several elements determines the degree such violations are tolerable, and the degree of tolerance also depends on the research design. For example, in common-item nonequivalent equating using the IRT model, issues affecting the quality of equating may include the reliability of each test form, the quality of the test items, the selection of anchor items etc. The quality of test equating can be seriously undermined by a combination of inadequate test equating design and unsatisfied assumptions (Jodoin, 2003; Goldstein and Wood, 1989; Klein and Jarjoura, 1985; Beguin et al 2000; Skyes et al 2002).

A number of IRT models have been developed so that different models can be applied

according to the feature of the data and the needs of the analysis. In large-scale test equating designs, unidimensional IRT models are most often used because the practice is simpler and more economical than multidimensional IRT, even though sometimes the unidimensional assumption is not satisfied. In this study, in order to test the model's tolerance to multidimensionality, we will use unidimensional IRT equating although the evidence indicates that the data are multidimensional. The model applied here is three-parameter logistic model (3-PL), as presented in equation (2.1), which is widely used in multiple-choice large-scale testing. The definitions of the symbols in the equation are: P (Xij=1)—probability that person j with ability level θj can answer item i correctly; θ—examinee ability; b—item difficulty; a—item discrimination; c—lower asymptote or guessing parameter.

$$P(X_{ij}=1 \mid \theta_j, b_i, a_i, c_i) = c_i + (1-c_i)\frac{\exp[a_i(\theta_j - b_i)]}{1+\exp[a_i(\theta_j - b_i)]} \quad (2.1)$$

B. IRT in Vertical Scaling

In theory, item parameters calibrated with IRT models are independent of the examinees' ability level. The a_i, b_i and c_i in equation (2.1) are invariant parameters (Lord, 1980, p34). The difficulty parameter (b_i) is on the same scale as the examinees' ability levels. The distribution of examinees' ability is usually set as a standardized normal distribution. Another feature of item parameter calibration with the IRT model is called *indeterminacy*, which means "the choice of origin for the ability scale is purely arbitrary" (Lord 1980, p36). Thus the IRT parameters calibrated from the two forms require adjustment in order to be put in the same scale, because these parameters were calibrated based on different examinee groups.

The quality of IRT equating is not only decided by quality of the test items and whether the data collected from the examinees fit the IRT model, but also by the appropriateness of the design of test equating. Best equating results could be obtained when

the two forms satisfy the requirements of equating mentioned in Chapter One. However, this study focuses on vertical scaling—equating between two forms that are different in difficulty levels, and more often than not, also different in test domains. The process therefore requires tolerance to both lack of fit of the IRT model (multidimensionality) and that of the unsatisfied requirements in equating.

According to Kolen and Brennan (2004), vertical scaling refers to the "process used for associating performance on each test level to a single score scale, and the resulting scale is a *developmental score scale.*" Because tests of different levels—and quite inevitably, different constructs—are involved in vertical scaling, issues such as domains measured, definition of growth, multidimensionality, and others. need to be considered. Vertical scaling is much more sophisticated than equating and it involves more decisions in the design for equating. It is challenging that in testing practice, large-scale assessment often requires the scaling procedure to be simple and involve as little computation as possible.

Most of the studies that approach the issues in vertical scaling use a common-item equating design. Considering the challenges of vertical scaling, this study proposes two designs that are seldom mentioned in psychometrics literature. These designs may serve as better alternatives to the well-recognized common-item design.

a. Common-item vertical scaling

As is indicated by its name, in common-item equating, the two forms have some items in common. According to the invariant item (Lord, 1980) feature of the IRT model, the common items are supposed to function identically even when the examinees are different. Thus, based on the parameters of the common items, the two forms are linked. As for how many common items is enough for adequate linking accuracy, no theoretical conclusion can be drawn based on solid research. Conventionally, the linking items should take at least 20% of the total items in a form (Kolen and Brennan, 2004). In this study, three levels of test

length were applied to investigate the effect of test length. When the total number of unique items in the two forms was 120, 20 items were used as linking items; when the total number of items was 96, 16 items were used as linking; and when the total items was 72, 12 items were used as linking. Other than the requirements of the percentage that should be considered when selecting the linking items, the linking items are supposed to have high discrimination value, with stable function among different samples. What is more, the linking items should represent all the domains of the test forms. Due to practical considerations such as test security, linking items sometimes may not be able to satisfy these requirements.

Compare to concurrent IRT equating, two-step IRT equating is more often seen in the literature, especially in studies at the early stage of IRT equating. In two-step equating, the first step is to calibrate item parameters and examinee ability of the two forms separately. Then based on two sets of the parameters calibrated for the linking items from the two test forms, a linear or non-linear function is developed so that the two sets of parameters can be transformed to be equivalent to each other. The parameters of all the other items are then transferred in the same scale by the same mathematics function. Because the IRT ability estimates are on the same scale as the item difficulty parameters, the examinees' ability estimates of the two groups can also be transferred to the same scale based on this mathematical function.

Before BILOG-MG was available, BILOG is often used in IRT calibrating and equating. BILOG can be used for concurrent calibration of the item parameters when the two groups taking the two forms are randomly equivalent, but it is not strictly appropriate to use BILOG to concurrently estimate groups of different latent ability distributions. By using BILOG-MG, common-item non-equivalent group vertical scaling using IRT model becomes very convenient. BILOG-MG accomplishes multiple-group, common-item IRT equating concurrently for all the groups, with all item parameters calibrated concurrently. Research

indicates that the concurrent BILOG-MG equating using marginal maximum likelihood (MML) estimator is comparable or even superior to that of the two-step equating methods (Hanson and Beguin, 1999). However, DeMars (2002) shows that if group ability level is not taken into consideration, item parameter estimation is biased using MML estimation.

b. Common-group vertical scaling

The common-group equating design that is studied in this dissertation refers to the following: two test forms (with no item in common) are given to three groups of people, Group 1 takes Form 1 only and Group 2 takes Form 2 only, Group 3 takes the items on both forms. This method is different from the single-group design described in Kolen and Brennan (2004). The usual single-group design has one group of examinees answer the two test forms. However, in practice it is often expensive to have a sufficient number of examinees take the two tests. The common-group design does not require Group 1 and Group 2 to be equivalent, and Group 3 can be different from the other two groups. In this study, data is analyzed using BILOG-MG, and equating is performed concurrently using MML estimation.

Compared to common-item equating, studies on common-group equating are rare (Hambleton & Swanminathan, 1985, p.205; Hambleton et al., 1991, p.128; Noguchi, 1986, Noguchi, 1990; Toyoda, 1986, Ogasawara, 2001). This kind of equating links two test forms based on the same group of examinees that take both forms. Considering the number/percentage of examinees should be included in the common-group equating, no theory has been available for reference. In this study, the common-group equating is designed to be compared with the common-item equating; the strength of linking should be comparable between the designs. For example, in common-item equating, when the total number of items is 120, the number of common items is 20 and the number of examinees is 5,000 (2,500 for each group), a total number of 20*5,000=100,000 cells link the two forms together. To have the same number of cells linking the two forms, in common group equating,

the number of common examinees should be 100,000/120≈830. This principle was applied in the common-group equating of this study.

When linking items are unavailable, or when the statistical assumptions or requirements of common-item design cannot be fulfilled, common-group equating can be considered. Although issues such as fatigue exist in this design, common-group equating still serves as a possible alternative for the common-item equating design. In vertical scaling where the assumptions of IRT common-item equating are not fulfilled, common-group equating can possibly be a better choice. Harris (1991) compared spiraling design and single group design in vertical scaling, and found that across different examinee populations, the single group design exhibit more stability. The result of this study may or may not be applied to the common-group design here, for in the single-group design, two forms were equated by one group of people that were administered both forms. The common-group design described here has been seldom studied.

c. Common-item/common-group combined equating

This type of equating design combines the characteristics of the two equating methods introduced above: the two test forms share some common items and there is also a group of examinees that takes all the items from both test forms. However, the number of common items is only half as many as the common-item design, and the number of common examinees is also only half as many as the common-group design. This equating design has been used in large-scale testing practice but is not documented in publications (Y. M. Thum, personal communication, Nov. 18th, 2005).

This method is studied here because it may serve as an alternative practice when the number of common items and common examinees cannot satisfy the requirements of the common-item or common-group methods. Because this method combines the features of common-item and common-group equating, it contributes to the theory of equating design,

especially to the comparison between the common-item and common-group equating.

II. Issues of Multidimensionality in IRT Equating

As stated above, appropriate use of the IRT models requires the tenability of assumptions of unidimensionality and local independence. In test equating, each of the individual test forms should satisfy the assumptions. The testing data of each form should adequately fit the IRT model. However, in testing practice, these assumptions can be very stringent and impractical. In the past, a number of studies have been published on the effect of multidimensionality on IRT equating.

Jodoin (2003) used simulation data to investigate the impact of the violation of unidimensionality for individual test forms and inconsistency between the dimensional structure of the reference and focal forms. His conclusion was that low levels of dimensional inconsistency between the forms are reasonably well tolerated, but multidimensionality in either test form is not. Jodoin (2003) used IRT ability scores; he did not discuss the effect when anchor items do not represent all the domains of the form(s). Multiple studies have used IRT-true score equating functions to analyze the effect of test multidimensionality (Bogan & Yen, 1983; Bolt, 1999; Camilli, Wang & Fesq, 1992; Cook & Douglass, 1982; Cook, Dorans, Eignor, & Petersen, 1985; Dorans & Kingston, 1985; Kolen & Whitney, 1982; Snieckus & Camilli, 1993; Stocking & Eignor, 1986; Wang, 1985; Yen, 1984). Their results disagree with those of Jodoin (2003). The majority of these studies concluded that although multidimensionality of the latent ability space did affect the quality of IRT true-score equating, the impact often appeared to be minimal and of little practical significance, especially when correlations among the dimensions are high. Goldstein and Wood (1989) stated that the impact of multidimensionality on the quality of IRT equating is likely to be negligible as long as the same linear composite of latent traits, or reference composite (Wang, 1985), underlies the item response on both tests.

However, these studies did not clearly state whether the effect of the linking item design was considered. Although no explicit "requirements" are stated for linking item design, accepted practice calls for the set of common items to be proportionally representative of the total test forms in content and statistical characteristics (Kolen & Brennan, 2004). Previous research indicates that in linear equating, inadequate common item content representation can impact test scores when examinee groups taking alternate forms differ considerably in achievement level (Klein and Jarjoura, 1985). A later study by Beguin et al (2000) using simulated data noted a large effect of multidimensionality on IRT equating for nonequivalent groups. According to Sykes's et al (2002) research on a mixed-format math examination, equating by using anchors containing items that loaded more heavily on the first or the second dimension resulted into different standard errors.

The testing program used in this research is designed to measure three different types of English language ability—grammar, vocabulary and reading. Previous studies suggest that the data indicate a multidimensional pattern (Yamashiro and Yu, 2005a; Yamashiro and Yu 2005b). Further, the reading items are in the form of testlets, with a set of items focusing on the same reading passage. The grammar and vocabulary items are individual items. Due to security considerations, the anchor test of the ECPE G/V/R/ section contains only grammar and vocabulary items, no reading items. Based on the results from previous research, such anchor test designs are subject to systematic error (Sykes et al 2002). By comparing the equating results based on common-item equating design, common-group equating design and common-common design the study will estimate how much anchor test's lack of representative may affect the common-item equating, and whether common-group equating can circumvent the problem.

In this study, exploratory MIRT and exploratory factor analysis with oblique rotation on three factors/dimensions were applied to investigate the test's dimensionality. Tate (2003)

comprehensively summarized and compared the empirical methods of assessing the structure of tests with dichotomous items. About ten methods from exploratory and confirmatory families were included in this study, the methods were also categorized as parametric vs. non-parametric based on conditional item covariance. The results of this study indicated that for the most part, all methods performed reasonably well over a relatively wide range of conditions; exceptions only occurred when the test data departed appreciably from the assumptions or there is inherent limitation of a method. Compare with nonparametric methods, parametric modeling provides parsimonious and description of data structure. Factor analytic and MIRT methods were listed as parametric methods in Tate (2003). The MIRT method used to assess test dimension is the Normal-Ogive Harmonic Analysis Robust Method (NOHARM) developed by McDonald (2000) and programmed by Fraser and McDonald (1988).

III. Issues in Vertical Scaling

When different achievement tests are administered to different grades to assess growth in achievement, vertical scaling becomes inevitable. In vertical scaling, two test forms composed of items that have different difficulty levels are taken by two groups of examinees differing in ability. The results of vertical scaling can be unstable for multiple reasons such as: equating design, test dimensionality, test characteristics, DIF in different groups etc. These issues are discussed in the following sessions, and possible solutions are also introduced.

A. Issues of structure or dimension shift

When two test forms are composed of items from the same battery but are different in difficulty levels, there is a tendency for easier items to denote different constructs than higher difficulty items, even though they are designed to test the same constructs. A substantial amount of research suggests that when the same achievement battery measures achievement

at different levels, the content, complexity and difficulty of the assessment tasks also change (Linn, 1993; Mislevy, 1992; Yen, 1985, 1986). Even in a single form, differences in scores at the lower end of the scale may represent a different constructs from the differences in scores at the higher end of the construct (Reckase, 1989). Even when the two forms are carefully constructed to be parallel, different constructs can be empirically identified between the forms (Reckase, 1998). Dorans (1990) emphasized that forms to be equated should measure the same mix of content so that construct invariance could be achieved. However, in vertical scaling, this requirement is purposely violated. When the two forms cannot be considered to have the same construct, all the issues concerning test multidimensionality in equating would affect the vertical scaling results.

B. Issues of DIF

Among the issues that arise with vertical scaling, differential item functioning (DIF) should be considered seriously, especially when IRT models are applied. In vertical scaling, items that can be included in a battery should function identically between examinee populations that are of different ability levels (Kolen and Brennan, 2004). In common-item equating, only the linking items are administered to both examinee populations, and thus most of the items cannot be tested for DIF. As psychometricians are striving to improve the accountability of vertical scaling, different equating designs should be compared and evaluated. In common-group equating design, all the items are administered to part of the examinees from both groups. Thus, it is possible to estimate the DIF effect on the equating results.

Harris (1991) compared the results of Angoff's design I (spiraling design) and design II (single group design) in vertical scaling, and concluded that the single group design exhibits more stability across different samples. However, the increase in stability here is at the cost of more items administered to more examinees. In this study, vertical scaling results

will be compared using common-item or common-group as the linking design. This study can serve as reference to a seldom approached vertical scaling design that is of great potential.

C. Difficulty difference between the forms

Vertical equating is used to equate forms that differ in difficulty level. A major question is how much difference is reasonable between the two adjacent forms. No research has been found that directly explores this issue. Most studies on vertical equating use exams whose structures and designs are usually decided based on factors like test specifications, curriculums, policies etc. other than based on the requirements of vertical scaling. Item difficulty differences between the two forms, or the ability differences between the two groups, are seldom reported. Table 2.1 lists the item difficulty difference or group ability difference from several vertical scaling studies. The numbers provided here set a reference for how much growth (or difference) one may expect from the two groups in vertical scaling.

Pomplun et al (2004) and Kolen and Brennan (2004) reported the averaged item difficulty parameter after the forms were equated. In Pomplun's (2004) study, the differences in averaged item difficulty between the two adjacent grades are around 0.5-1.5 SDs; while the differences in item difficulty reported by Kolen and Brennan (2004) are more likely to be around 0.5 SD, even though both examined math achievement tests of a similar grade range. Russell (2000) focused on the ability growth between the two grades; and data on three subject areas (math, reading and language) were reported. Ability growth ranged from about 0.3 to about 1.5 SD, with bigger growth expected between lower grade levels. The ability difference levels reported in Jodoin (2003) are not comparable with those reported by other studies, because the forms were administered to students from the same grade, while different students were tested each year. The differences in averaged ability between years are very small (less than 0.1), indicating that, for the same grade, little change was observed in

students' ability from year to year. Based on the literature about item difficulty differences in vertical scaling, this dissertation assigns the forms to be different by 0.5, 1.0 and 1.5 SDs of averaged item difficulty.

D. The effect of test length

It is well-known from the literature that longer tests are usually of higher reliability, as long as the items all target the same trait or ability. To obtain sound accuracy in equating, each test form should have good reliability (above 0.85 in most high-stake exams) and a stable estimate of the IRT parameters. Existing literature offers some guidance on the test length needed to obtain reasonable estimates of IRT parameters. Lord (1980) clearly stated that test length and sample size, in combination, affects the quality of parameter estimates. Swaminathan and Gifford (1983) reported that multiple-choice tests below 15 items gave poor parameter estimates, and the inadequacy in item number could not be compensated by increasing sample size. Hambleton and Cook (1983) recommended a minimum of 200 examinees and 20 items to obtain stable testing results. Few studies have been found targeting the effects of test length on equating. Fitzpatrick and Yen (2001) suggested that a test should have at least eight 6-point items or at least twelve 4-point items. This study investigated constructed-response tests.

The situation becomes more complicated when multidimensional tests are equated. Sometimes tests that have only 20 items are seen in multidimensional IRT equating (Kim, 2001). It is assumed that both forms should have enough items to meet the requirement of reliability. Moreover, additional items may be needed for accurate equating results. While longer tests are favored when reliability is considered, shorter tests are preferred when cost is considered. This dissertation study explored the effects of test length using forms that contain 36, 48 and 60 items. It intends to address the question of whether shorter tests will perform equally well as longer tests in vertical scaling when reliability is adequate.

IV. Evaluating the errors of equating

A. Comparing the parameters' true values and those obtained through equating

A lot of equating studies use generated data to evaluate a certain equating method in which the true item parameters values are known. Typically in these studies, the standard errors of item parameters are calculated through the squared difference between the parameters obtained by equating and the true parameters on which the data were created -- for example the study by Hansen and Beguin (2002). Some studies that compare the quality of different equating methods by directly comparing the parameter obtained through these methods using scatter plots or correlation coefficients (Li, Griffith and Tam, 1997).

When real data is used and the true parameters are unknown, error estimation can be challenging. In this dissertation study, we use real data and the true parameters are unknown. We consider the parameters obtained using the original data (about 30,000 examinees' responds to 130 items) as the "real parameters". Sub-samples of about 5000-6000 were drawn from the original data: the items were split into two forms for each of the equating design. Item parameters obtained through each equating design were compared with the "real parameters" to reflect the quality of equating.

B. Standard error of equating

Standard error of equating usually refers to the errors due to sampling, instead of systematic errors. The two methods most commonly used in estimating standard errors of equating are the bootstrap and delta methods. The delta method is a set of "procedures [that] result in an equation that can be used to estimate the standard errors using sample statistics" (Kolen and Brennan, 2004, p234). This analytic method usually includes a process of time-consuming development of the equations, and it often results in very complicated equations. In this dissertation, standard errors of equating are estimated by a method similar

to the bootstrap method.

The bootstrap method calculates the standard deviation of equated scores over hypothetical replications of an equating procedure in samples from a population. In one hypothetical replication, a specified numbers of examinees would be randomly sampled. Then the Form Y equivalents of Form X scores would be estimated at various score levels using a particular equating method. The standard error of equating at each score level is the standard deviation, over replications, of the Form Y equivalent at each score level on Form X. Standard errors typically differ across score levels.

Bootstrapping is very computationally intensive, in which many samples are drawn from the data at hand and the equating functions are estimated on each sampling. In this dissertation study, a method similar to the bootstrapping but less computationally intensive was used. This method does not repeatedly sample subjects, makes the standard error estimate more reliable than bootstrapping. A large sample size of the testing data (about 30,000 for both forms) allows the examinees to be randomly divided into groups. For each test, the two forms were randomly paired to form ten paired samples. IRT equating of different designs was applied to each paired sample to obtain the ability scores. The standard deviations were calculated between the ten values to obtain standard error at each score level. This method of standard error calculation will be introduced with more detail in the following chapter.

Table 2.1. References on Difficulty Difference between Forms

Reference	Test Program	Criteria	Item difficulty difference	Group Ability difference
Pomplun M. et al (2004)	Grade 2 to 6 math exam	WINSTEPS estimated item difficulty	From grade 2-6, the averaged item difficulty levels are: -1.35, -1.02, -0.17, 0.36 and 1.61 respectively	
Russell M. (2000)	Grade 1 to 8 math	Expected growth size		Ability difference between two adjacent graded levels are: 1.38, 1.25, 0.89, 0.68, 0.53, 0.42, 0.38, 0.4 respectively
	Grade 1 to 8 reading	Expected growth size		N/A, 0.93, 0.79, 0.67, 0.52, 0.39, 0.39, 0.36
	Grade 1 to 8 language	Expected growth size		1.46, 1.10, 0.89, 0.58, 0.50, 0.32, 0.29, 0.29
Jodoin, M. G. et al (2003)	Mathematics of a certain grade tested in 1998, 1999 and 2000	State average ability measured by concurrent EAP ability estimate		From year 1998 to 2000, the averaged ability levels are: -0.12, -0.03, 0.04 respectively
Kolen M. J. and Brennan, R. L. (2004), p405-406	ITBS math test for grade 3 to 8	Equated item parameter by ICL EAP estimate	From grade 3-8, averaged item difficulty levels are: -0.29, 0.008, 0.627, 1.24, 1.81, 2.37 respectively.	

Chapter 3. Methods

I. Data Description

The data for this study is from a real examination program that tests the English proficiency of ESL (English as Second Language) learners. The exam is administered abroad annually in about 20 countries; at each administration, approximately 30,000 people take the exam. The whole exam program contains four sections: Writing, Speaking, Listening and the Grammar/Vocabulary/Reading (GVR) sections. The Writing and Speaking sections are performance assessment sections, while the Listening and GVR sections are composed of multiple-choice questions.

In this study, the data from the GVR section in one administration is used. The original data contains dichotomous responses of 130 GVR items administered to 29,935 examinees. Among the items, 50 test the learners' grammar ability, 50 test their vocabulary proficiency and 30 test their reading ability. The items testing grammar and vocabulary are independent items, while the reading items are given in the form of testlets. Not all the items were included in this study. The items were selected according to their quality (based on the values of classical test theory item parameters) and the test length of the equating designs. The numbers of items that were selected from each section for each level of test length are shown in Table 3.1.

Table 3.1. Number of Unique Items

Total number of unique ms in the two forms	120			96			72		
* Number of item in each form	60			48			36		
Section	G	V	R	G	V	R	G	V	R
Number of items in each section	23	24	13	19	19	10	15	15	6

*The item numbers listed are for common-group design, in common-item and common-common designs, more items were used in each form due to common items.

A. The Items' Unidimensional and Multidimensional IRT Parameters

All the items were first treated as one test administrated to one group of examinees. The examinees' scores and the item parameters for all the items were calibrated for the three-parameter logistic (3-PL) IRT model using BILOG-MG (Zimowski, 2003) with maximum marginal likelihood estimation. Item parameter estimates are presented in Chapter 4.

The examination program for this study is designed to test three aspects of English language proficiency—Grammar, Vocabulary and Reading. Previous statistical analysis on the data from this examination program and other ESL programs developed with similar test specification confirms that the test items are sensitive to differences in skills on three dimensions. Evidence of the multidimensionality of the item response data comes from exploratory factor analysis on the dichotomous response data, exploratory factor analysis on the scores of item clusters in each of the G, V, R subsections (Yamashiro & Yu, 2005), and structural equation modeling analysis on another exam that was developed with similar test specifications (Johnson, Yamashiro and Yu, 2004).

Multidimensional IRT parameters were estimated using TESTFACT (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, R., 2003) and NOHARM (Fraser, 1986), results from the two methods are very similar. The parameters estimated through the NOHARM program are

presented in Chapter 4. The lower-asymptote parameters calibrated in the previous step were used in the NOHARM control file, a three-dimensional, exploratory solution was specified. One item from each of the G, V, R section was selected as anchor item for dimension 1, 2, and 3 respectively. The d-parameter and the discrimination parameters on each of the three dimensions are shown in Chapter 4.

B. Factor analysis

To further check the dimension of the response data, three-factor exploratory analysis with oblique rotation were done on the dichotomous testing data using SPSS. The factor loadings and the correlations between factors are presented in Chapter 4.

C. Goodness of fit

Approximate chi-square indices of fit were computed by BILOG-MG for each item in the item calibration phase, using the responses from all examinees to the 120 item, 96 item and 72 item data sets. For the purpose of computing these chi-squares, the scale score continuum was divided into 20 intervals (the maximum number of intervals allowed by BILOG-MG). For each item, within each of the intervals, the actual and the expected percentage of item endorsements were computed. The chi-square indices reveal the discrepancy between the two percentages. The bigger the chi-square, the less likely the model fits the actual responses to the item. The expected percentage of item endorsement at each ability level was calculated based on BILOG-MG3 parameter estimation, with EAP (expected a posteriori) estimation of theta. The degree of freedom for chi-square indices equaled the number of intervals, in this case, was 20. Because the sample size is very large and chi-square is sensitive to sample size, model fit cannot be solely decided by the p-value of chi-square. The rule of thumb that chi-square equal or less than three times degree of freedom was used to evaluate item model fit, together with the p-value ($p \geqq 0.01$) of chi-square.

II.Equating Designs

This research study investigates the difference in equating results among three different equating designs, given that the forms are of different lengths and have different difficulty differences. The original data is from one test that has 130 items. According to the test length assigned to test equating design, 120, 96 and 72 items were selected from these items. The items were then split into two groups—Form 1 and Form 2. Which item was assigned to which form was mainly decided based on the category of the item (grammar, vocabulary or reading) and the item's IRT difficulty parameter. For each level of test length, the items were selected so that the averaged difficulty differences between the two forms were designed to be 0.5, 1.0 or 1.5 standard deviation(s) of the IRT ability score. As introduced in the previous chapter, among the studies reported different difficulty differences between the forms in vertical scaling, the difference between adjacent levels can be as small as 0.3 SD to as big as 1.4 SD's, depending on the examinee of the exam and the grade levels of the forms (Russell, 2000; Pomplun, Omar and Custer, 2004; Jodoin, Keller and Swaminathan, 2003). In most of the studies listed here, the averaged item difficulty difference between adjacent forms in vertical scaling is expected to be between 0.5 and 1.0. No study was found that reported the vertical scaling of ESL exams. In this dissertation, difficulty difference between the two forms were assigned to be 0.5, 1.0 and 1.5, for the reason that a difference of 0.5 or 1.0 can reflect situations in real tests; while a difference of 1.5 exaggerated the difference a little bit, so that if the effect of difficulty difference was subtle, it could be detected.

Three equating designs were studied: common-group equating, common-item equating and the combination of common-group and common-item equating. Different test lengths were tried and it was found that reliability drops to lower than 0.85 when test length for each form is less than 36. It was considered that the same number of anchor items should be selected from the grammar and vocabulary sections, because in the examination program, the

same numbers of items were in these two sections. The number of common items in the common-item design is twice as much as that of the common items in the common-common design. Considering these requirements, three levels of test lengths were selected for comparison in this study: 60, 48 and 36.

The dissertation investigates the main effect of three factors: equating design, test length and difficulty difference between forms; and there are three levels in each factor. Thus a total number of 3*3*3=27 designs of equating were structured and analyzed in this study.

A. Common-Group Equating

In this design, two sub-samples shared about 20% of the total examinees (Group 2), the responses of these examinees on both Form 1 and Form 2 were applied to final analysis. The responses of 40% examinees that had relatively lower ability level (Group 1) on the lower level form were included, while the responses of another 40% examinees that had relatively higher ability level (Group 3) on the higher level form were selected. The ability distribution of the combination of Group 1 and 2 is normal; the ability distribution of the Group 2 and 3 combined is also normal. No common items are taken by Group 1 and Group 3.

The examinees in each group were selected based on their ability scores using a MATLAB program. Examinees in Group 2 (N=1000) were first selected so that the mean of the group is zero with standard deviation equals 0.5. To select the examinees of the designed distribution in ability scores, the first step was to create 1000 numbers of normal distribution with mean=0 and SD=0.5. The created numbers were divided into 40 bins that had equal intervals. These bins were then used as template to select ability scores from the 19,935 cases in the original data, examinees were selected until all the bins were full. The MATLAB commands for selecting Group 2 for common-group and common-common equating are shown in Appendix 1.

Then, according to the difference in averaged item difficulty, Group 1 and Group 3 were formed. For example, when the two forms' averaged item difficulty were different by 1.0, Group 1 and Group 2 combined together should have distribution of mean=-0.5 and SD=1, total number of 3000 cases; while Group 3 and Group 2 combined together should have distribution of mean=0.5 and SD=1, also with total number of 3000 cases. Thus Group 1, with 2000 cases and a distribution in which a group of N{0, 0.5} (normal distribution with mean=0 and SD=0.5, 1000 cases) is deleted from a group of N{-0.5, 1} (normal distribution with mean=-0.5 and SD=1, 3000 cases). The 2000 cases that satisfied this distribution were also selected using the MATLAB program. The MATLAB commands that were used to select cases in Group 1 or Group 3 for common-group and common-common designs are shown in Appendix 2.

B. Common-Item Equating

In this design, about 20% of items from each form were selected as common items (as illustrated in Figures 3.1-3.3). However, the linking items do not represent all the subsections of the exams, because only grammar items and vocabulary items were selected; in real test, reading items can not be used as anchor item for security reasons. Anchor items were selected from different levels of difficulty (i.e. approximately equal number of items were drawn from top 30% difficult ones, middle 30% difficult ones and lower 30% difficult ones). The averaged difficulty differences were kept the same between the two forms after adding in the common items.

Two sub-samples were drawn from the original data, with one sub-sample contained examinees of lower ability scores, and the other contained higher ability examinees. The two sub-samples were both normally distributed in IRT ability scores with standard deviation equals 1. The means of the two sub-samples were decided based on the item difficulty difference. When the averaged item difficulty was different by 0.5, the mean of ability scores

of the two groups were -0.25 and 0.25; when difficulty difference was 1.0, the mean of ability scores were -0.5 and 0.5; when difference was 1.5, the mean of ability scores were -0.75 and 0.75. Each sub-sample contained 3000 cases that were drawn from the original data. The MATLAB commands used for case selecting here is very similar with the ones used in selecting Group 2 for the common-group design (Appendix 1). Only the values of the means needed to be changed to obtain the target distribution. For each pair of sub-samples, no common examinees were shared.

C. Common-group and common-item combined equating

As its name indicated, this design combined the features of the two above mentioned equating designs. About 10% of the items were shared by the two forms as common items, and about 10% of the total examinees were shared by the two groups as common-group. The item/common item numbers at each test length is given in Figures 3.1-3.3. The common items were selected from diffierent levels of item difficulty as described in common-item design. Unlike the Group 2 described in the common-group design that have 1000 examinees, the common groups in the common-common design have 500 examinees. Group 1 or 3 each has 2000 examinees. The 500 examinees were drawn to have normal distribution with mean of 0 and SD of 0.5. The 2000 examinees of Group 1, when combined with the 500 examinees of Group 2 have normal distribution with SD=1 and mean=-0.25, -0.5 or -0.75 according to the difficulty difference between the forms in the equating design. The 2000 examinees of Group 3, when combined with the 500 examinees of Group 2, have normal distribution with SD=1 and mean=0.25, 0.5 or 0.75 according to the difficulty difference betweeen the forms in the equating design. The examinees of this design were also drawn with MATLAB software using similar commands shown in Appendix 1 and 2.

Figure3.1. Three designs when total item=120

	Common-group		Common-item			Common-common		
	Easy	Hard	Easy		Hard	Easy		Hard
Number Item	60	60	50	20	50	55	10	55
Lower ility group								
Higher ility group								

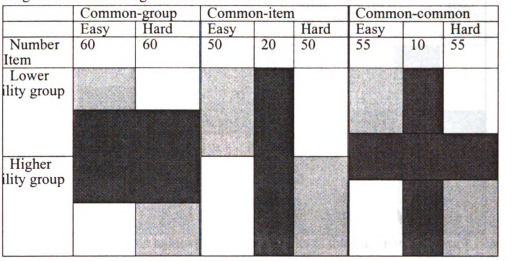


Figure3.2. Three designs when total item=96

	Common-group		Common item			Common-common		
	Easy	Hard	Easy		Hard	Easy		Hard
Number items	48	48	40	16	40	44	8	44
Lower ility oup								
Higher ility oup								

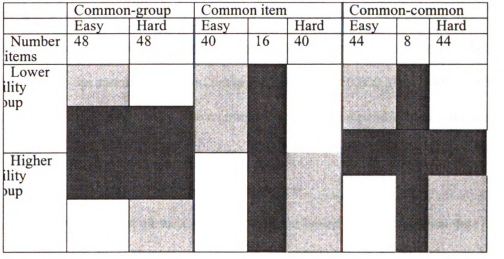
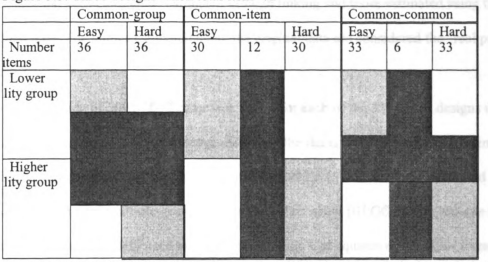


Figure 3.3. Three designs when total item=72



At each test length, the numbers of common cells in the data matrix for each design were kept the same. For example, when the total item number was 120, common-group equating overlapped part had 1000*120=120,000 of common cells in the data matrix; the common-item design had the overlapped part of 6000*20=120,000 common cells in data matrix; while the common-common design had 500*120=60,000 plus 5500*10=55,000, a total of 115,000 common cells. The overlapped parts are comparable in size for the three designs.

D. Data Analysis and Evaluation of Different Designs

The ability scores of all the examinees were calibrated from the original data (29935*130) matrix, using BILOG-MG3 with maximum marginal likelihood (MML). The proficiency estimate for each examinee obtained this way is considered as the estimate closest to the examinee's real ability, and is thus used as standard. It is called the "real score". For each design at each test length with specific item difficulty difference, certain items were selected; the responses for these items from all the 29,935 examinees were included in the data. Thus altogether 27 sets of data were prepared.

Two criteria were used to evaluate the quality of different equating designs. One criterion compares the examinees' scores between the "real score" and the scores obtained through the

equating design data matrix that are described in Figures 3.1-3.3. The other criterion compares the item parameters estimated through linking and those estimated using the original data. The values estimated using the original data are considered the "real parameter values".

A sub-sample of 5000-6000 cases was drawn for each of the 3*3*3=27 designs using the MATLAB program as described above. Although the items represented different dimensions of English language ability, the original data (the 29935*130 matrix) was calibrated as if it was unidimensional multiple-group dichotomous data using BILOG-MG3; because this method is the most widely used in testing application. Chi-squares of the items were used as model-fit index, the chi-squares are presented in chapter 4, the values indicate that the data fit 3-PL unidimensional IRT model reasonably well. The data matrices described in Figures 3.1-3.3 were also calibrated as if they were unidimensional using BILOG-MG3. In each equating design, item parameters and examinee ability scores of all the groups were calibrated simultaneously using MML estimation. The calibrated item parameters and the examinees' ability scores were plotted against the corresponding parameters that were calibrated with the original data. The correlations and the average squared differences between the "real scores" and the equated scores of each design were also calculated.

III. Standard Error of Equating

As introduced before, 27 sets of data were created, each had different item sets according to the design of test length and difference in averaged test difficulty; but all data sets contain responses from 29,935 examinees. To calculate the standard error of equating in each design, each data set was divided randomly into 10 groups of examinees. The examinees in each group were then randomly divided into two sub-groups (each has about 1500 cases) for equating. The two sub-groups shared 500 common examinees in common-group design and

shared about 250 examinees in common-common design; but in common-item design, they shared no common examinee. Although the two forms are different in averaged difficulty level, the two sub-groups were not designed to be different in ability level.

Each of the ten data sets was calibrated using the 3-PL IRT model with BILOG-MG, MML estimation were used to estimate the ability scores and the item parameters. The sampling method used in standard error calculating was different from the one that used before, here the 10 groups were randomly divided using the original data; however, the in the equating design that was introduced in the previous section, the samples were drawn based on their ability score distribution and the averaged ability scores are different between the two sub-groups. Separate random samples were used to calculate the standard error of equating because the number of examinees in the original data is limited. It is impossible to draw ten groups (with at least 1500 examinees in each group so that 3-PL IRT estimation is sufficiently accurate) and each contains two sub-groups that satisfy the distribution requirements as described before.

The "real scores" for all examinees were divided into 80 levels between -4 and 4, with an interval of 0.1. For the examinees whose "real scores" were within each interval, their equated scores calculated in the sub-samples were determined. The standard deviation of the equated scores from the same interval provided an estimate of the standard error of equating at this score level.

Chapter 4. Results

I. Parameters, dimensions and model fit analysis

The discrimination, difficulty, and lower asymptote (a, b and c) parameters are shown in table 4.1. The item IDs indicate the subsection and sequence number of each item—grammar (G), vocabulary (V) and reading (R) subsections respectively. Thus, "G1" is the first grammar item. Table 4.2 gives the multidimensional IRT item parameters estimated by NOHARM. The highest discrimination parameter is bolded. When two discrimination parameters for an item differ by 0.02 or less, both parameters are bolded. The design of the G/V/R structure is reflected through the MIRT a-parameters. Most of the Grammar items' highest a-parameter estimates are on the first dimension—a1, while the Vocabulary and Reading items mostly have their highest parameter estimates on the third and the second dimension, respectively. The result of a factor analysis with oblique rotation shows a similar pattern (Table 4.3). The percentage of items that load high on each factor or dimension is listed in Table 4.4.

Table 4.1. IRT Parameter Estimates for All Items

Item	a	b	c	Item	a	b	c	Item	a	b	c	Item	a	b	c
G1	0.939	-1.570	0.244	**G36**	1.447	-0.658	0.375	**V21**	1.519	-1.074	0.184	**R1**	0.910	-2.015	0.114
G2	1.799	-1.105	0.174	**G37**	0.953	-1.144	0.090	**V22**	1.758	-1.188	0.113	**R2**	0.526	1.204	0.500
G3	1.142	-2.325	0.106	**G38**	1.611	-0.312	0.179	**V23**	1.179	0.135	0.138	**R3**	1.011	1.159	0.083
G4	0.783	-2.758	0.064	**G39**	1.595	-0.721	0.106	**V24**	2.155	-1.359	0.203	**R4**	1.565	-0.329	0.186
G5	1.445	-1.631	0.461	**G40**	0.983	-1.195	0.037	**V25**	2.319	2.289	0.275	**R5**	0.959	0.156	0.188
G6	1.348	-1.707	0.341	**G41**	1.090	0.557	0.158	**V26**	1.383	-1.208	0.136	**R6**	1.556	1.269	0.383
G7	1.207	-1.935	0.079	**G42**	1.651	0.382	0.316	**V27**	1.772	0.366	0.375	**R7**	1.381	-0.046	0.500
G8	1.592	-0.409	0.189	**G43**	2.534	1.339	0.408	**V28**	1.709	-0.323	0.292	**R8**	0.769	-1.553	0.228
G9	1.055	-0.430	0.321	**G44**	1.295	0.696	0.206	**V29**	1.103	0.650	0.345	**R9**	0.754	-1.379	0.241
G10	1.518	-0.848	0.077	**G45**	0.472	1.067	0.167	**V30**	1.357	-1.003	0.182	**R10**	1.475	-0.665	0.205
G11	1.572	-0.598	0.202	**G46**	2.076	0.520	0.267	**V31**	1.342	0.157	0.254	**R11**	0.993	-2.451	0.112
G12	1.549	-0.929	0.204	**G47**	1.155	-1.392	0.094	**V32**	1.819	-0.073	0.367	**R12**	0.27	-5.55	0.237
G13	2.121	-1.048	0.07	**G48**	0.788	-1.439	0.371	**V33**	0.703	-2.314	0.034	**R13**	1.179	-2.439	0.155
G14	0.749	-1.748	0.053	**G49**	1.374	0.192	0.185	**V34**	1.713	-0.256	0.197	**R14**	2.174	0.16	0.238
G15	0.856	-2.104	0.099	**G50**	1.145	0.533	0.053	**V35**	1.261	-0.149	0.294	**R15**	0.600	-1.101	0.238
G16	1.613	0.137	0.297	**V1**	1.055	-0.562	0.192	**V36**	1.183	0.787	0.130	**R16**	1.454	-0.299	0.413
G17	1.752	-0.947	0.28	**V2**	1.596	-1.564	0.226	**V37**	0.986	0.223	0.18	**R17**	1.146	0.305	0.267
G18	1.553	0.214	0.441	**V3**	0.902	-0.782	0.434	**V38**	1.694	-1.164	0.182	**R18**	0.887	-2.201	0.131

(Table 4.1 continued)

G19	1.854	0.457	0.184	**V4**	1.018	-1.744	0.049	**V39**	0.985	0.054	0.322	**R19**	1.057	2.565	0.161
G20	0.921	-2.017	0.069	**V5**	1.184	-2.677	0.082	**V40**	1.535	0.320	0.401	**R20**	1.693	0.543	0.299
G21	1.179	-1.451	0.108	**V6**	1.737	-1.056	0.179	**V41**	1.821	1.081	0.123	**R21**	1.079	0.041	0.247
G22	1.171	-0.770	0.069	**V7**	1.242	0.263	0.313	**V42**	1.790	-0.488	0.235	**R22**	1.243	0.858	0.461
G23	0.616	-0.421	0.04	**V8**	1.630	-1.353	0.244	**V43**	1.488	0.080	0.147	**R23**	1.227	1.105	0.203
G24	0.993	-0.646	0.445	**V9**	1.184	-0.667	0.148	**V44**	1.855	-0.286	0.259	**R24**	1.176	0.505	0.290
G25	0.880	0.443	0.326	**V10**	1.523	-0.163	0.277	**V45**	1.098	0.175	0.159	**R25**	1.067	-0.961	0.383
G26	1.319	0.944	0.152	**V11**	1.985	0.992	0.350	**V46**	0.650	0.392	0.065	**R26**	1.013	-1.499	0.234
G27	1.436	-1.121	0.105	**V12**	1.912	-1.612	0.109	**V47**	1.422	0.815	0.239	**R27**	0.602	-0.831	0.085
G28	0.978	-1.022	0.137	**V13**	1.456	-0.88	0.161	**V48**	1.457	0.155	0.232	**R28**	1.086	-0.592	0.230
G29	1.099	-0.807	0.253	**V14**	1.377	-0.592	0.209	**V49**	1.623	-1.019	0.252	**R29**	0.952	-1.607	0.108
G30	0.501	-1.595	0.032	**V15**	1.159	-0.578	0.153	**V50**	2.011	1.531	0.169	**R30**	0.987	0.316	0.303
G31	1.782	0.978	0.128	**V16**	1.462	-0.083	0.099								
G32	1.862	-0.974	0.149	**V17**	1.210	-0.431	0.107								
G33	1.064	-1.271	0.108	**V18**	1.203	-0.270	0.283								
G34	1.323	-0.370	0.335	**V19**	1.438	-1.048	0.189								
G35	1.906	-0.921	0.115	**V20**	1.335	-0.471	0.163								

Table 4.2. MIRT Parameter Estimates for of All Items

item	d	a1	a2	a3	item	d	a1	a2	a3	item	d	a1	a2	a3	item	d	a1	a2	a3
G1	1.09	**0.46**	0.15	0.13	**G36**	0.97	**0.51**	0.16	0.29	**V21**	1.12	**0.51**	0.28	**0.51**	**R1**	1.16	**0.33**	0.30	0.23
G2	1.30	**0.80**	0.29	0.34	**G37**	0.74	**0.46**	0.15	0.22	**V22**	1.33	**0.64**	0.29	0.66	**R2**	0.47	0.06	**0.12**	0.09
G3	1.58	**0.58**	0.20	0.14	**G38**	0.54	**0.63**	0.16	0.37	**V23**	0.13	**0.38**	0.15	**0.37**	**R3**	-0.47	**0.32**	0.15	0.25
G4	1.31	**0.44**	0.14	0.03	**G39**	0.80	**0.73**	0.2	0.36	**V24**	1.79	**0.77**	0.35	0.67	**R4**	0.55	**0.48**	0.34	0.41
G5	1.71	**0.61**	0.16	0.22	**G40**	0.78	**0.67**	0.12	0.08	**V25**	-0.53	-0.06	0.08	**0.11**	**R5**	0.20	**0.30**	0.27	0.21
G6	1.56	**0.56**	0.17	0.26	**G41**	-0.05	**0.39**	0.06	0.28	**V26**	1.09	**0.54**	0.26	0.40	**R6**	0.02	0.10	**0.20**	0.22
G7	1.39	**0.55**	0.18	0.28	**G42**	0.26	**0.37**	0.11	0.35	**V27**	0.38	0.27	0.14	**0.46**	**R7**	0.76	**0.25**	**0.25**	0.23
G8	0.62	**0.61**	0.20	0.35	**G43**	-0.03	0.12	0.08	**0.20**	**V28**	0.71	**0.49**	0.18	**0.47**	**R8**	0.93	0.22	**0.31**	0.22
G9	0.67	**0.42**	0.11	0.19	**G44**	-0.07	**0.35**	0.16	0.25	**V29**	0.22	**0.26**	0.12	0.20	**R9**	0.87	0.20	**0.32**	0.21
G10	0.86	**0.79**	0.22	0.28	**G45**	-0.04	**0.20**	0.06	0.09	**V30**	0.97	**0.48**	0.19	0.44	**R10**	0.82	**0.44**	**0.43**	0.41
G11	0.78	**0.65**	0.18	0.32	**G46**	0.08	**0.40**	0.17	0.37	**V31**	0.31	0.27	0.17	**0.48**	**R11**	1.51	0.29	**0.47**	0.26
G12	1.07	**0.78**	0.24	0.18	**G47**	1.03	**0.58**	0.17	0.23	**V32**	0.65	0.37	0.17	**0.50**	**R12**	1.10	0.02	**0.21**	0.07
G13	1.44	**1.19**	0.34	0.30	**G48**	1.03	**0.29**	0.11	0.19	**V33**	0.97	**0.29**	0.13	0.16	**R13**	1.80	0.37	**0.55**	0.28
G14	0.84	**0.45**	0.16	0.05	**G49**	0.17	**0.47**	0.11	0.34	**V34**	0.57	0.43	0.25	**0.65**	**R14**	0.30	0.43	0.36	**0.49**
G15	1.13	**0.42**	0.11	0.18	**G50**	-0.26	**0.60**	0.11	0.21	**V35**	0.52	**0.33**	0.23	**0.34**	**R15**	0.66	0.14	**0.25**	0.19
G16	0.38	**0.50**	0.16	0.25	**V1**	0.58	**0.38**	0.15	0.31	**V36**	-0.24	0.23	0.12	**0.47**	**R16**	0.82	0.26	**0.39**	**0.39**
G17	1.21	**0.67**	0.17	0.40	**V2**	1.59	**0.58**	0.27	0.50	**V37**	0.16	0.18	0.14	**0.47**	**R17**	0.25	0.19	0.29	**0.36**
G18	0.56	**0.35**	0.09	0.30	**V3**	0.91	0.24	0.12	**0.30**	**V38**	1.32	0.55	0.32	**0.63**	**R18**	1.28	0.25	**0.47**	0.22
G19	-0.03	**0.50**	0.14	0.38	**V4**	1.07	**0.40**	0.20	0.31	**V39**	0.44	0.20	0.12	**0.36**	**R19**	-0.73	0.11	0.11	**0.13**
G20	1.14	**0.47**	0.16	0.16	**V5**	1.76	**0.42**	0.26	0.28	**V40**	0.44	0.26	0.15	**0.36**	**R20**	0.15	0.24	**0.44**	0.32

(Table 4.2 continued)

G21	1.12	**0.65**	0.17	0.17	**V6**	1.23	**0.59**	0.32	0.57	**V41**	-0.56	0.29	0.14	**0.48**	**R21**	0.346	0.24	**0.39**	0.27
G22	0.62	**0.61**	0.18	0.23	**V7**	0.33	**0.30**	0.18	0.28	**V42**	0.80	**0.54**	0.19	**0.54**	**R22**	0.350	0.11	**0.38**	0.15
G23	0.21	**0.29**	0.08	0.18	**V8**	1.44	**0.59**	0.24	0.48	**V43**	0.20	0.42	0.18	**0.51**	**R23**	-0.251	0.19	**0.32**	0.22
G24	0.90	**0.33**	0.1	0.21	**V9**	0.63	**0.48**	0.18	0.29	**V44**	0.68	0.47	0.27	**0.55**	**R24**	0.188	0.20	**0.38**	0.24
G25	0.29	**0.29**	0.11	0.13	**V10**	0.57	0.32	0.18	**0.60**	**V45**	0.15	0.28	0.19	**0.40**	**R25**	1.219	0.27	**0.86**	0.13
G26	-0.31	**0.38**	0.05	0.31	**V11**	0.00	0.16	0.18	**0.30**	**V46**	-0.06	0.21	0.10	**0.28**	**R26**	1.352	0.35	**0.89**	0.09
G27	1.04	**0.68**	0.19	0.32	**V12**	1.78	**0.70**	0.36	0.60	**V47**	-0.08	0.20	0.18	**0.40**	**R27**	0.416	0.14	**0.44**	0.15
G28	0.74	**0.48**	0.13	0.21	**V13**	0.92	**0.53**	0.23	0.47	**V48**	0.28	0.34	0.12	**0.50**	**R28**	0.707	0.3	**0.63**	0.21
G29	0.80	**0.51**	0.12	0.20	**V14**	0.72	**0.47**	0.16	0.43	**V49**	1.17	**0.55**	0.28	0.42	**R29**	1.191	0.36	**0.83**	0.09
G30	0.53	**0.37**	0.09	-0.04	**V15**	0.60	0.33	0.19	**0.51**	**V50**	-0.67	0.09	0.09	**0.42**	**R30**	0.308	0.21	**0.44**	0.14
G31	-0.48	**0.43**	0.10	0.35	**V16**	0.24	**0.49**	0.16	**0.50**										
G32	1.25	**0.99**	0.27	0.23	**V17**	0.45	**0.43**	0.17	**0.43**										
G33	0.91	**0.59**	0.20	0.12	**V18**	0.59	0.27	0.16	**0.49**										
G34	0.71	**0.49**	0.12	0.23	**V19**	1.05	**0.50**	0.22	0.46										
G35	1.12	**0.86**	0.26	0.39	**V20**	0.59	0.38	0.22	**0.54**										

Table 4.3. Dichotomous Factor Analysis of All Items with Oblique Rotation

Item	1	2	3	Item	1	2	3	Item	1	2	3	Item	1	2	3
G1	**0.384**	0.049	-0.020	**G36**	**0.313**	-0.010	0.162	**V21**	0.144	0.043	**0.333**	**R1**	0.142	**0.199**	0.105
G2	**0.402**	0.054	0.073	**G37**	**0.323**	0.020	0.094	**V22**	0.153	-0.000	**0.377**	**R2**	-0.020	**0.114**	0.074
G3	**0.445**	0.070	-0.060	**G38**	**0.342**	-0.040	0.198	**V23**	0.152	-0.020	**0.313**	**R3**	0.173	0.037	**0.188**
G4	**0.429**	0.078	-0.140	**G39**	**0.389**	-0.020	0.14	**V24**	0.202	0.024	**0.309**	**R4**	0.162	0.138	**0.226**
G5	**0.422**	-0.000	0.050	**G40**	**0.549**	-0.010	-0.11	**V25**	-0.190	0.080	**0.171**	**R5**	0.136	**0.192**	0.091
G6	**0.364**	0.009	0.104	**G41**	**0.248**	-0.100	0.237	**V26**	**0.230**	0.057	**0.229**	**R6**	-0.080	0.147	**0.220**
G7	**0.346**	0.010	0.125	**G42**	0.159	-0.050	**0.310**	**V27**	-0.010	-0.050	**0.458**	**R7**	0.085	**0.166**	0.144
G8	**0.326**	0.006	0.173	**G43**	-0.020	-0.000	**0.263**	**V28**	0.176	-0.030	**0.343**	**R8**	0.027	**0.249**	0.123
G9	**0.320**	-0.010	0.095	**G44**	**0.201**	0.043	0.170	**V29**	0.146	0.021	**0.169**	**R9**	0.014	**0.268**	0.110
G10	**0.453**	0.015	0.041	**G45**	**0.174**	-0.000	0.050	**V30**	0.184	-0.010	**0.320**	**R10**	0.101	**0.222**	0.206
G11	**0.377**	-0.010	0.137	**G46**	0.166	0.002	**0.291**	**V31**	-0.030	-0.020	**0.465**	**R11**	0.034	**0.349**	0.091
G12	**0.497**	0.055	-0.060	**G47**	**0.395**	0.016	0.065	**V32**	0.052	-0.030	**0.437**	**R12**	-0.080	**0.245**	0.011
G13	**0.488**	0.037	-0.030	**G48**	**0.198**	0.009	0.140	**V33**	**0.203**	0.054	0.087	**R13**	0.078	**0.377**	0.056
G14	**0.426**	0.085	-0.130	**G49**	**0.255**	-0.060	0.256	**V34**	0.021	-0.010	**0.480**	**R14**	0.072	0.135	**0.313**
G15	**0.330**	-0.010	0.081	**G50**	**0.438**	-0.05	0.058	**V35**	0.093	0.093	**0.262**	**R15**	-0.03	**0.210**	0.143
G16	**0.331**	0.009	0.119	**V1**	0.188	0.000	**0.246**	**V36**	-0.050	-0.060	**0.490**	**R16**	-0.040	0.239	**0.270**
G17	**0.343**	-0.040	0.208	**V2**	0.203	0.026	**0.296**	**V37**	-0.110	-0.040	**0.512**	**R17**	-0.080	0.178	**0.311**
G18	**0.180**	-0.060	0.272	**V3**	0.076	0.007	**0.269**	**V38**	0.096	0.041	**0.389**	**R18**	0.022	**0.380**	0.051

(Table 4.3 continued)

G19	**0.249**	-0.050	0.270	**V4**	0.197	0.052	**0.210**	**V39**	-0.030	-0.010	**0.390**	**R19**	0.011	0.070	**0.127**
G20	**0.367**	0.045	0.014	**V5**	**0.211**	0.122	0.142	**V40**	0.032	0.008	**0.356**	**R20**	-0.030	**0.318**	0.176
G21	**0.470**	0.021	-0.02	**V6**	0.159	0.045	**0.330**	**V41**	-0.000	-0.050	**0.471**	**R21**	-0.000	**0.294**	0.150
G22	**0.407**	0.020	0.054	**V7**	0.113	0.061	**0.232**	**V42**	0.172	-0.040	**0.374**	**R22**	-0.070	**0.376**	0.031
G23	**0.210**	-0.020	0.141	**V8**	0.228	0.004	**0.288**	**V43**	0.096	-0.040	**0.414**	**R23**	-0.000	**0.265**	0.131
G24	**0.229**	-0.010	0.147	**V9**	**0.278**	0.026	0.167	**V44**	0.090	0.029	**0.389**	**R24**	-0.030	**0.315**	0.124
G25	**0.235**	0.034	0.054	**V10**	-0.040	-0.04	**0.528**	**V45**	0.023	0.037	**0.365**	**R25**	-0.000	**0.594**	-0.150
G26	0.214	-0.110	**0.287**	**V11**	-0.070	0.077	**0.334**	**V46**	0.045	-0.020	**0.298**	**R26**	0.064	**0.588**	-0.200
G27	**0.393**	-0.000	0.116	**V12**	0.200	0.054	**0.288**	**V47**	-0.060	0.030	**0.414**	**R27**	-0.050	**0.421**	-0.000
G28	**0.351**	-0.010	0.086	**V13**	0.190	0.009	**0.311**	**V48**	0.044	-0.080	**0.465**	**R28**	0.033	**0.477**	-0.030
G29	**0.388**	-0.020	0.064	**V14**	0.187	-0.030	**0.324**	**V49**	0.227	0.060	**0.237**	**R29**	0.079	**0.572**	-0.190
G30	**0.437**	0.055	-0.210	**V15**	0.008	-0.010	**0.452**	**V50**	-0.180	-0.060	**0.528**	**R30**	0.032	**0.404**	-0.030
G31	0.221	-0.070	**0.282**	**V16**	0.165	-0.050	**0.374**								
G32	**0.518**	0.035	-0.050	**V17**	0.151	-0.020	**0.336**								
G33	**0.461**	0.074	-0.080	**V18**	-0.030	-0.030	**0.478**								
G34	**0.352**	-0.020	0.109	**V19**	0.180	0.005	**0.318**								
G35	**0.406**	0.007	0.111	**V20**	0.032	0.004	**0.440**								

In the MIRT analysis, 98% of the grammar items have the highest a-coefficient on a1; 60% of the vocabulary items have the highest a-coefficient on a2; and 76.7% reading items have the highest a-coefficient on a3. In the factor analysis with oblique rotation, 94% of grammar items and the same percentage of vocabulary items have the highest loading on the first and third factor respectively, and 76.7% of reading items have the highest loading on the second factor. About 16.0% of total variance is extracted by the first factor.

Table 4.4. Percentage of the Highest Loading Items on One Dimension/Factor

	Dimension		
%	1	2	3
Grammar	98	------	------
Vocabulary	------	------	60
Reading	------	76.7	------
	Factor		
%	1	2	3
Grammar	94	------	------
Vocabulary	------	------	94
Reading	------	76.7	------

As noted in the previous chapter, a portion of the 130 items in the original test were selected according to each equating design. The item/common item numbers for each design are shown in Figures 3.1-3.3 in the previous chapter. Three levels of test length were selected for the equating designs. At each level of test length (120 items, 96 items and 72 items), the same set of items were used for all the equating designs that differ in difficulty differences or methods of equating; however, the items were grouped differently when designs are different. Item model fit was estimated at each level of test length according to methods presented in the previous chapter, and the results, exhibited in Table 4.5, are discussed in the next chapter.

Table 4.5. Percentage of Items Fitting the Model

Test length	chi-square≦60 (3*df)	P>0.01
120 items	87%	75%
96 items	90%	70%
72 items	86%	67%

*for all the items, the degree of freedom (df) for chi-square estimate is 20

II. Item selection for each design

The difference in the average difficulty between the two forms of each design is shown in Table 4.6. As the original items were actually developed for one test, and were not intended to be dramatically different in difficulty, it was challenging to select items and separate them into two forms whose average difficulty levels were different by 1.5 units on the θ-scale. Thus for the designs targeted to have a difficulty level difference of 1.5, the targets are not met; the differences between the two forms are particularly smaller in the common-item designs and also smaller when more items are included in each form (such as in data sets that have 120 total items). This may affect the results of the study. The item numbers in each section (G, V and R) maintain the same ratio as in the original test; item/common item numbers of each section are shown in Table 4.7.

The chi-squares of item fit for the linking items are displayed in tables 4.8-10. For designs that are of different difficulty difference between the test forms, the linking items in common-item and common-common designs are also different in a few items. The linking items used in each of the common-item designs are shown in tables 4.8-10; each table is for different test length. The linking items in the common-common design were included in the correspondent common-item. The items that do not fit the 3-PL IRT model according the rule of thumb that was mentioned before (chi-square<3*df) are marked. Most of the linking items (90% in average) fit the 3-PL IRT model.

Table 4.6. Difference between the Averaged Item Difficulty

	N_{item}=120			N_{item}=96			N_{item}=72		
Target difference	0.50	1.00	1.50	0.50	1.00	1.50	0.50	1.00	1.50
Common -item	0.48	0.95	1.14	0.52	0.92	1.29	0.52	0.99	1.34
Common -common	0.50	0.95	1.21	0.48	0.97	1.40	0.48	1.03	1.46
Common -group	0.51	0.99	1.28	0.51	0.99	1.45	0.51	1.07	1.51

Table 4.7. Item/Common-Item Numbers for Each Design

Total item=120						
Design	Common-group		Common-item		Common-common	
	Number of Items	Items in anchor	Number of Items	Items in anchor	Number of Items	Items in anchor
Grammar	23	0	28	10	27/25*	6
Vocabulary	24	0	29	10	26/28*	6
Reading	13	0	13	0	13	0
Total	60	0	70	20	66	12
Total item=96						
Design	Common-group		Common-item		Common-common	
	Number of Items	Items in anchor	Number of Items	Items in anchor	Number of Items	Items in anchor
Grammar	19	0	23	8	21	4
Vocabulary	19	0	23	8	21	4
Reading	10	0	10	0	10	0
Total	48	0	56	16	52	8
Total item=72						
Design	Common-group		Common-item		Common-common	
	Number of Items	Items in anchor	Number of Items	Items in anchor	Number of Items	Items in anchor
Grammar	15	0	18	6	17/16*	3
Vocabulary	15	0	18	6	16/17*	3
Reading	6	0	6	0	6	0
Total	36	0	42	12	39	6

Table 4.8 Linking Item Fit for the Equating Design of 120 Item Tests (df=20)

Item	Chi-Square	Difficulty Difference			Item	Chi-Square	Difficulty Difference		
		0.5	1.0	1.5			0.5	1.0	1.5
G51	22.6			X	V108	20.9	X		
G52	20.9	X			V109	91	X		
G55	28.7	X	X		V113	31.9	X		
G58	22.4		X	X	V120	24.8			X
G59	32.4		X	X	V126	12.3	X	X	
G60	27.4			X	V127	76.1*	X		
G67	24.8	X			V128	30.6		X	X
G69	44.4	X			V132	35.7	X		X
G72	38.6			X	V134	72*			X
G73	100.0*		X		V135	43.2			X
G74	44.3	X			V139	28.5		X	X
G75	79.2*	X	X		V142	49.4	X	X	X
G78	33.8	X		X	V143	26.9	X	X	
G81	55.1	X			V144	40	X		X
G83	69.7*	X		X	V145	42.7		X	
G84	20.6	X	X	X	V146	17.6		X	
G88	39.8		X	X	V147	39.4	X		
G90	60.4		X	X	V148	58.5		X	X
G92	28.8		X		V149	42.6		X	X
G94	33.5		X						

*Chi-square>3*df, item does not fit IRT 3-PL model.

Table 4.9 Linking Item Fit for the Equating Design of 96 Item Tests (df=20)

Item	Chi-Square	Difficulty Difference			Item	Chi-Square	Difficulty Difference		
		0.5	1.0	1.5			0.5	1.0	1.5
G52	34.2	X			V101	25.1			X
G58	25.5		X	X	V106	19.1	X	X	
G60	21.1	X	X	X	V114	21.6	X	X	X
G61	32.3		X	X	V117	45.1		X	X
G62	30.4		X		V120	31.7		X	X
G69	47.1	X			V126	19.5	X	X	
G72	39.3			X	V128	24.5			X
G77	34.9			X	V132	41.1	X	X	
G79	32.8			X	V134	83.7*	X	X	
G81	50.5	X			V135	54.8			X
G82	29.7	X			V137	30.2	X		
G83	70.7*		X		V139	33.9	X		X
G85	103.0*	X		X	V142	54.0	X		
G86	35.1		X		V148	66.0*		X	
G88	53.8	X	X		V149	37.9			X
G89	33.8		X	X					
G90	50.3	X							

*Chi-square>3*df, item does not fit IRT 3-PL model.

Table 4.10 Linking Item Fit for the Equating Design of 72 Item Tests (df=20)

Item	Chi-Square	Difficulty Difference			Item	Chi-Square	Difficulty Difference		
		0.5	1.0	1.5			0.5	1.0	1.5
G58	25.0	X			V101	23.8		X	
G60	26.8	X	X	X	V106	29.7	X		
G61	37.7	X			V118	19.5		X	
G75	79.2*		X		V120	23.5		X	X
G79	35.0	X			V128	30.5	X		X
G83	69.2*	X	X	X	V132	41.1		X	X
G85	94.4*			X	V134	81.1*		X	X
G86	44.8		X	X	V135	59.3	X		
G88	36.3	X	X	X	V137	38.9	X		
G89	25.4		X	X	V142	51.4			X
					V143	36.5	X		
					V148	55.9	X	X	
					V149	37.9			X

*Chi-square>3*df, item does not fit IRT 3-PL model.

III. Regressions between the two sets of item parameters

Item parameters were estimated using the original data and the data sets for the different equating designs. The parameters estimated using the original data are regarded as "real" values and those estimated using the data samples designed for each equating method are called equated parameters. Figures 4.1-4.9 illustrate the scatter plots between the "real" values and the equated values of the parameters. The graphs indicate that difficulty (b) parameter estimation is the most stable across different designs. In general, the values estimated through the common-group designs tend to be high and those estimated by the common-item design tend to be low. The values estimated by common-common design fall in the middle. Compared with the scatter plots of the b parameters, the scatter plots of the discrimination (a) parameters show more variance in their estimates; and even bigger variance in the lower asymptote (c) parameters. The same trend is also indicated in the correlation coefficients.

Table 4.11 demonstrates the correlation coefficients for the a, b and c (slope, difficulty and asymptote) parameters between their "real" and equated values. The correlation coefficients are highest for the difficulty parameters (around 0.97-0.99), lower for slope parameters (around 0.90-0.93) and lowest for lower asymptote parameters (0.6-0.8). Table 4.12 lists the significance of ANOVA analysis between the means of each level. None of the parameters has significant difference in correlation coefficients between different designs. "a" and "b" parameters are significantly different when test lengths are different, "c" parameters are significantly different when item difficulty between the forms are different. Tables 4.13 and 4.14 list the slope and interception from the regressions respectively.

Table 4.11. Correlations Between the Two Sets of Parameters

Differ-	Common-item			Common-common			Common-group		
ence	a	b	c	a	b	c	a	b	c
	120 items								
0.5	0.920	0.981	0.815	0.914	0.969	0.724	0.928	0.979	0.804
1.0	0.907	0.981	0.835	0.920	0.984	0.825	0.920	0.972	0.804
1.5	0.863	0.970	0.775	0.901	0.967	0.757	0.907	0.983	0.816
	96 items								
0.5	0.906	0.967	0.827	0.894	0.949	0.666	0.898	0.966	0.672
1.0	0.849	0.973	0.701	0.899	0.983	0.788	0.888	0.972	0.741
1.5	0.916	0.992	0.852	0.862	0.987	0.803	0.868	0.975	0.754
	72 items								
0.5	0.815	0.980	0.718	0.820	0.976	0.652	0.877	0.986	0.765
1.0	0.914	0.988	0.770	0.882	0.987	0.831	0.920	0.989	0.841
1.5	0.906	0.992	0.900	0.879	0.988	0.863	0.850	0.983	0.820

Table 4.12. ANOVA Significance of the Correlation Coefficients

Levels of Comparison	**Test Length**			**Difficulty Difference**			**Designs**		
Parameters	a	b	c	a	b	c	a	b	c
P_{ANOVA}	0.041*	0.024*	0.323	0.489	0.083	0.021*	0.810	0.735	0.581

The difference between the correlations at different levels is significant ($p<0.05$)

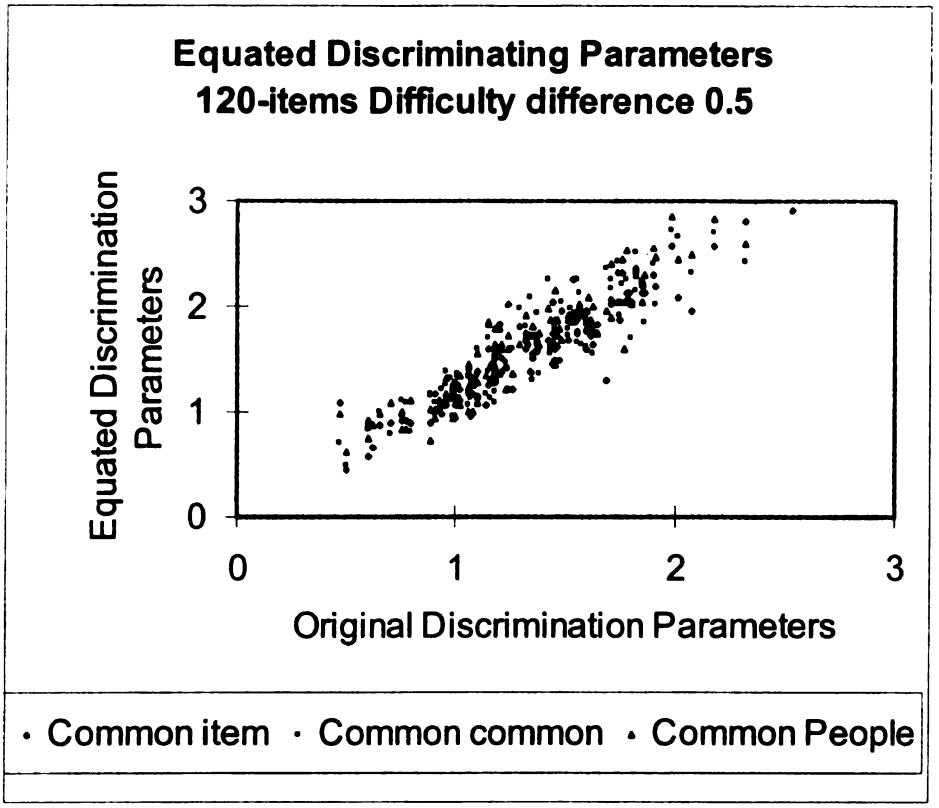
Table 4.13. Slope of the Regression Function

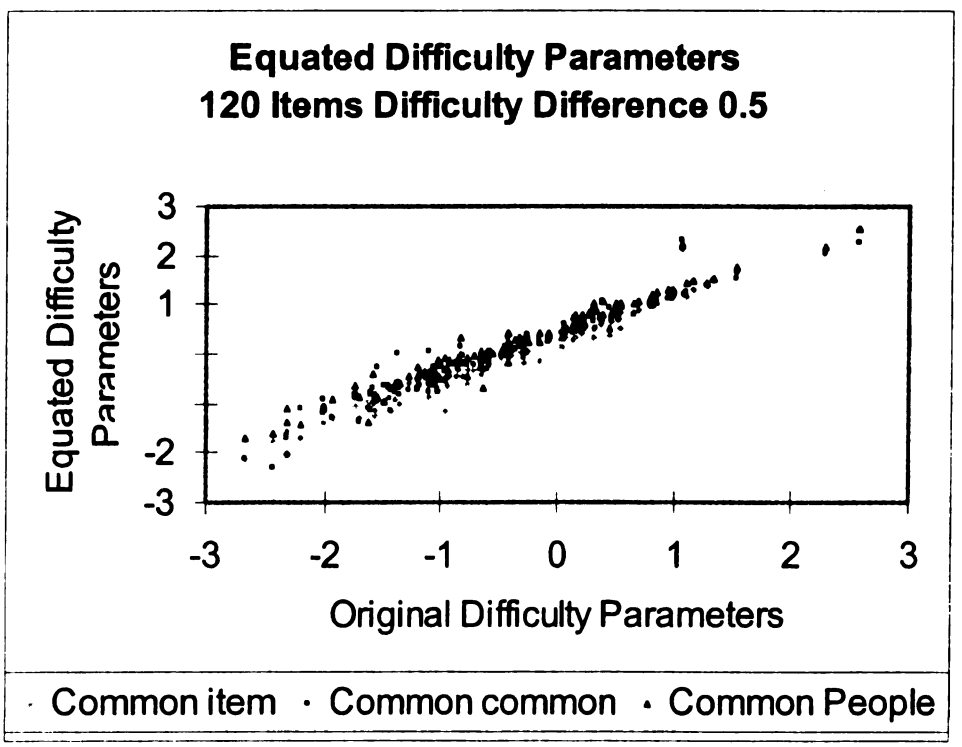
Difficulty	**Common-item**			**Common-common**			**Common-group**		
ference	a	b	c	a	b	c	a	b	c
	120 items								
0.5	1.13	0.86	0.71	1.13	0.82	0.64	1.26	0.81	0.66
1.0	1.10	0.81	0.69	1.04	0.83	0.68	1.18	0.82	0.59
1.5	1.08	0.84	0.71	1.00	0.89	0.62	1.13	0.89	0.67
	96 items								
0.5	1.06	0.90	0.70	1.08	0.83	0.61	1.26	0.75	0.67
1.0	1.05	0.89	0.67	1.09	0.83	0.63	1.07	0.83	0.59
1.5	1.09	0.86	0.69	1.09	0.83	0.66	1.13	0.89	0.63
	72 items								
0.5	0.95	0.89	0.70	0.97	0.86	0.67	1.13	0.80	0.61
1.0	1.10	0.84	0.73	1.25	0.85	0.73	1.18	0.86	0.82
1.5	1.03	0.83	0.80	1.12	0.82	0.70	1.28	0.87	0.79

Table 4.14. Intercept of the Regression Function

Difficulty	Common-item			Common-common			Common-group		
Difference	a	b	c	a	b	c	a	b	c
	120 items								
0.5	0.05	0.32	0.09	0.10	0.40	0.16	-0.00	0.51	0.10
1.0	0.14	0.52	0.09	0.18	0.62	0.10	0.03	0.85	0.13
1.5	0.16	0.90	0.14	0.14	0.98	0.11	0.04	1.30	0.12
	96 items								
0.5	0.05	0.44	0.15	0.12	0.45	0.17	0.09	0.54	0.11
1.0	0.11	0.65	0.16	0.08	0.60	0.09	0.10	0.82	0.11
1.5	0.12	0.73	0.08	0.17	0.90	0.09	0.06	1.26	0.12
	72 items								
0.5	0.24	0.34	0.07	0.30	0.33	0.14	0.12	0.47	0.10
1.0	0.10	0.47	0.06	-0.10	0.58	0.06	0.02	0.83	0.06
1.5	0.28	0.78	0.08	0.17	0.91	0.09	-0.10	1.26	0.09

Figure 4.1 Item parameters for test length 120, difficulty difference 0.5





(Figure 4.1 Continued)

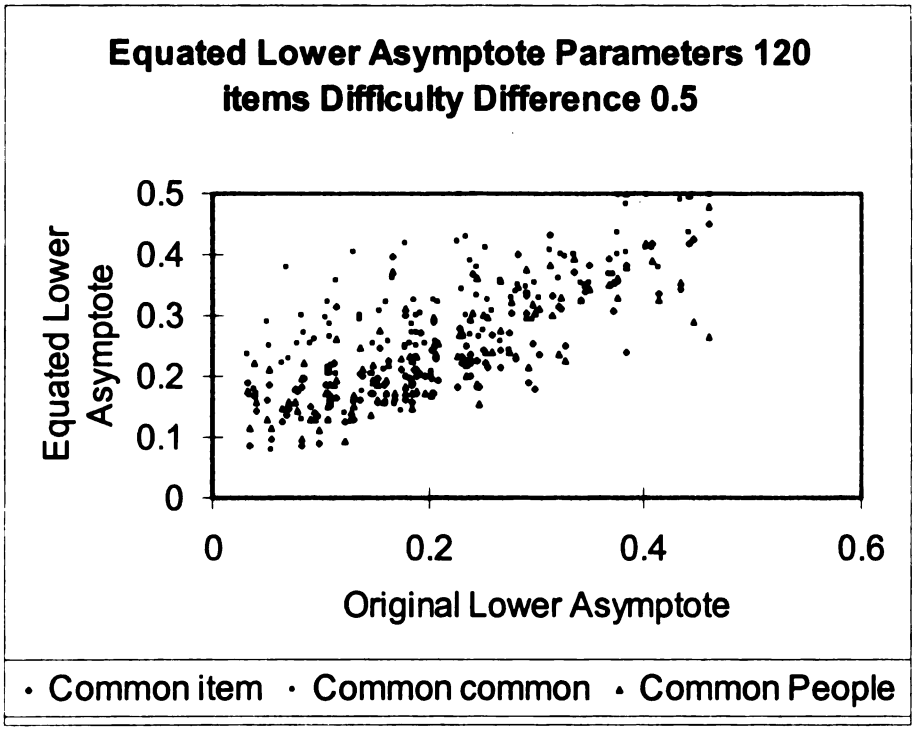
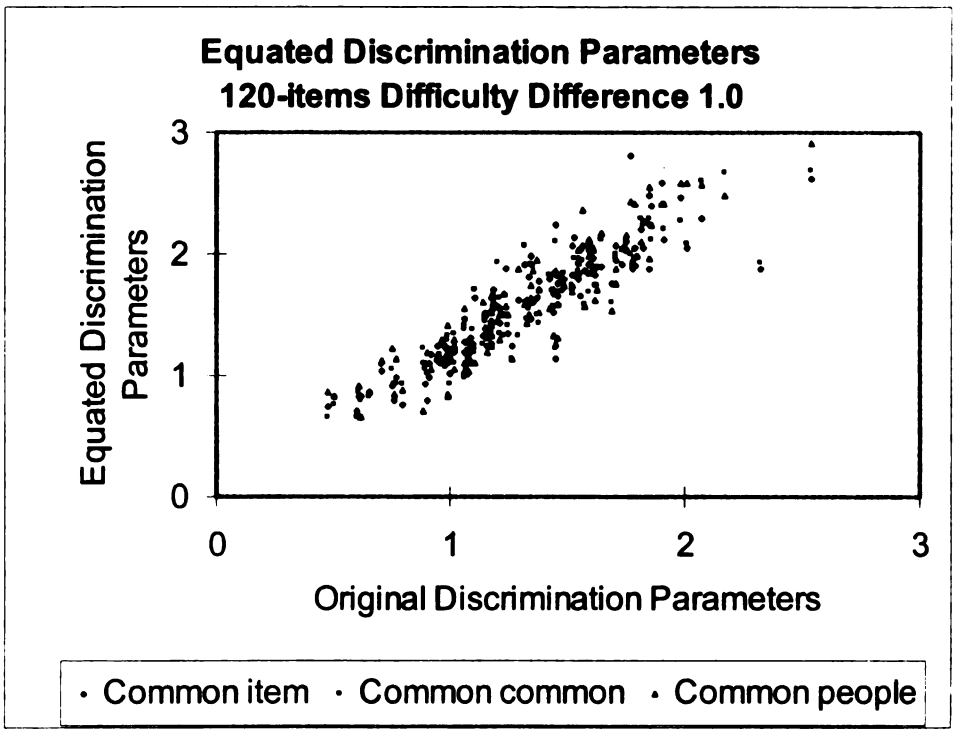


Figure 4.2. Item parameters for test length 120, difficulty difference 1.0



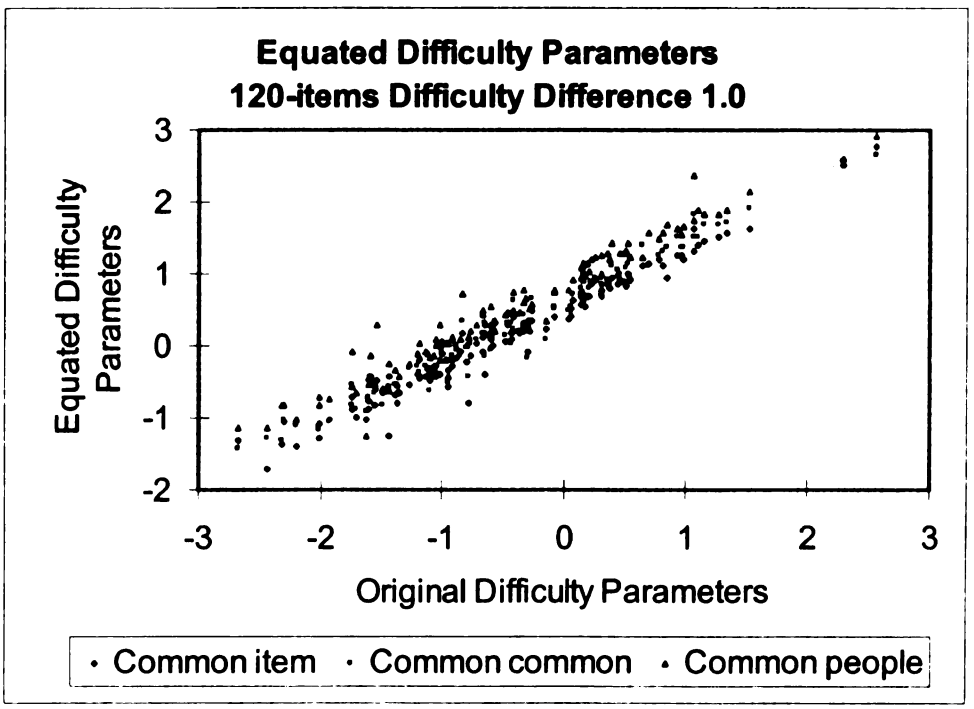


Figure 4.2 (Continued)

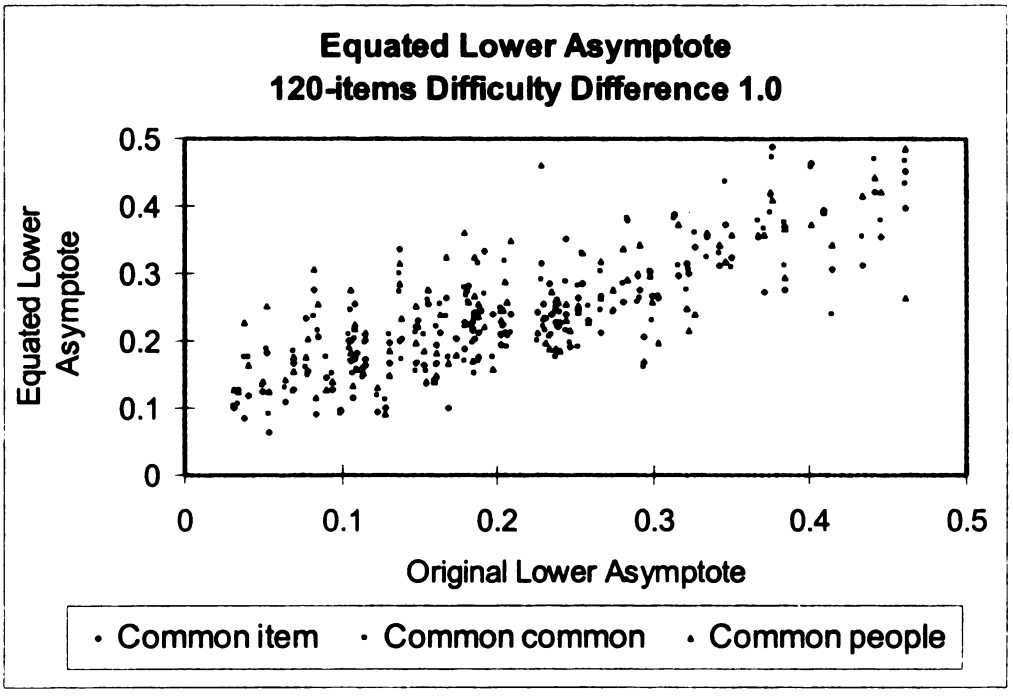
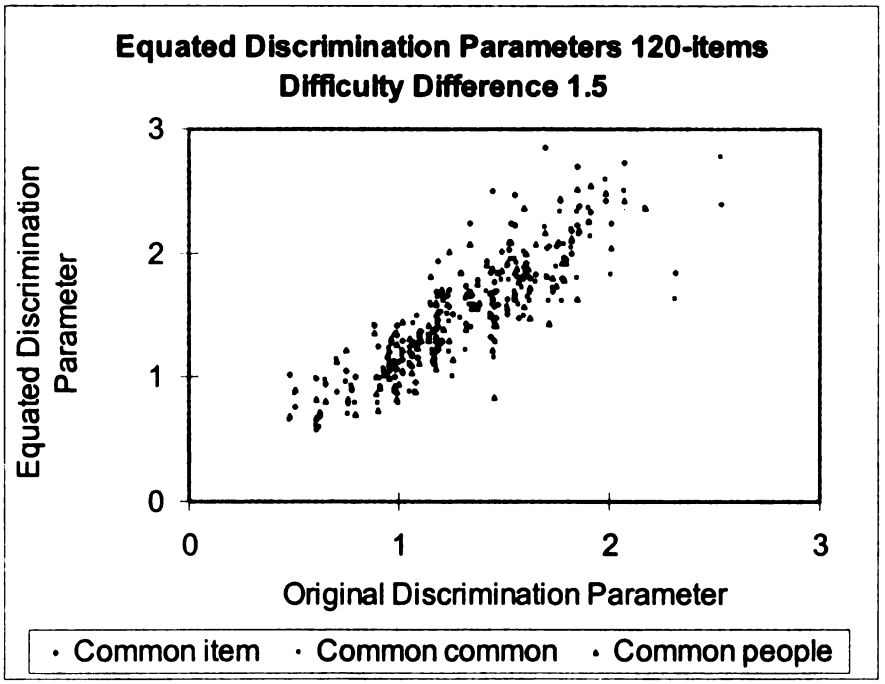


Figure 4.3. Item parameters for test length 120, difficulty difference 1.5



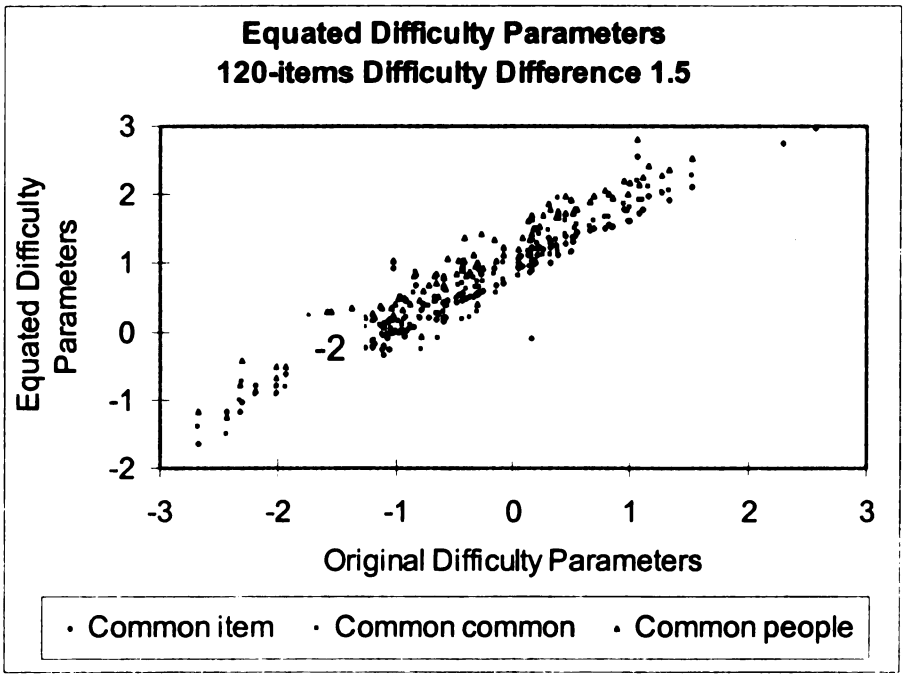


Figure 4.3 (Continued)

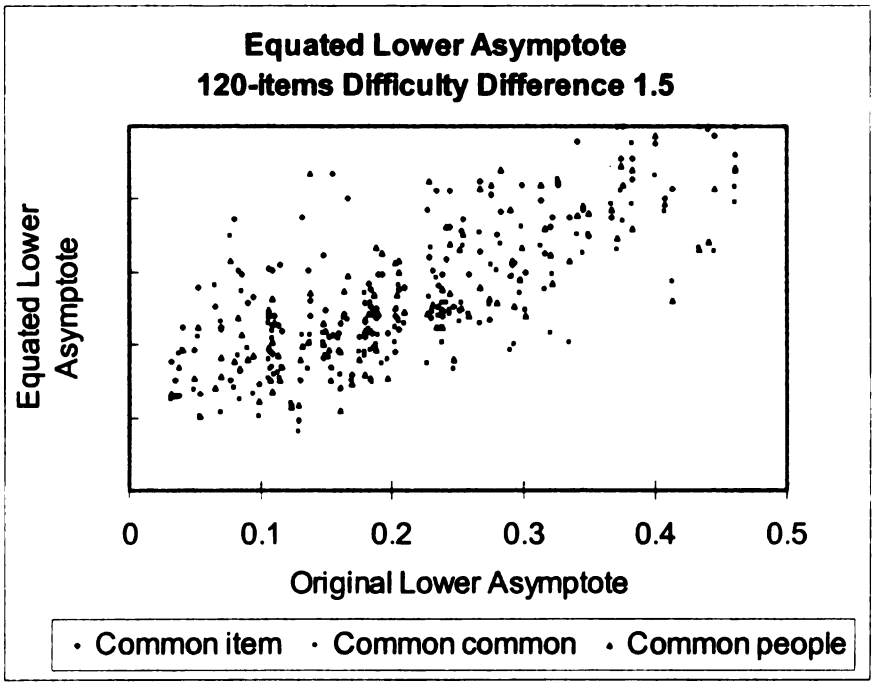
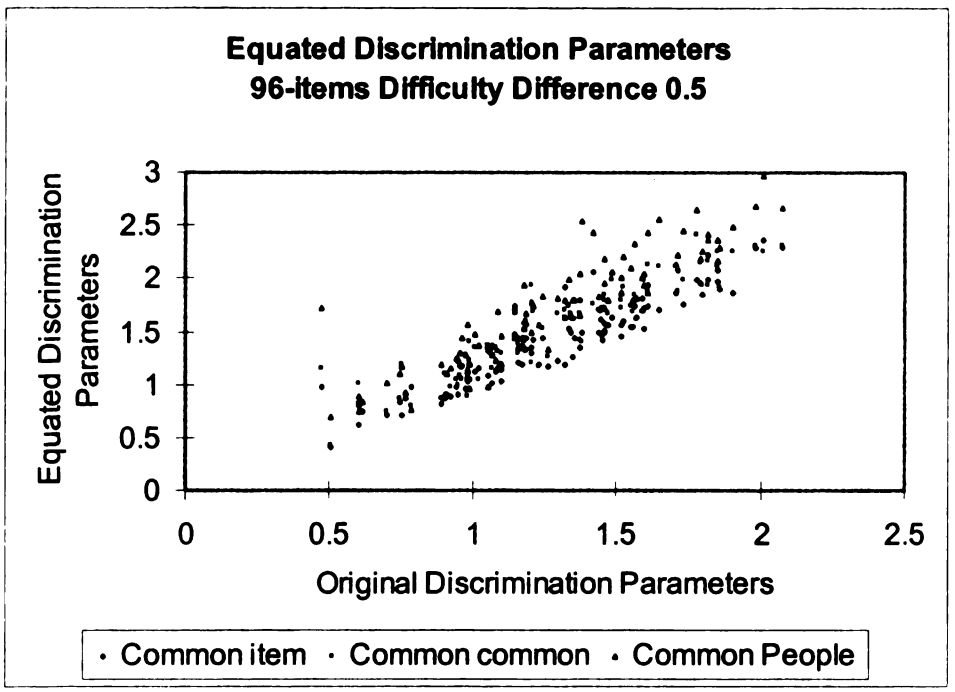


Figure 4.4. Item parameters for test length 96, difficulty difference 0.5



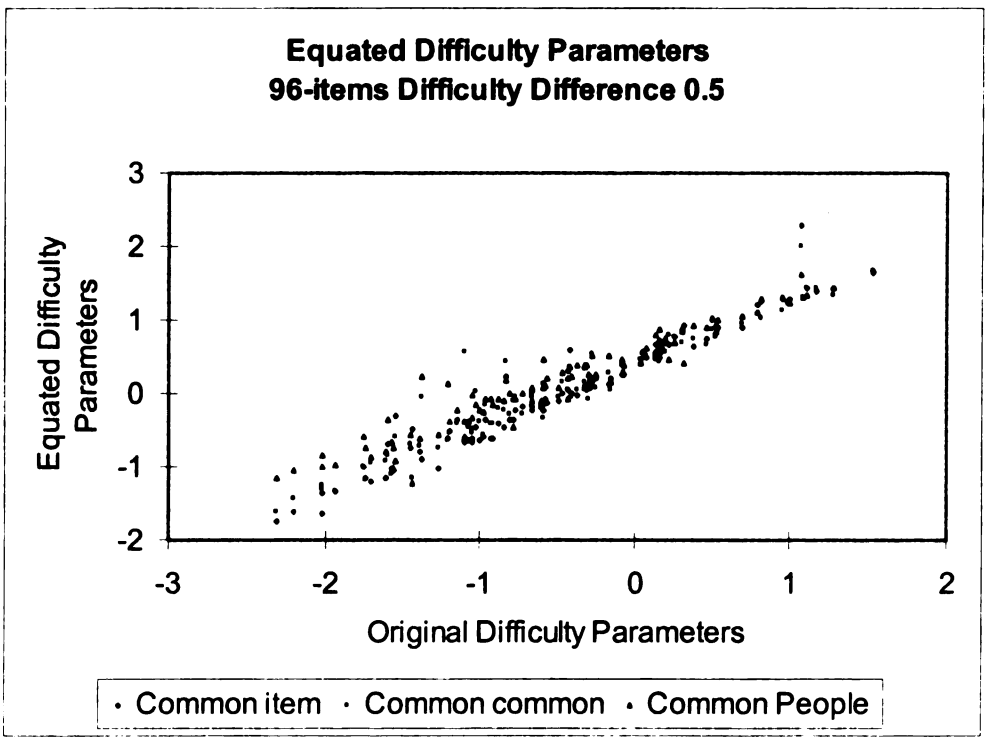


Figure 4.4 (Continued)

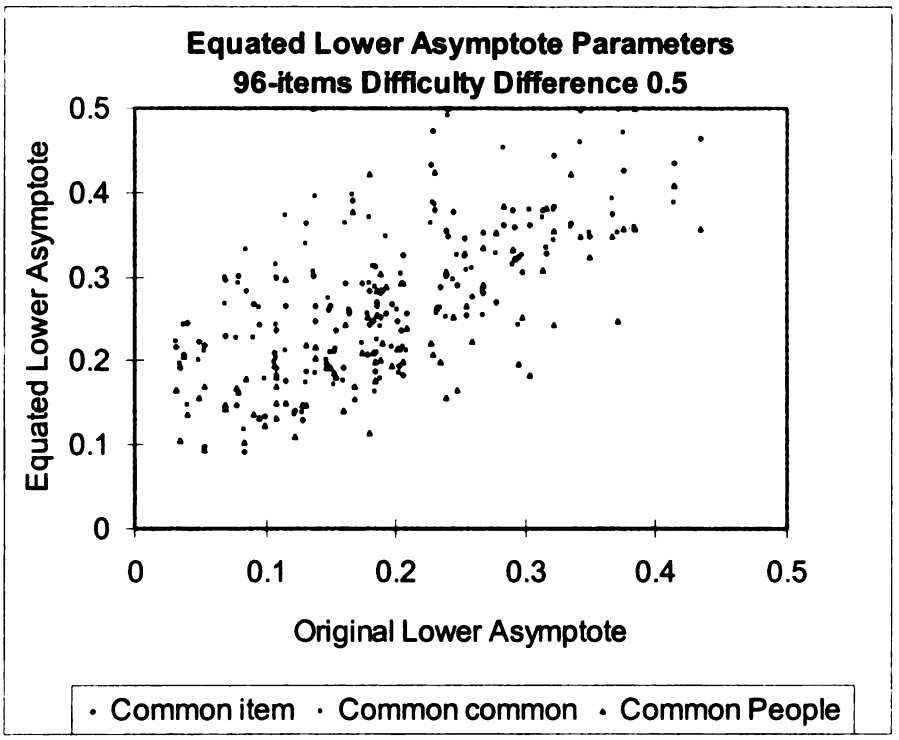
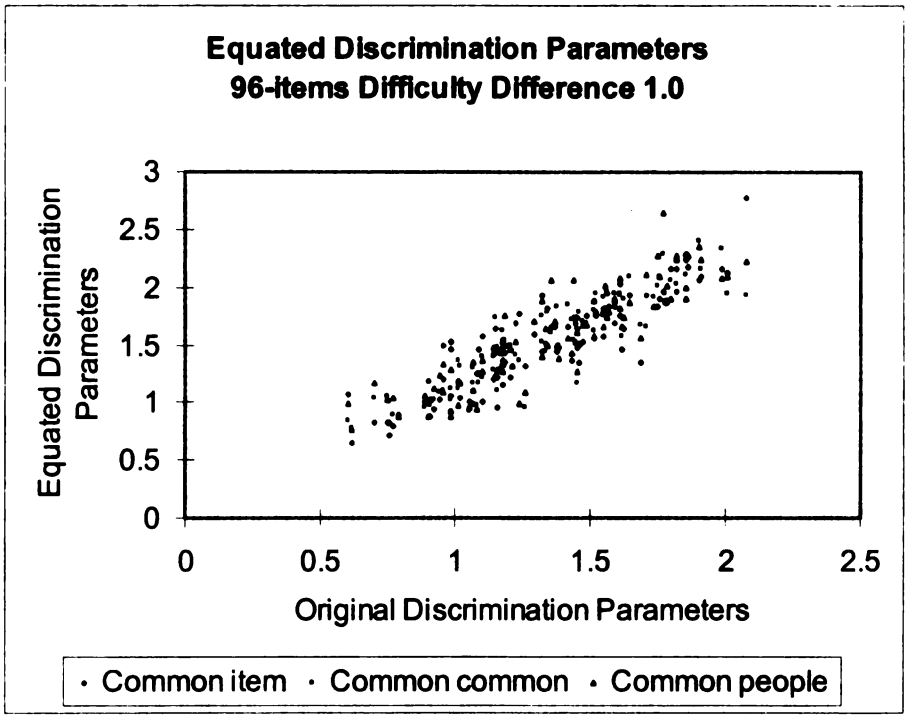


Figure 4.5. Item parameters for test length 96, difficulty difference 1.0



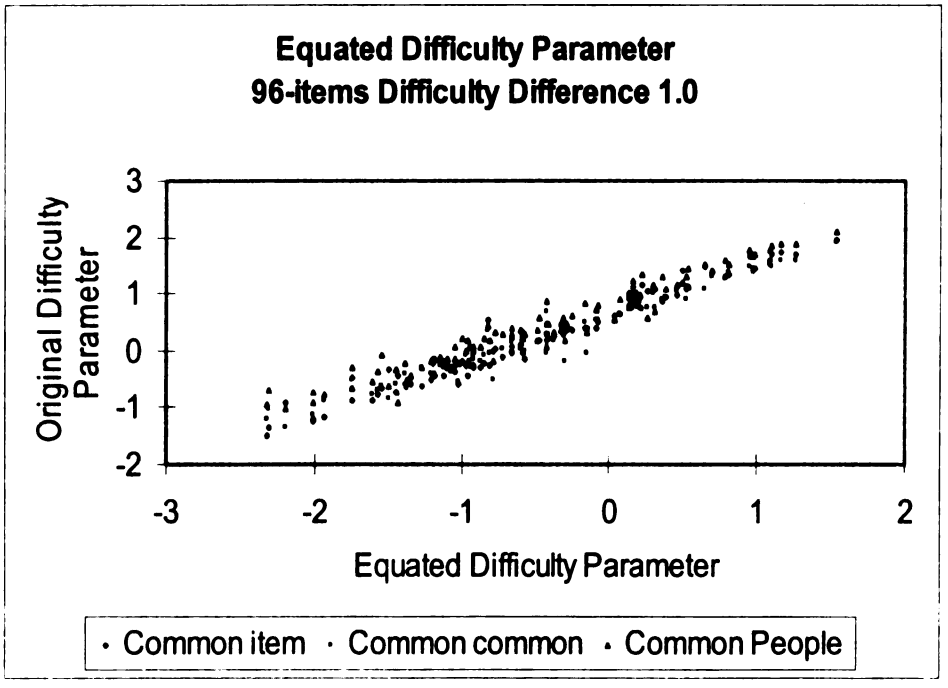


Figure 4.5 (Continued)

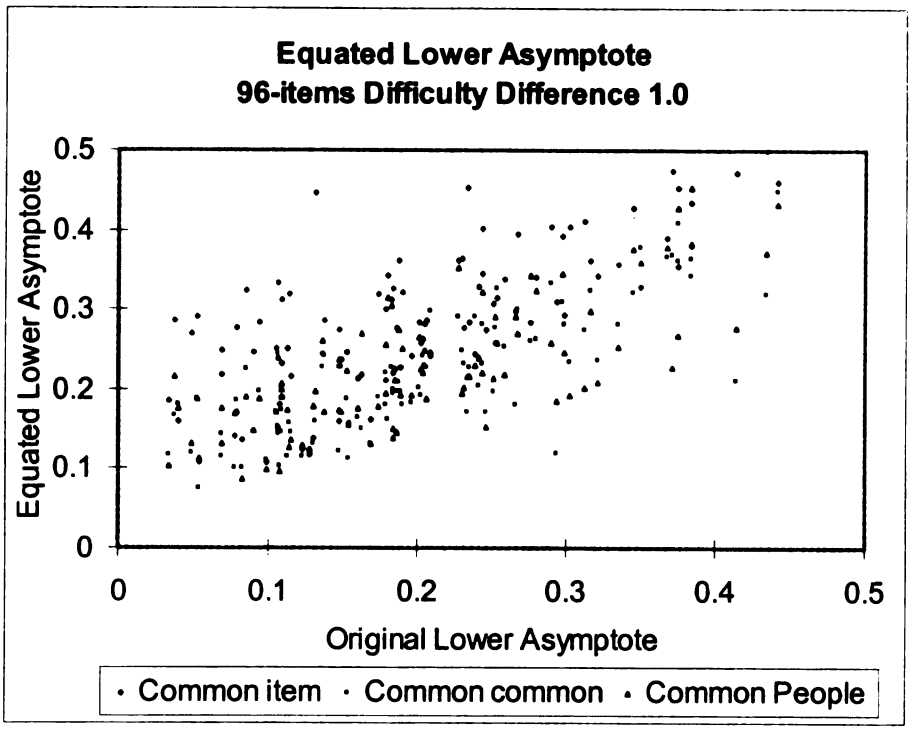
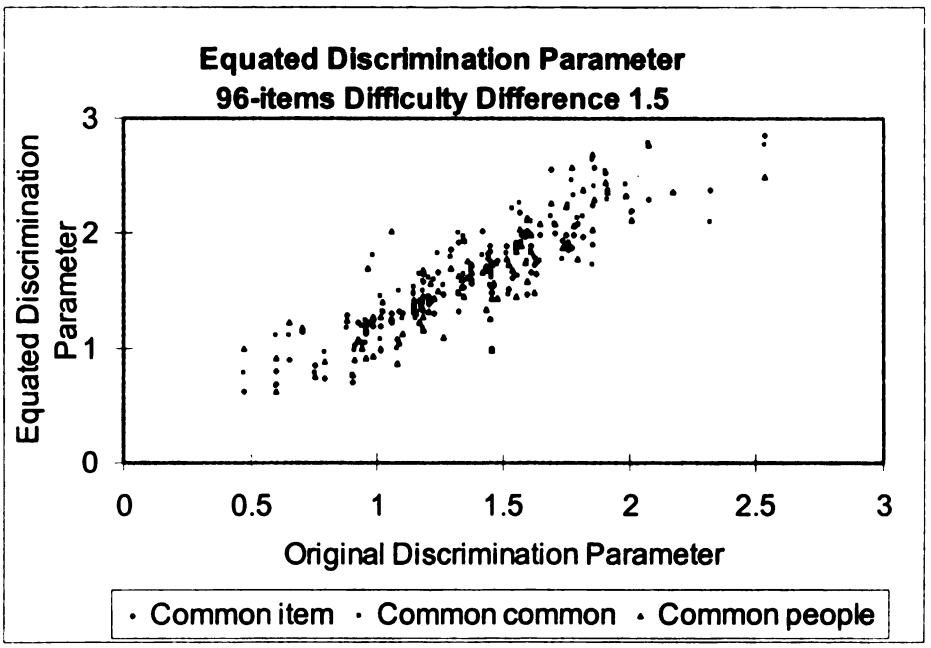


Figure 4.6. Item parameters for test length 96, difficulty difference 1.5



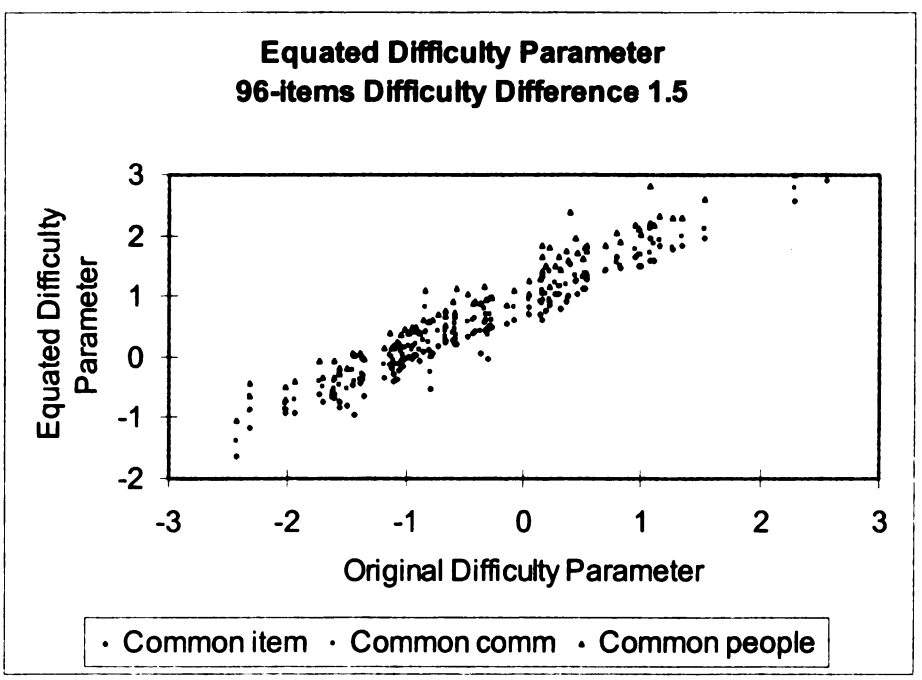


Figure 4.6 (Continued)

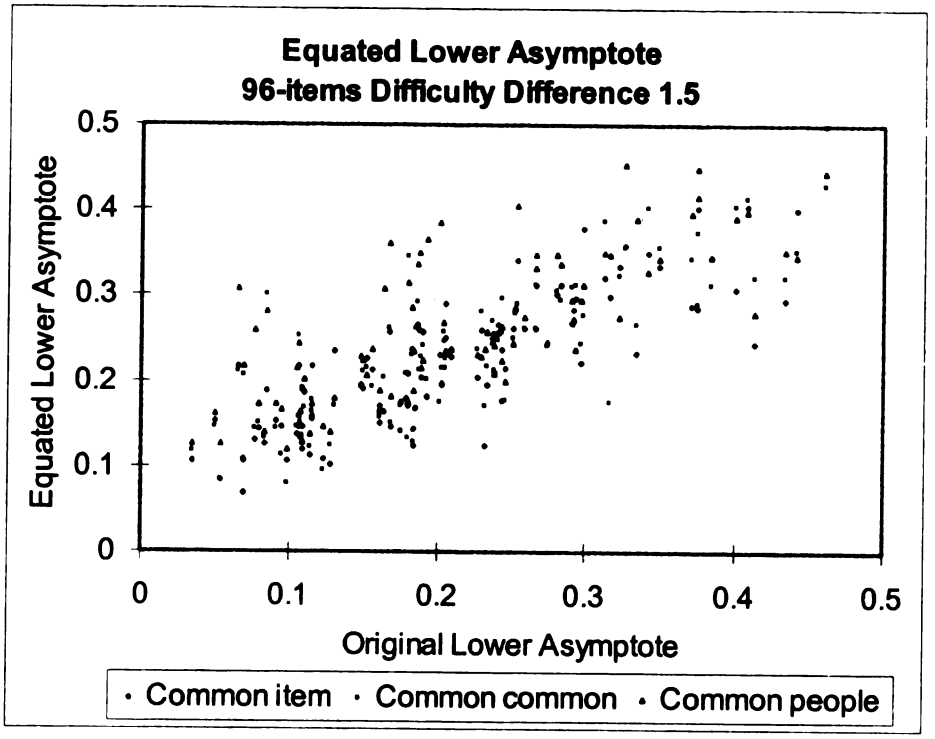
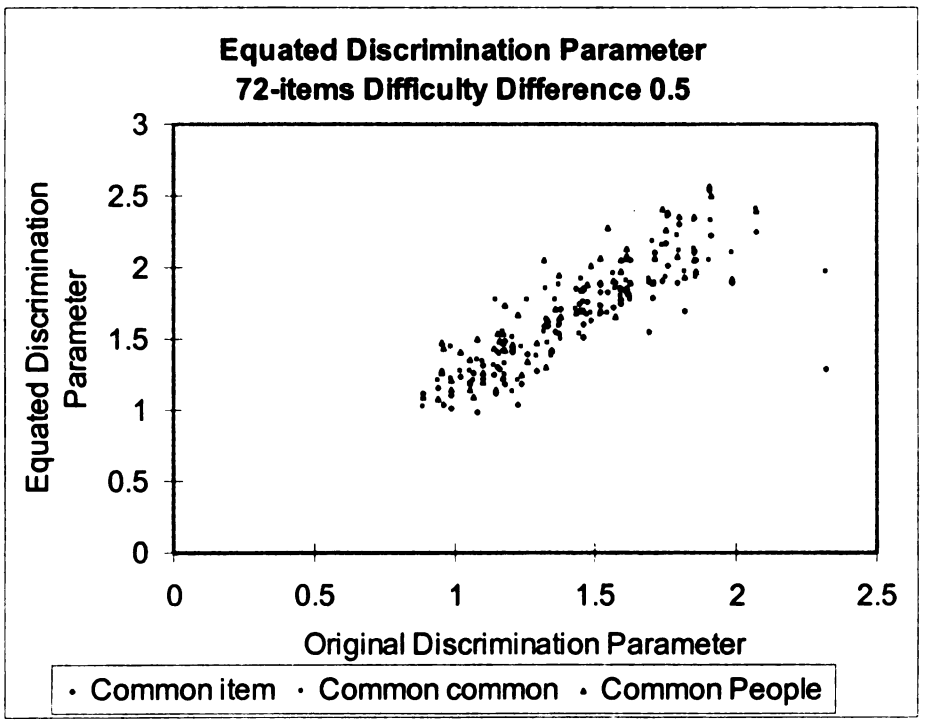


Figure 4.7. Item parameters for test length 72, difficulty difference 0.5



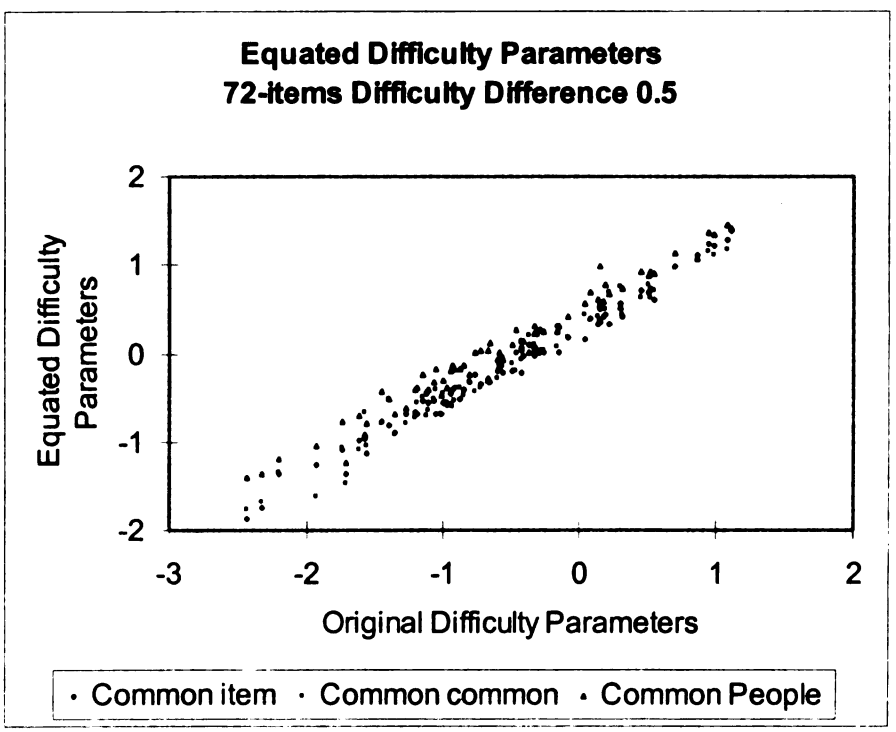


Figure 4.7 (Continued)

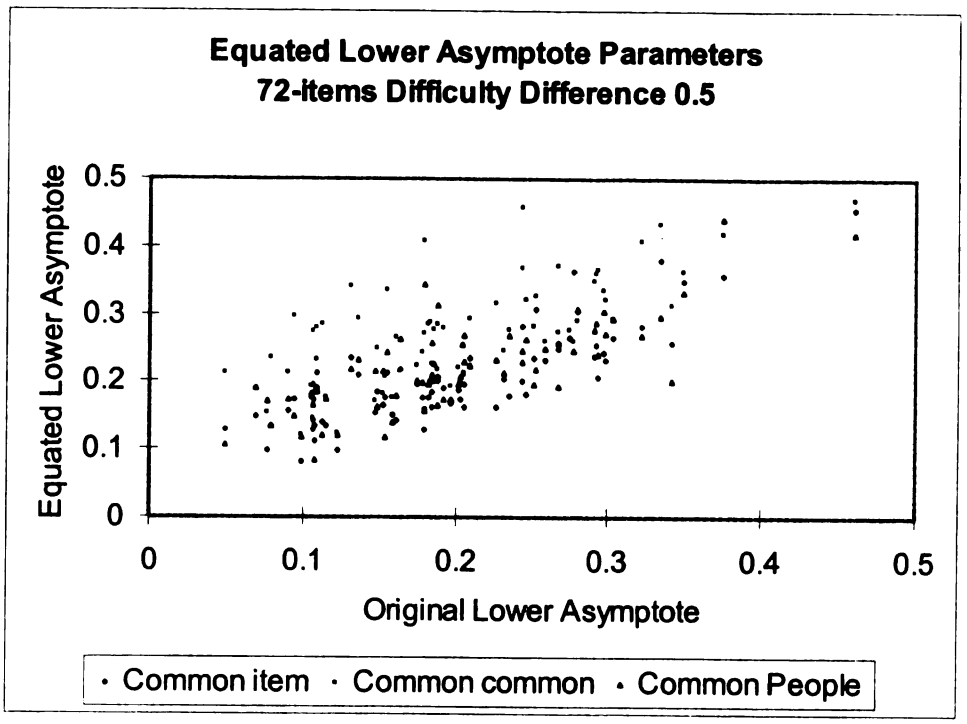
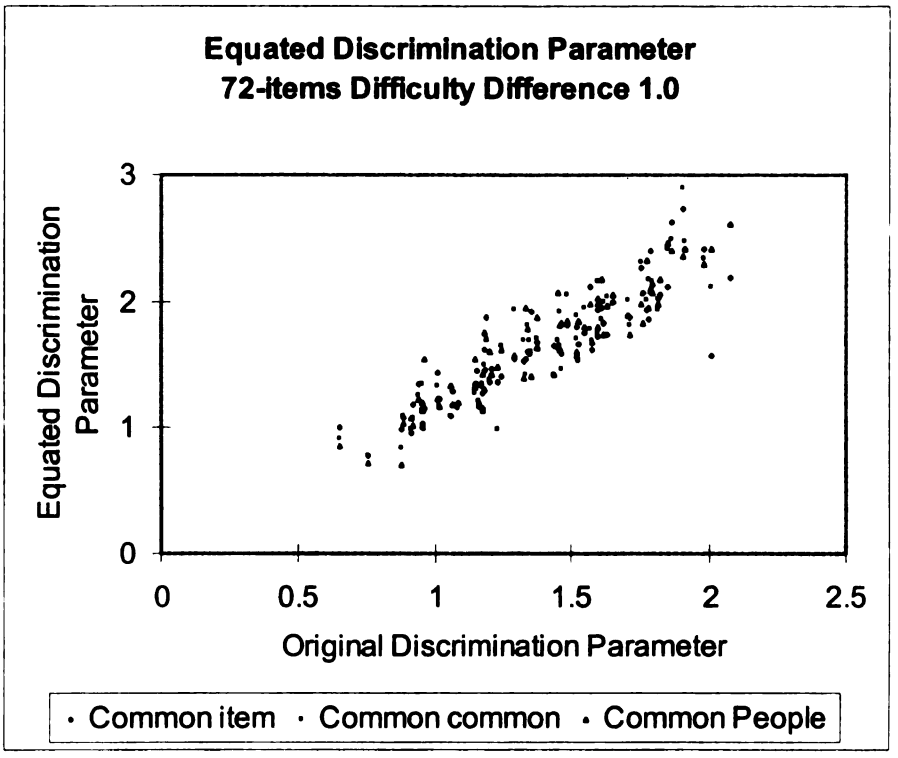


Figure 4.8. Item parameters for test length 72, difficulty difference 1.0



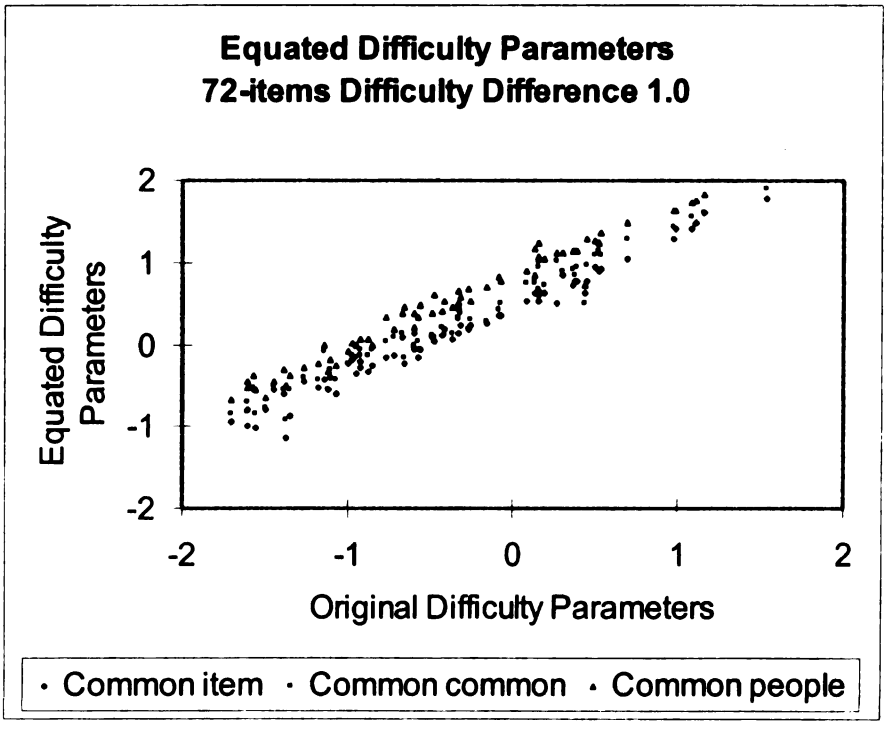


Figure 4.8 (Continued)

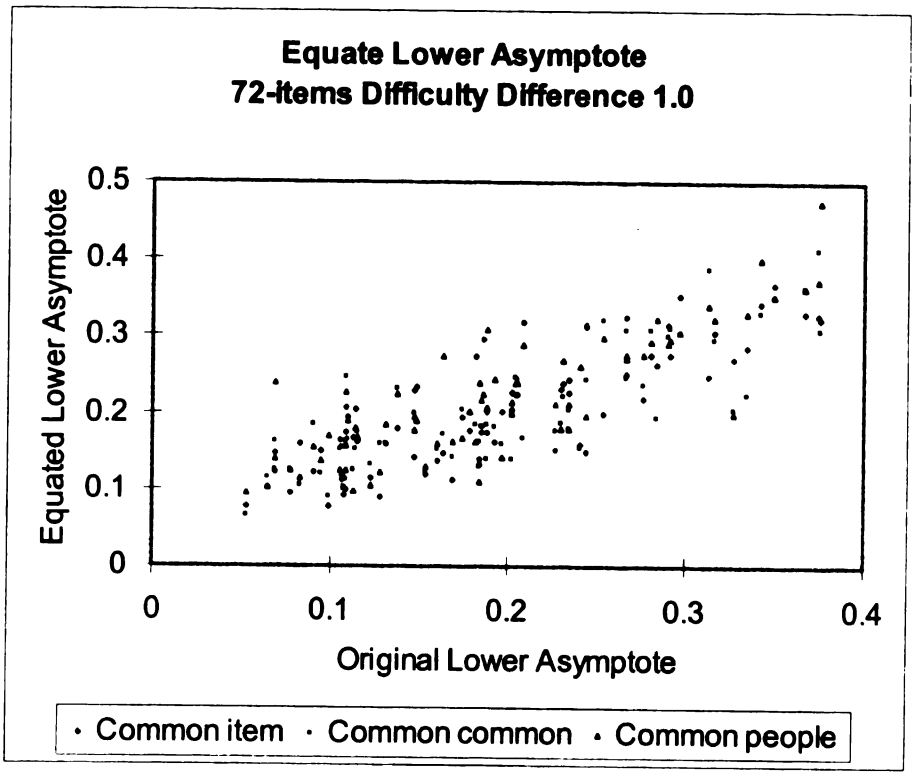
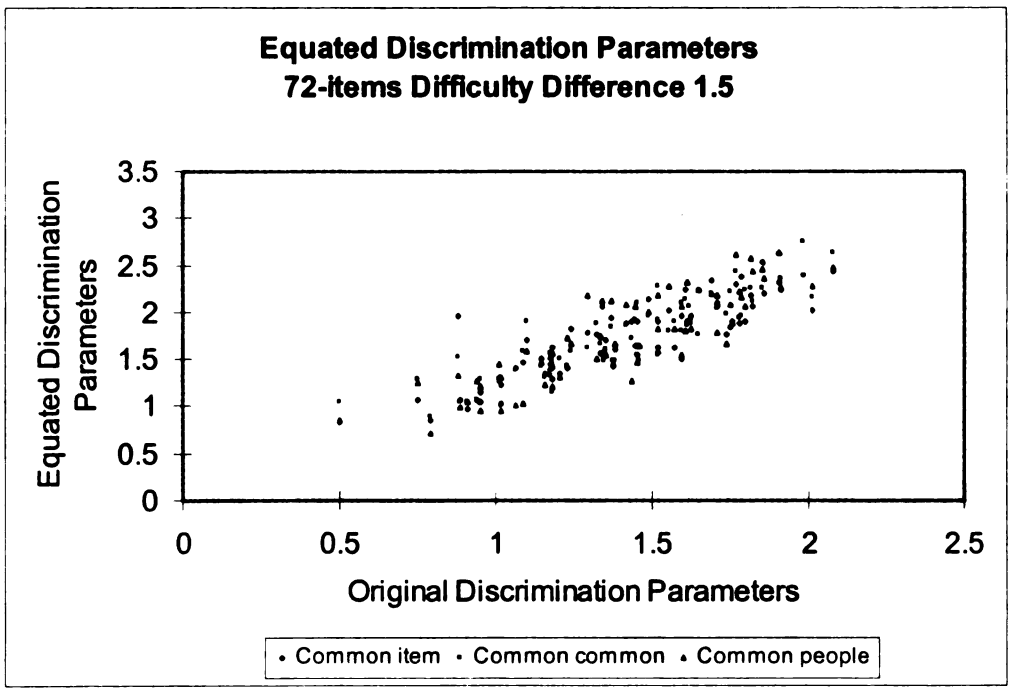


Figure 4.9. Item parameters for test length 72, difficulty difference 1.5



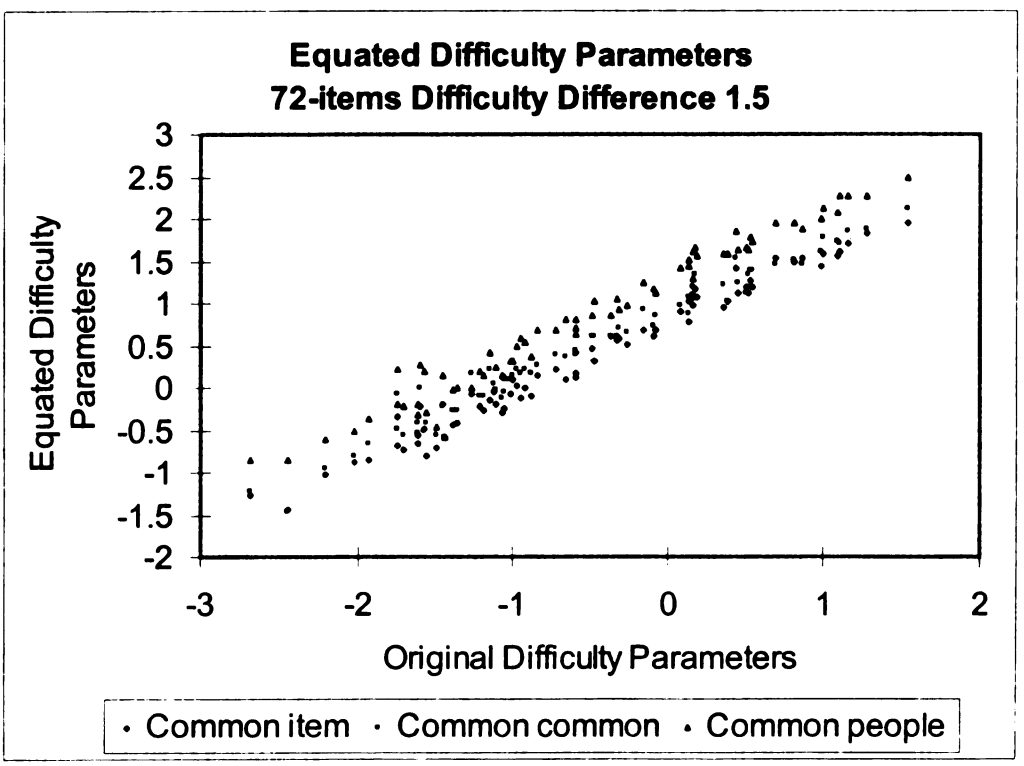
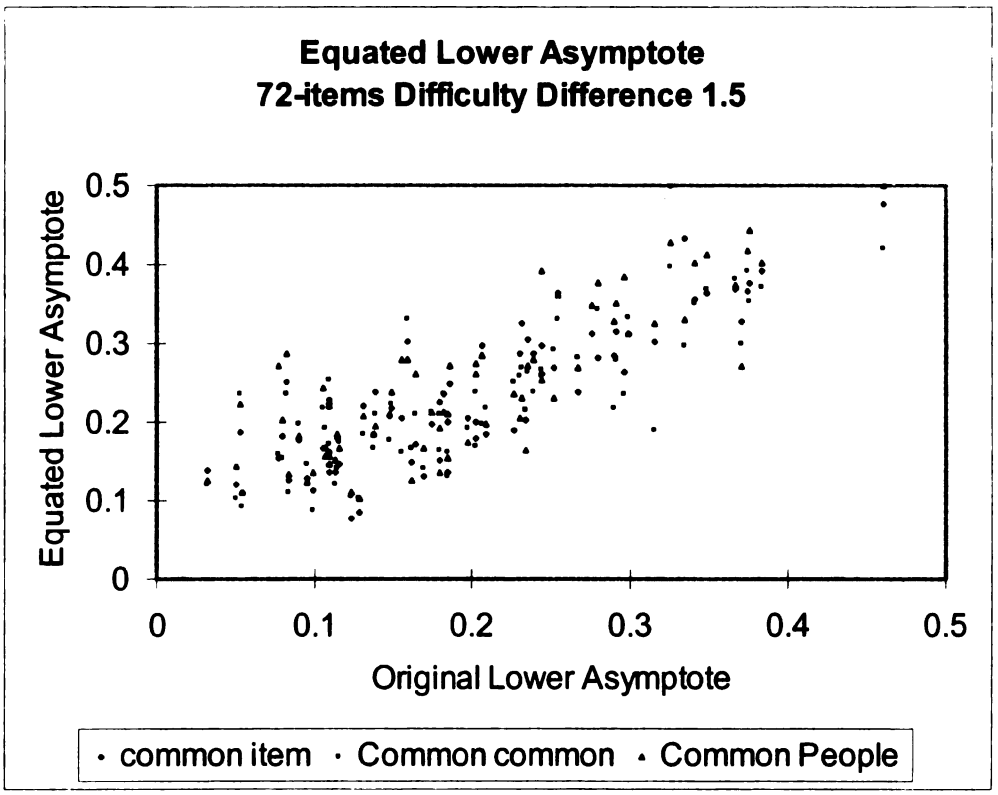


Figure 4.9 (Continued)



IV. Regression between the "real scores" and equated scores

As noted in chapter 3, sub-samples were selected from the original data set according to each equating design (as is shown in Figures 3.1-3.3). Some of the responses in each of the sub-samples were deleted according to the designed data matrices. The data were then processed through BILOG-MG3 as vertical equating data. BILOG-MG3 calibrates item parameters and ability scores concurrently with marginal maximum likelihood (MML) estimation.

The "real scores" in each data set was divided into intervals of 0.1 standard deviation, the mean of the equated scores for examinees whose "real scores" fall in the corresponding interval were calculated. The scatter plots between the mean of the equated scores and the interval of the "real scores" for each design are shown in Figures 4.10-4.12. Figure 4.10 is for tests with 120 items in total, Figure 4.11 is for 96 items and Figure 4.12 is for 72 items. In each figure, the first column is for common-item design, second column for common-common design and the third for common-group design. The first row is for the designs when the two forms differ in averaged difficulty for 0.5 SD, the second row when they differ in 1.0 SD and the third when differ in 1.5 SD. BILOG-MG3 assigned the lower ability group as the control group, thus the means of thetas in lower ability groups are zero; the means of the thetas for higher ability groups are higher in about 0.5, 1.0 or 1.5 standard deviations according to the design of the equating. The correlations between equated and "real" scores are calculated and listed in Table 4.15. The results show no obvious difference across different designs and different test length. However, the correlation coefficients are significantly ($p<0.01$) different when item difficulties are different (Table 4.15).

Figure 4.10. "Real" vs. mean of the equated scores, test length=120

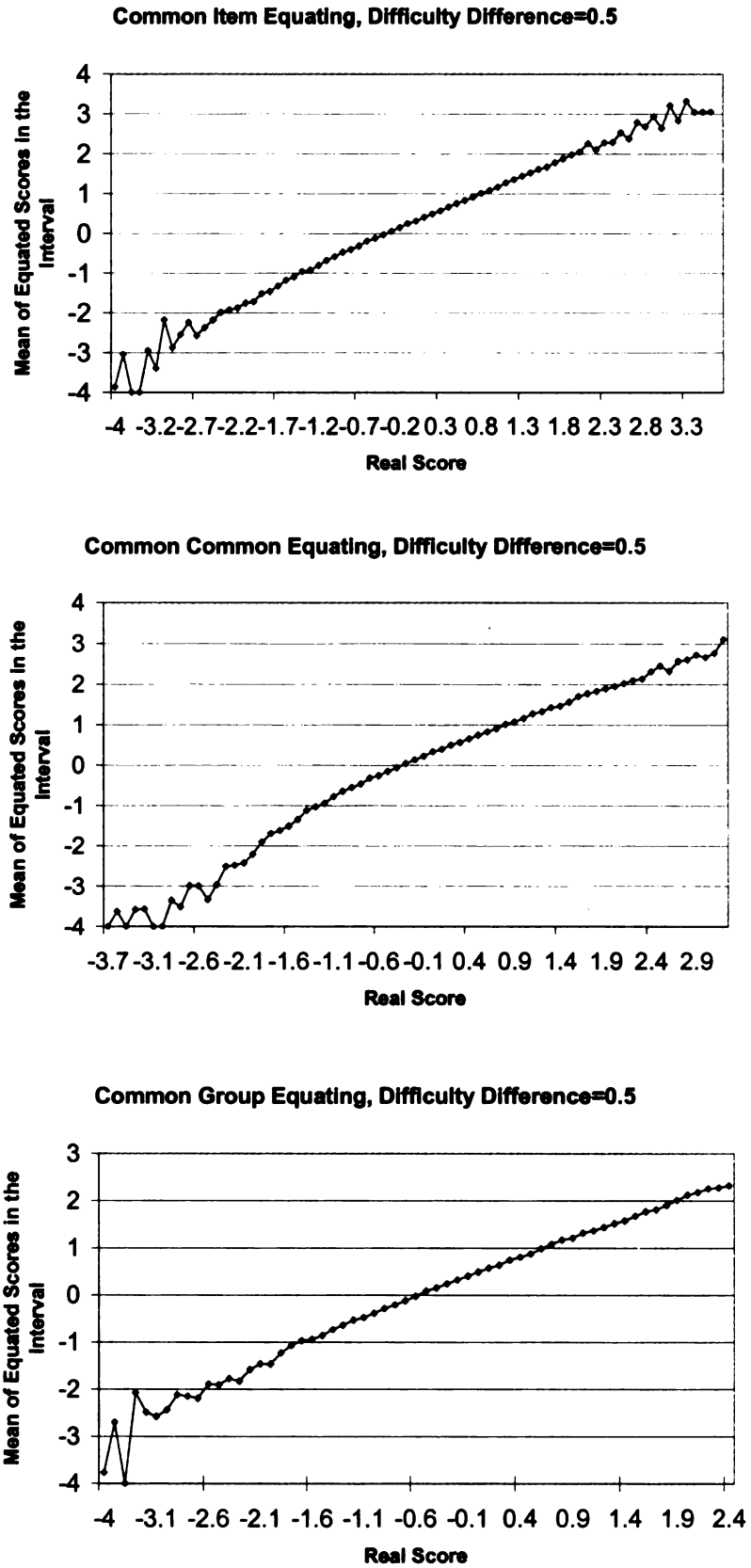
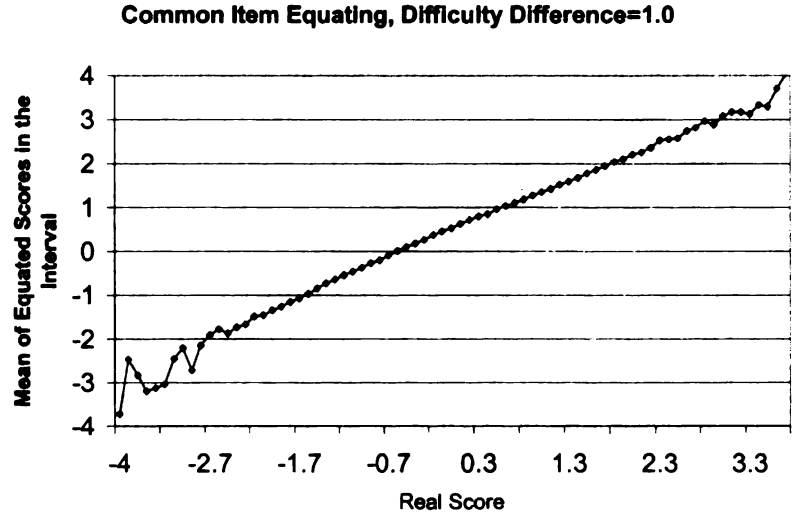
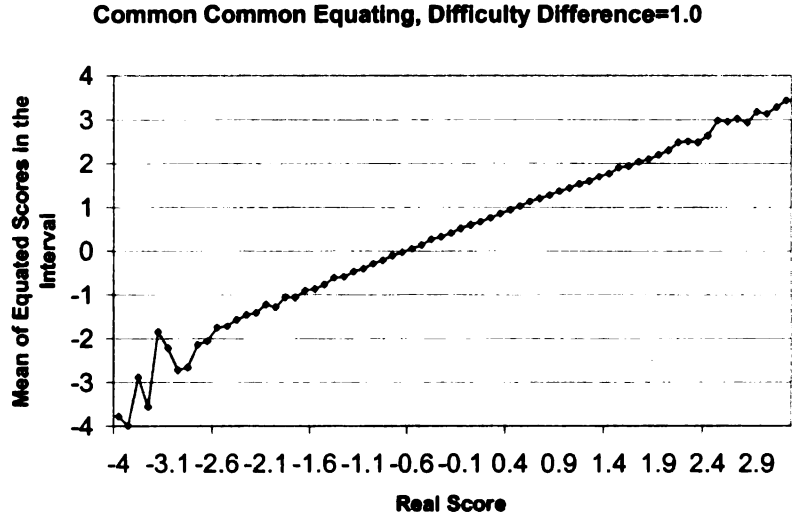


Figure 4.10 (continued)





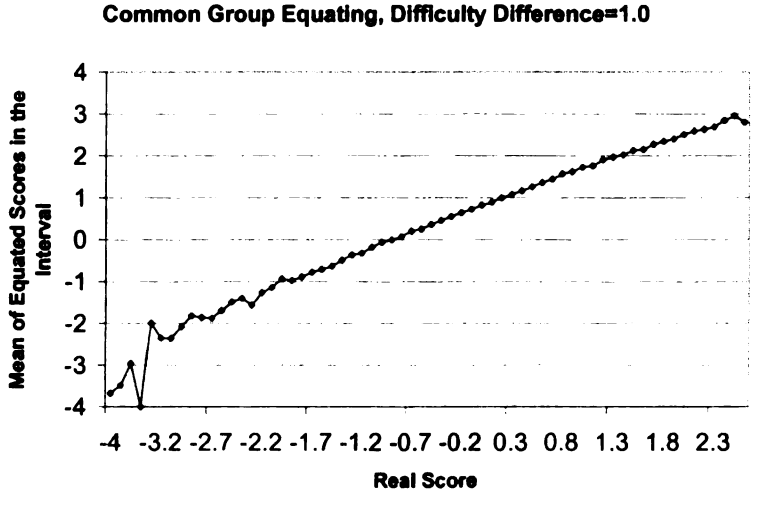
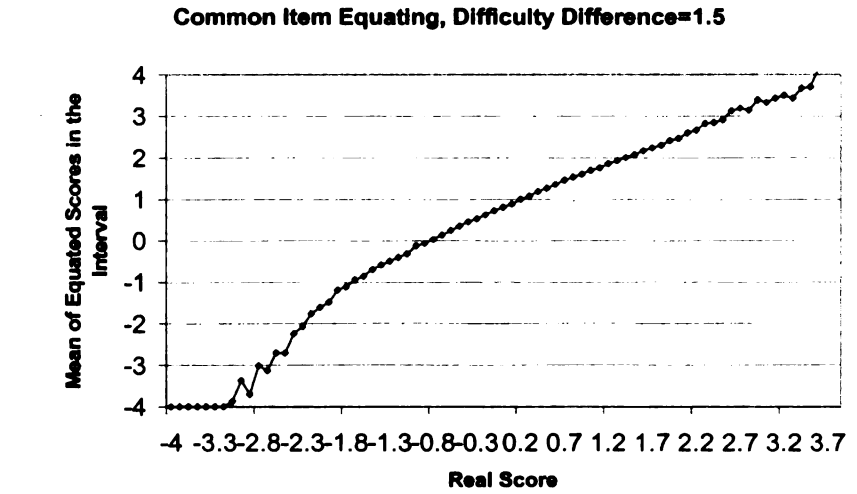
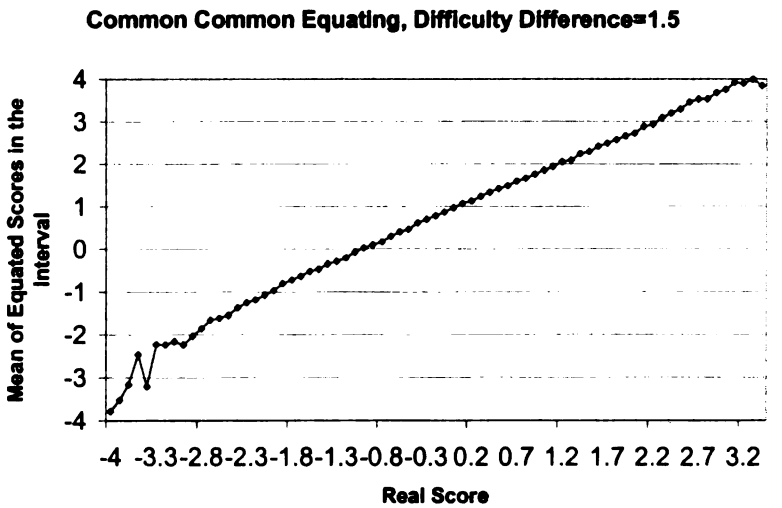


Figure 4.10 (continued)





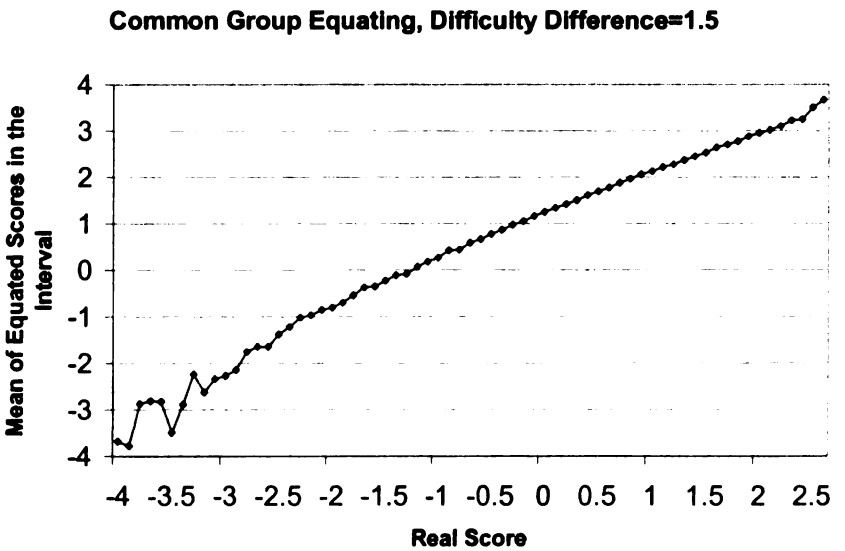
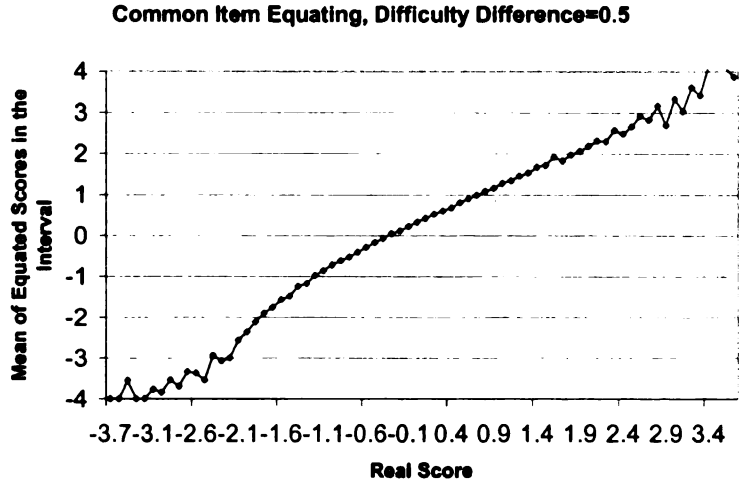
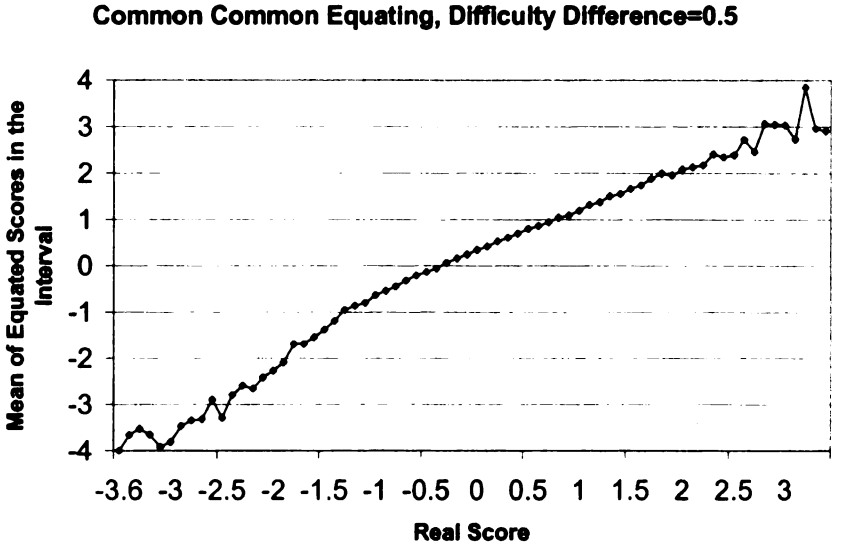


Figure 4.11 "Real" vs. mean of the equated scores, test length=96 items





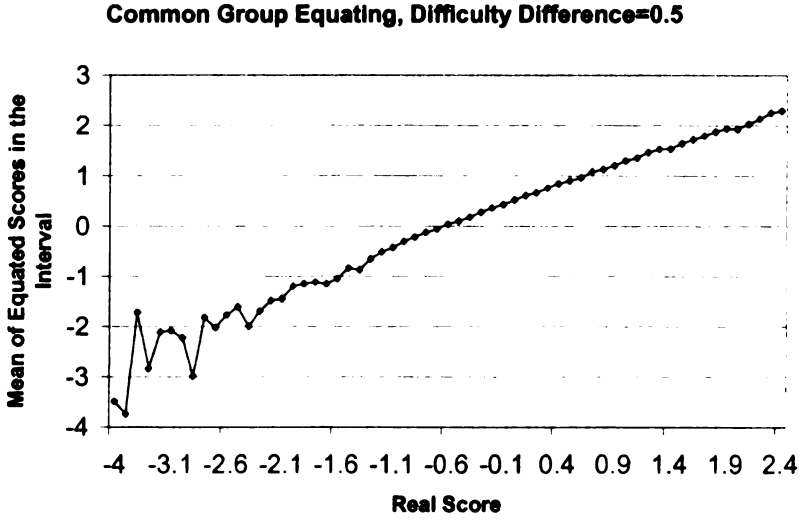
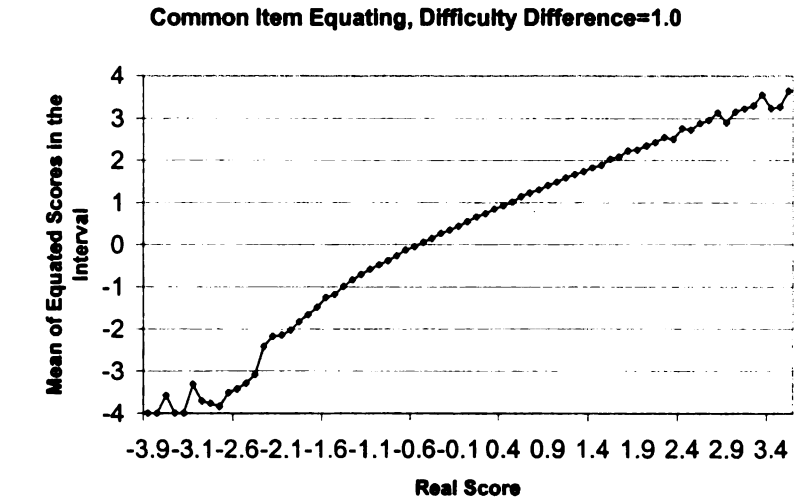
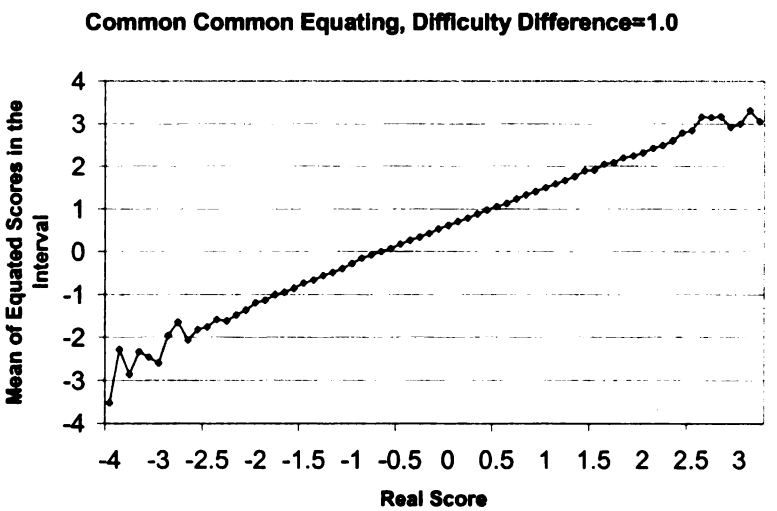


Figure 4.11 (continued)





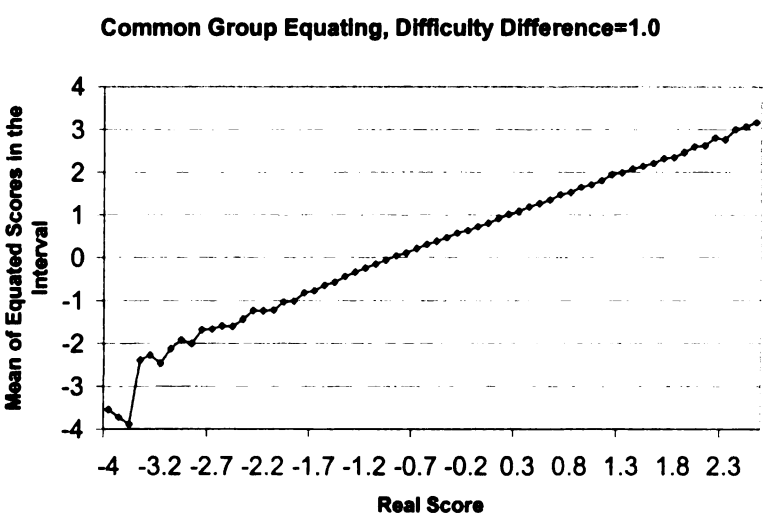
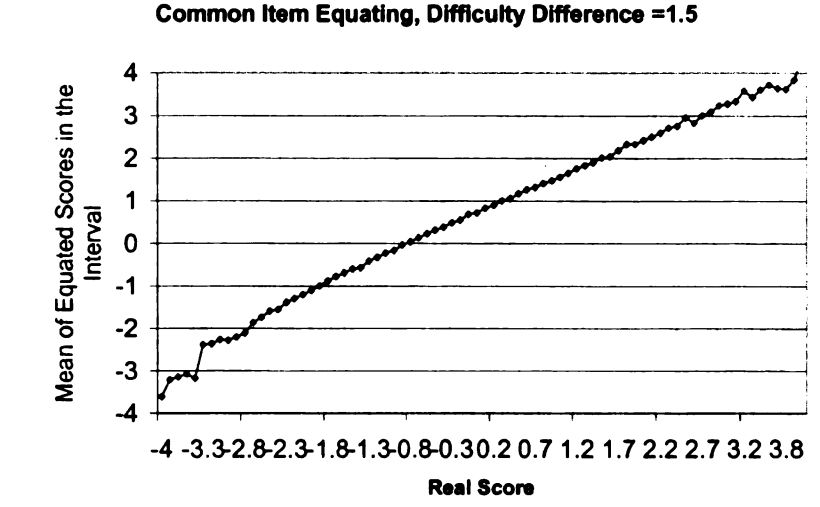
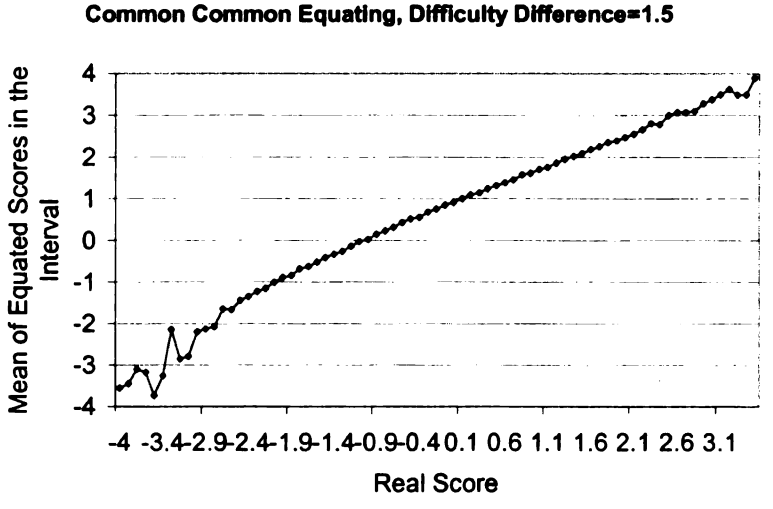


Figure 11 (continued)





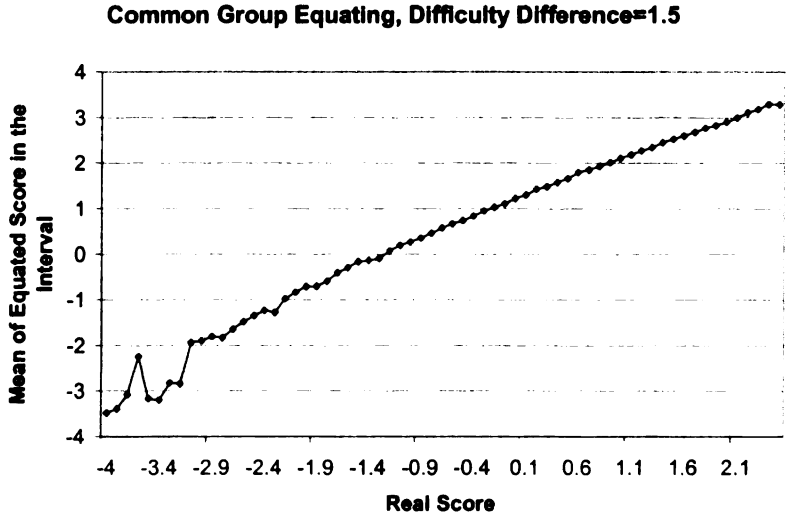
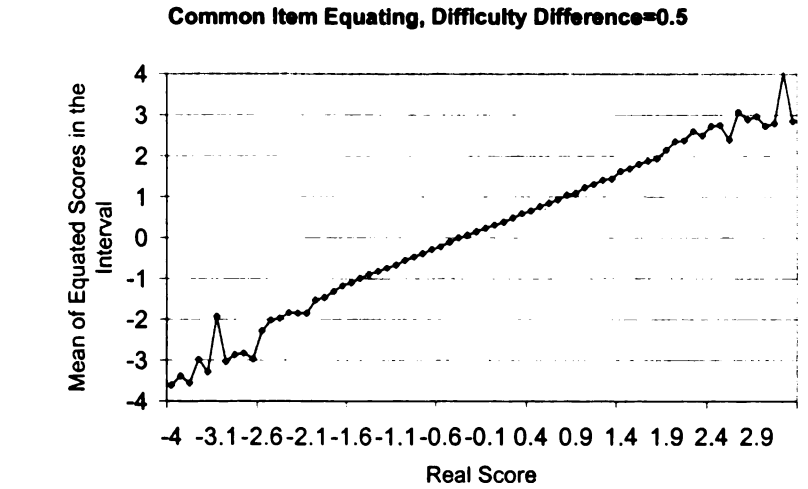
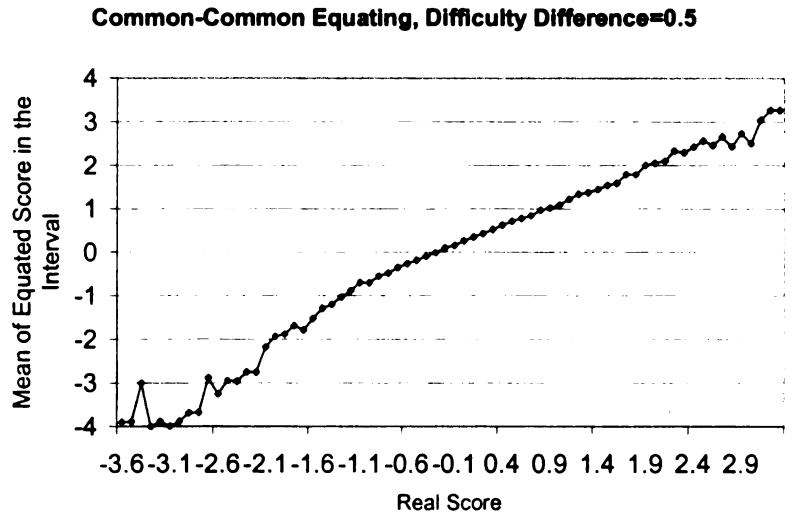


Figure 4.12 "Real" vs. mean of the equated scores, test length=72 items





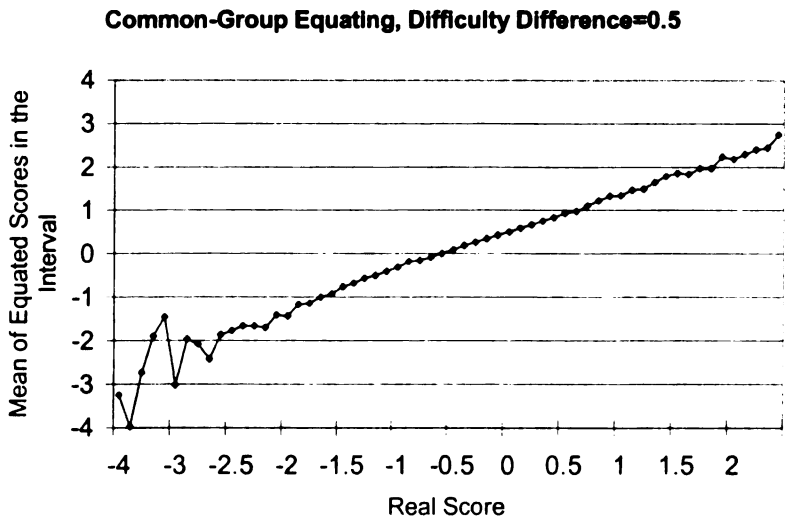
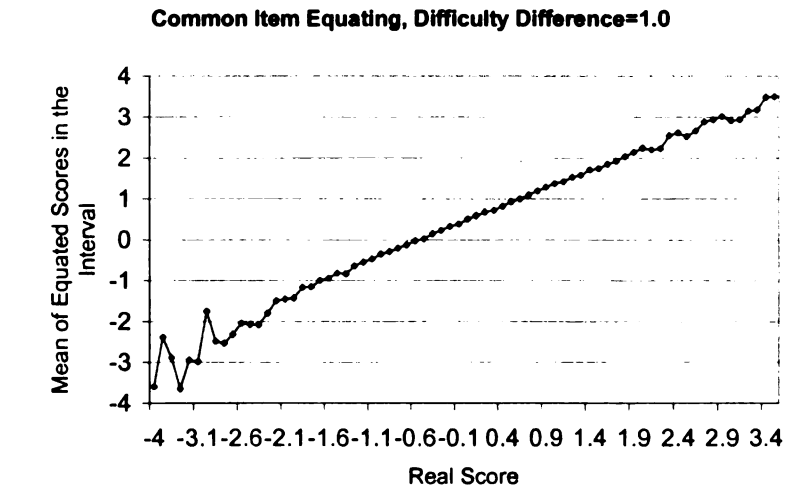
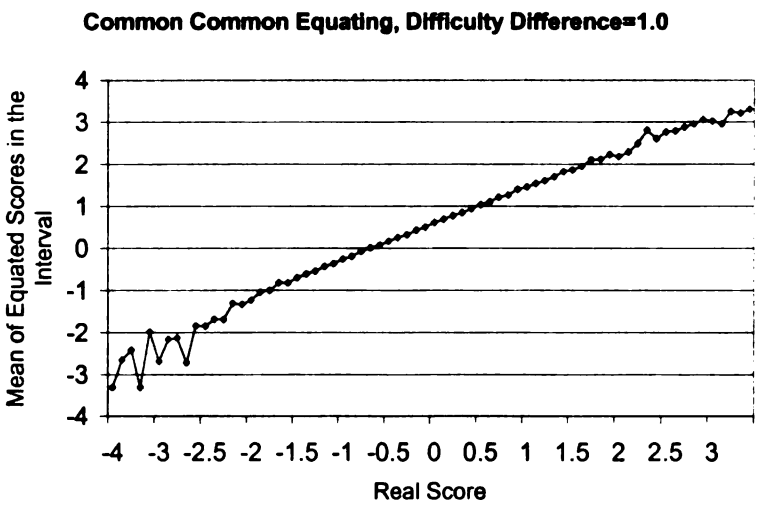


Figure 4.12 (continued)





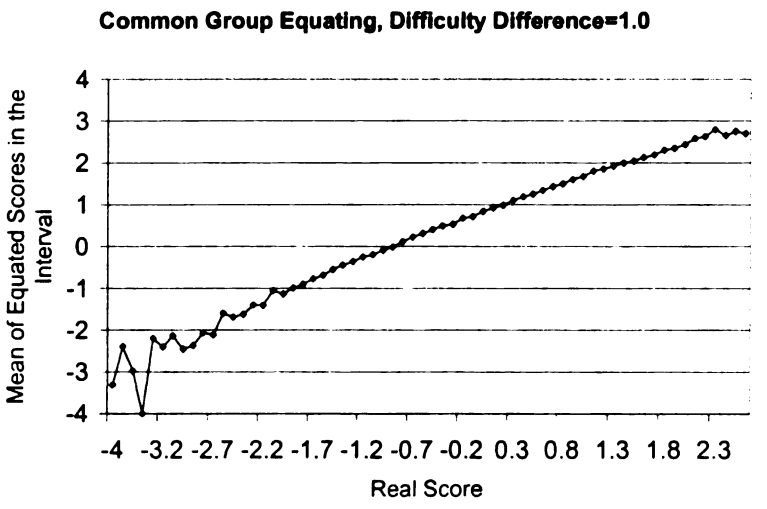
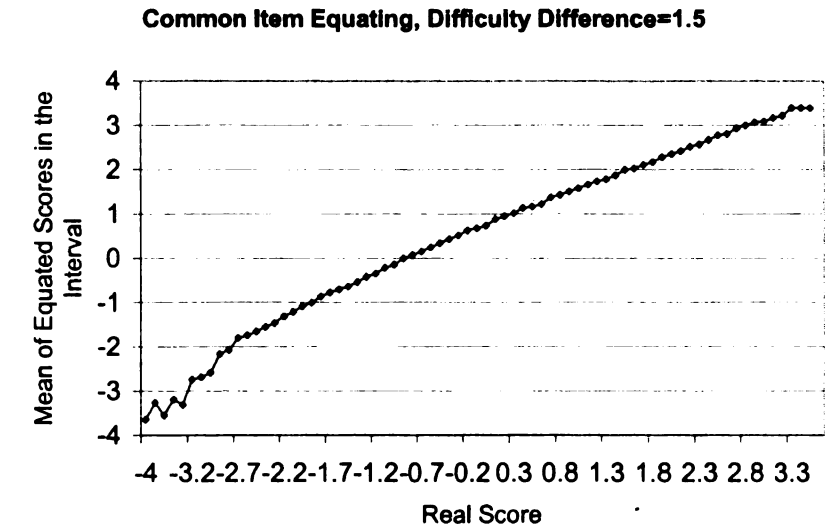
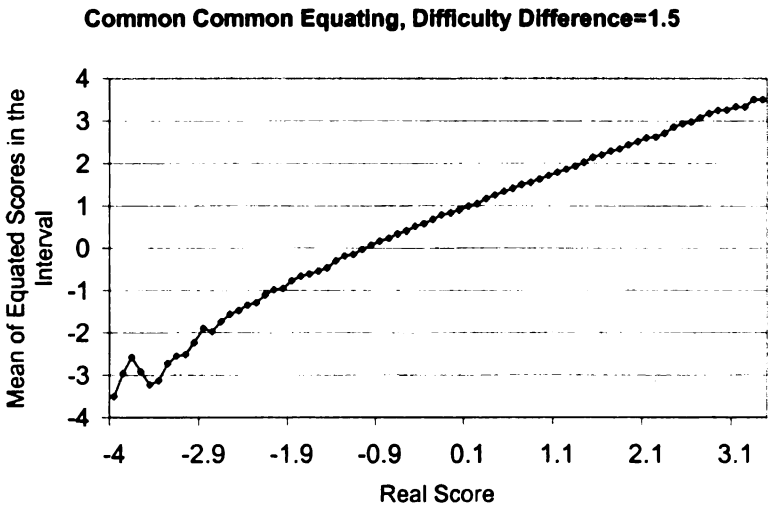
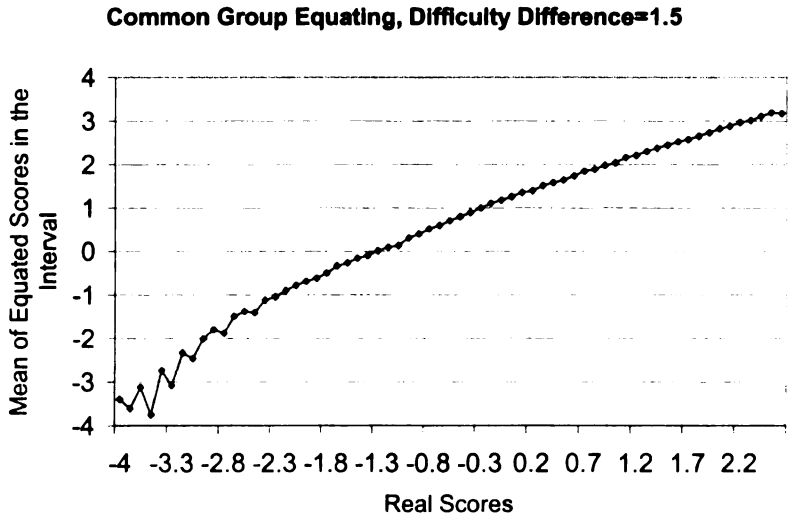


Figure 4.12 (continued)







The correlation coefficients are listed in Table 4.15. In common-group designs, the higher-level group and the lower-level group share more common examinees than those of the common-common and common-item designs; the total number of examinees in each common-group design is 5000. The common-common design has 5500 and the common-item design has 6000.

The square of the difference between "real score" and the adjusted equated score for each examinee was calculated, and the averaged value of the square difference for each equating design is listed in Table 4.16. The adjusted equated score equals the equated score reduced by 0.25, 0.5, or 0.75 for designs when difficulty level differences are 0.5, 1.0 or 1.5 respectively. The reason for using adjusted equated score instead of equated score will be discussed in Chapter 5. The values indicate that on average, the average squared differences are smaller for longer tests; and when the difficulty difference between the two forms increases, the average squared difference increases. However, the differences between different test lengths or different form difficulty levels are not significant. Equating of common-group designs has higher average squared differences than those of the common-common designs. The common-item designs have the lowest average squared differences. The difference in average squared differences between the equating designs is statistically significant.

Table 4.15. Correlation Coefficients between "Real Score" vs Equated Score

Difference	Common -item	Common -common	Common -group	Averaged by test length	Averaged by Difficulty Difference**
120 items				0.976	
0.5	0.977	0.960	0.975		0.966
1.0	0.980	0.980	0.975		0.974
1.5	0.971	0.985	0.980		0.978
96 items				0.970	
0.5	0.964	0.958	0.961		
1.0	0.966	0.972	0.976		
1.5	0.982	0.978	0.975		
72 items				0.972	
0.5	0.972	0.961	0.966		
1.0	0.976	0.974	0.967		
1.5	0.978	0.978	0.974		
Average	0.974	0.971	0.972		

**The averaged values between the levels are significant different ($P_{ANOVA}<0.001$)

Table 4.16 Adjusted Averaged Squared Difference

Difference	Common -item	Common -common	Common -group	Averaged by test length	Averaged by difficulty difference
	120 items			0.108	
0.5	0.058	0.094	0.114		0.104
1.0	0.061	0.068	0.155		0.112
1.5	0.097	0.080	0.246		0.173
	96 items			0.128	
0.5	0.105	0.105	0.158		
1.0	0.106	0.089	0.168		
1.5	0.070	0.098	0.253		
	72 items			0.153	
0.5	0.020	0.117	0.165		
1.0	0.081	0.099	0.181		
1.5	0.099	0.109	0.507		
Average**	0.077	0.096	0.216		

**The averaged values at each level are significantly different ($P_{ANOVA}<0.001$)

IV. Standard Errors of Equating

The testing data for each design was randomly divided into ten parts for standard error calculation according to the description in chapter 3, part III. The plots of standard errors of different equating designs are exhibited in figures 4.13-4.21. Three obvious trends are evident in the results. First, the plots indicate that in designs where the difficulty differences between the two forms are lower (if =0.5), the SE level tends to be lower and the range of lower SE are wider. Second, when test length and the level of difference in difficulty are kept the same, common-item equating tends to have lower SE between ability level -1 and 1; common-group equating tends to have higher SE in that range, although the difference in SE between the two designs are not large. Third, the SE level tend to be lower when the length of the forms is longer.

The averaged SE between ability scores of -1 and +1 are listed in Table 4.17. On

average, shorter test forms tend to have higher standard error of equating. What is more, common-item designs have lower standard error than common-common designs, which is again lower than the common-group design. However, none of the above differences are statistically significant. Higher difference in difficulty between the two forms also contributes to increased standard error of equating, and this trend is statistically significant ($p<0.01$). The trend agrees with what is shown by the average squared differences between the "real" and equated scores (Table 4.16).

Table 4.17. Averaged Standard Error between Scores -1.0 and 1.0

Difference	Common -item	Common -common	Common -group	Averaged by test length	Averaged by difficulty Difference**
	120 item			0.196	
5	0.170	0.170	0.184		0.199
10	0.192	0.204	0.210		0.229
15	0.193	0.211	0.227		0.253
	96 item			0.224	
5	0.183	0.198	0.212		
10	0.205	0.236	0.227		
15	0.232	0.249	0.273		
	72 item			0.261	
5	0.215	0.221	0.237		
10	0.247	0.261	0.278		
15	0.267	0.303	0.320		
Average	0.212	0.228	0.241		

**The averaged values at each level are significantly different $P_{ANOVA}<0.001$

Figure 4.13. Standard error, 120 items, difficulty difference=0.5

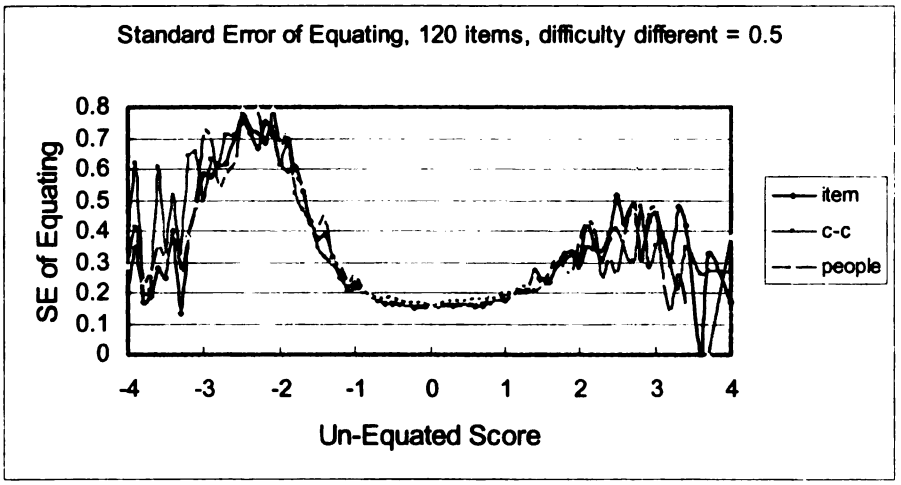


Figure 4.14. Standard error, 120 item, difficulty difference=1.0

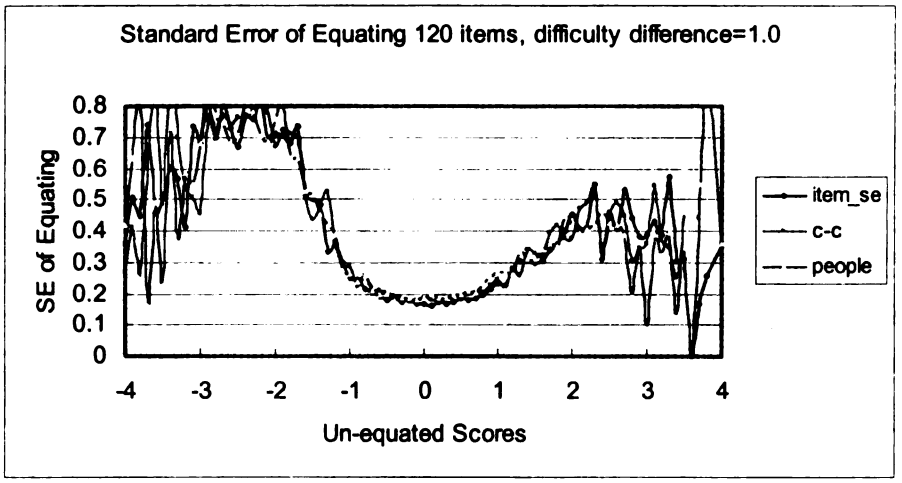


Figure 4.15. Standard error, 120 items, difficulty difference=1.5

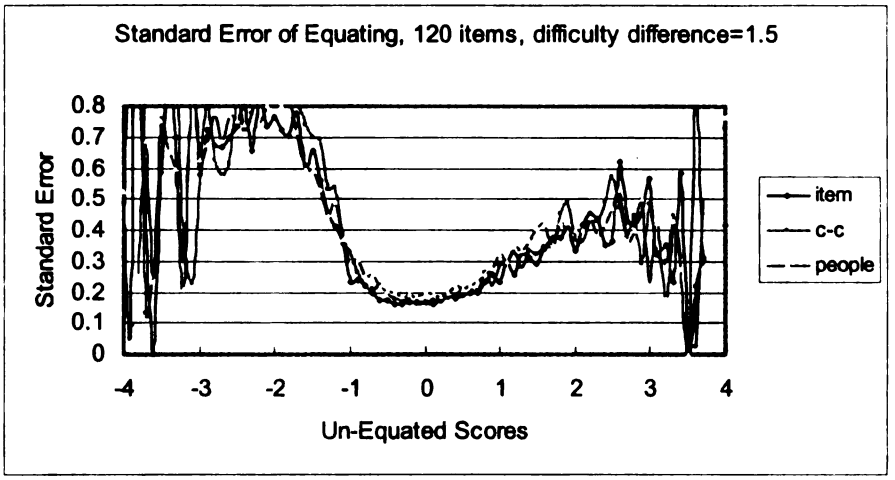


Figure 4.16 Standard error, 96 items, difficulty difference=0.5

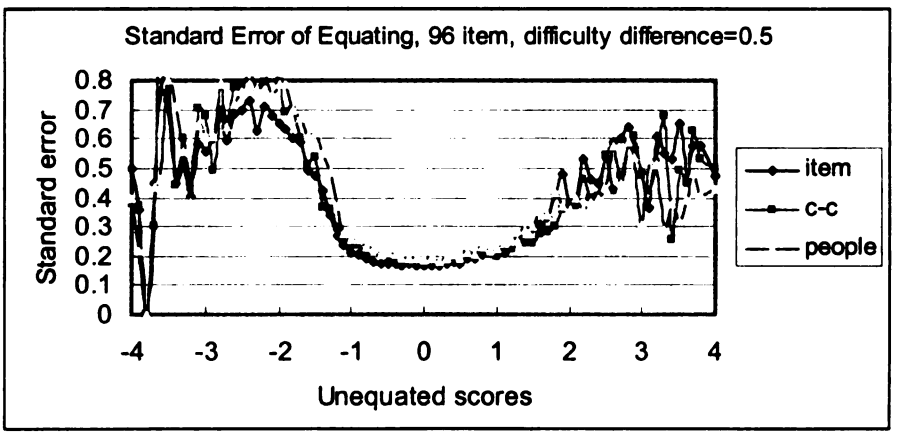


Figure 4.17 Standard error, 96 items, difficulty difference=1.0

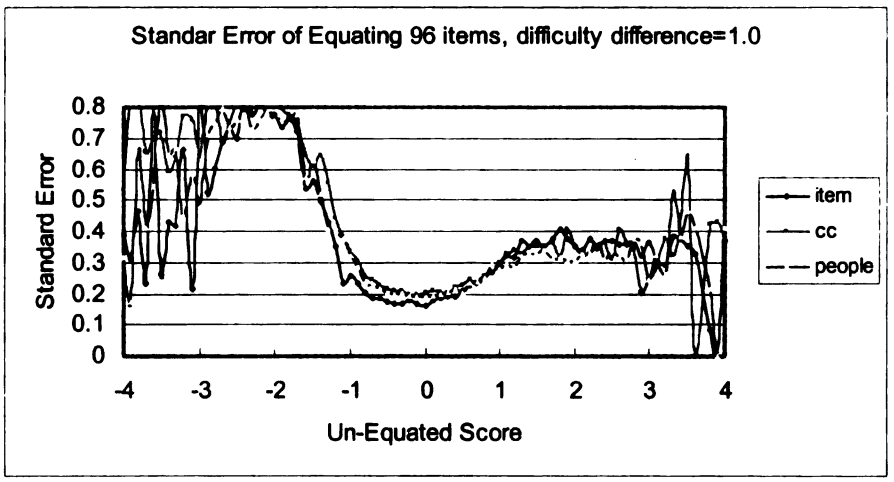


Figure 4.18 Standard error, 96 items, difficulty difference=1.5

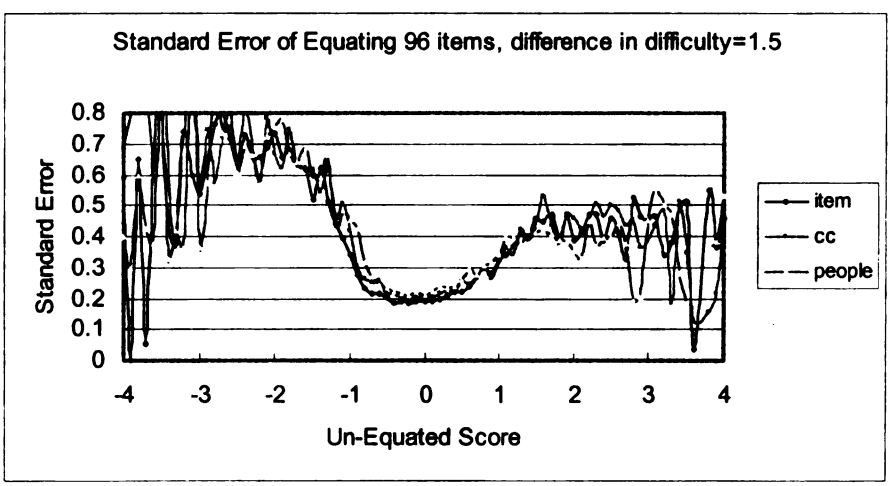


Figure 4.19 Standard error, 72 items, difficulty difference=0.5

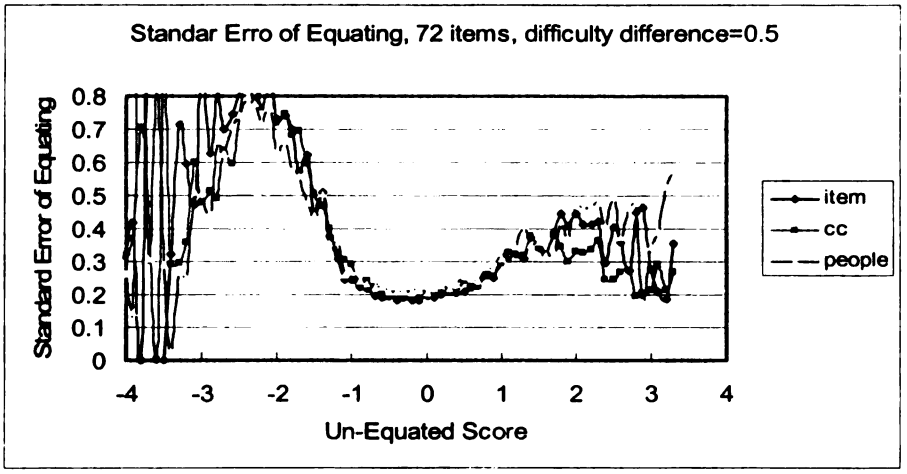


Figure 4.20 Standard error, 72 items, difficulty difference=1.0

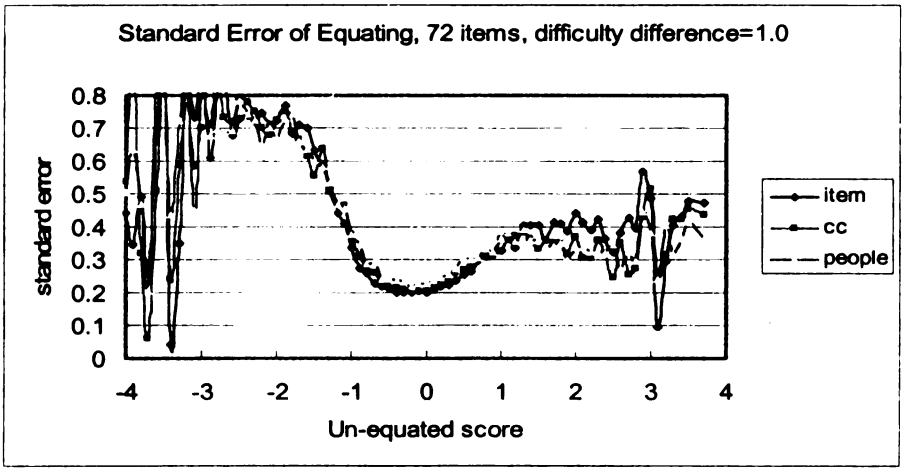
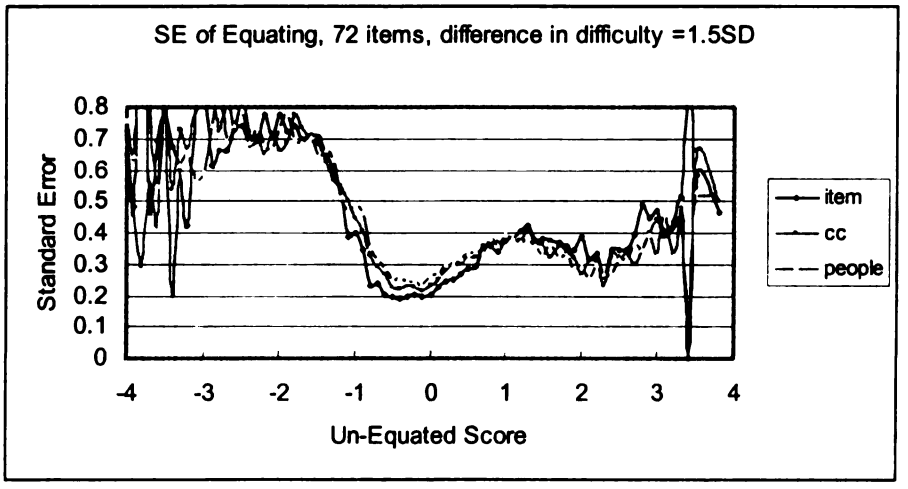


Figure 4.21 Standard error, 72 items, difficulty difference=1.5



Chapter 5. Conclusions and Discussions

I. Dimensionality and Model fit

A. Unidimensional or Multidimensional Structure

One major purpose of this dissertation study is to test the robustness of unidimensional IRT vertical equating when testing data is multidimensional. Thus the first step of data analysis starts from the dimensionality analysis. The direct reason that the data is considered to be multidimensional comes from the structure of the exam. The exam has three sections that are designed to probe different dimensions of language ability of ESL learners—grammar, vocabulary and reading. As IRT is a quantitative model used to interpreting testing data, the dimensionality of the data should be investigated in order to decide whether IRT model is appropriate.

Table 4.2 presents the NOHARM analysis result and 4.3 present dichotomous factor analysis results. The discrimination parameters (the a's) on different dimensions in Table 4.2 mostly agree with the factor loadings in Table 4.3. The summary in Table 4.4 indicates that the multidimensional/factor pattern agrees with the test design (in Grammar, Vocabulary and Reading sections). Although the same test dimensionality is reflected in factor analysis and MIRT, the percentage of items that load the highest disagrees between the two methods.

Factor analysis methods assume that the variables are normally distributed and do not allow guessing in the model, thus MIRT is more suitable for dichotomous testing data where guessing probably exists. Exploratory factor analysis and MIRT were applied in this dissertation, both with oblique rotation on three factors/dimensions, because the test were developed to measure three types of English language ability and previous results indicated the distinctive dimensions between Grammar, Vocabulary and Reading sections.

As already mentioned in chapter 2, the effect of violation in unidimensionality might not be substantial on IRT equating. Among the studies exploring the issue of robustness of IRT unidimensionality assumption, Reckase, Ackerman and Carlson (1988) provides substantial evidence theoretically and empirically that IRT unidimensionality assumption was robust. The study concluded that even though more than one dimension of ability was manifested in examinees' test performance, a set of items measuring the same weighted composite of abilities should be able to meet the assumptions of unidimensional IRT model. The studies by Yen (1984) and Dorans (1990) further support this argument. In this study, the effect of multidimensionality will be analyzed in combination with the effect of vertical equating.

B. IRT Goodness of fit

Although the assumption of unidimensionality is robust for IRT equating, model fit of the data is essential for stable results in parameter calibration and equating. According to van der Linden and Hambleton (1997), well-established statistical tests for two or three parameter IRT models do not exist. And the study further stated that "even if they did, questions about the utility of statistical test in assessing model fit can be raised, especially with large samples." McDonald (1989) even concluded that when sample size is big enough, an IRT model will be rejected by statistical tests.

Currently, model assessment methods that most often used are: judging item fit, judging person fit and compare the fit of different models (Embretson and Reise 2000). The item model-fit tool provided by BILOG-MG3 is a chi-square index. As described in chapters 3 and 4, the test data for this thesis were analyzed with item model fit for all the items at each level of test lengths. According to the p-value of chi-square analysis (Table 4.5), 65 to 75 percent of the items fit the model. Because chi-square analysis is sensitive to sample size, for

an exam that has about 30,000 examinees as here studied, fit analysis based on chi-square can be misleading. The rule of thumb for large-scale exam was then applied so that chi-square values less than 3 times of the degree of freedom is considered as non-significant (M. D. Reckase, personal communications). When this method is applied, 85 to 90 percent of the items fit the IRT model. The results indicate less item-fit for shorter tests—percentages of item-fit are the smallest for the 72 item tests. According to Stone and Zhang (2003), this can be a result of increased Type I error. The results of this study (Zhang, 2003) indicates that Type I error for the traditional item-fit method is big for short tests (less than 40 items), especially for large sample size.

As no well-established, well-recognized method is available in testing two- and three-parameter IRT model fit, checks of model fit from different perspectives are often recommended. This includes checks on the unidimensionality assumption. However, as indicated before, IRT equating can still be valid even when unidimensionality is not fulfilled. Other checks on data include item biserial correlations, test format and difficulty analysis and the test speededness analysis. The biserial correlations of the original test are high (more than 95% of the items have biserial correlations higher than 0.30), however, no data are available to check the test speededness.

Van der Linden and Hambleton (1997) suggest that if the model fit is acceptable, examinee ability estimates ought to be the same from different samples of items within the test. The results that will be discussed later in this chapter show the ability estimates based on a portion of the total items (120, 96 and 72 items). The estimates have high correlations with those based on the original data. The results support the goodness of fit for the IRT model (Figures 4.10-12). On the other hand, van der Linden and Hambleton (1997) also suggest that item parameter estimates ought to be about the same from different samples of examinees from the population of examinees for whom the test is intended. The results in the following

section also supports the model fit from this perspective, the correlations are high between item parameters that were obtained based on different equating designs and different sub-samples. In summary, the chi-square test of item fit, the biserial correlations and the ability estimates based on part of the items, or part of the samples, all support that the data-model fit of this study is satisfactory.

MIRT and factor analysis results provide evidences that the data of this study is multidimensional; however, item model fit analysis and other results indicate that the data fit the IRT model relatively well. Because IRT model requires unidimensionality of the data, when the dimensions are very distinct, the data would not fit the IRT model. Here in this study, the dimensions are correlated between each other with moderately high correlation coefficients (around 0.6-0.7, results not shown). This can be the reason that the items still meet the unidimensionality assumption. If the dimensions were more distinct, the IRT model fit might not be satisfied. When the data does not fit the IRT model, the calibration results obtained through unidimensional IRT would not be stable. In such cases, multidimensional IRT is suggested for item calibration and ability scoring.

II. Correlation between "real" and equated item parameters

Figures 4.1-4.9 present the scatter plots between item parameters calibrated using the original data (130 items by 30,000 examinees) and those calibrated using each equating design. Due to errors in IRT scoring and in equating, the "real" scores and the equated scores are not perfectly correlated, although the correlation is high. What is more, due to scale indeterminacy, we do not expect the regression between the "real" parameters and the equated parameters to cross the origin with slope equals one. The correlations are presented in Table 4.11 and the slope and the origin of each regression are presented in Tables 4.13 and 4.14. In the following sections, the results will be discussed from the perspectives of item parameter indeterminacy, error in parameter estimates and how to evaluate the errors in

parameter estimates.

A. Scale indeterminacy in equating

The BILOG-MG3 3-PL calibration used in this study defines the θ–scale as having a mean of 0 and a standard deviation of 1 for the set of data being analyzed. The IRT parameters are estimated based on this scale. In nonequivalent group equating, when the two groups have samples that are different in the distribution of ability (θ) levels, scale transformation has to be done so that the item parameters and ability levels can be interpreted according to the same scale. The linear relationship between the two scales can be expressed through a set of equations:

$$\theta_{1i}=A*\theta_{2i}+B \qquad (5.1)$$

$$a_{1j}=a_{2j}/A \qquad (5.2)$$

$$b_{1j}=A*b_{2j}+B \qquad (5.3)$$

A and B are called the equating coefficients. θ_{1i} represents the ability level of person "i" estimated by scale 1; θ_{2i} represents the ability level of the same person estimated by scale 2. a_{1j} is the slope parameter of item "j" estimated in scale 1, and a_{2j} is the slope of the same item estimated in scale 2; b_{1j} is the difficulty parameter of item "j" estimated in scale 1, and b_{2j} is the difficulty parameter of this item estimated in scale 2. If the data fits the IRT model perfectly, the same A and B should be applied to all the examinees on all the items. In practice, in common-item equating, A and B are calculated using the averaged value of the of the slope parameters and difficulty parameters across the common items.

According to the equations above, "A" equals the ratio between the two SDs (standard

deviation) of ability distributions. Mathematically, it is calculated as the ratio between averaged "a" parameters of the common items estimated through the data of the two groups. "B" coefficient is related to the difference between the mean of ability distributions, mathematically, it can be calculated through equations:

$$B=b_{1j}-Ab_{2j}=\theta 1_i-A\theta 2_i \qquad (5.4)$$

According to IRT scaling indeterminacy described before, the slopes of both a-parameter and b-parameter regressions (coefficient A's) are related to the ratio between the standard deviation of the two samples ($SD_{\theta 1}/SD_{\theta 2}$) from which the parameters were estimated. Generally speaking, the original data should have bigger variance in ability distribution than any of the equating samples. Among the equating designs, common-group equating has the smallest variance in ability since it has the biggest percentage of overlapped examinees shared by the two groups (20%). Common-common design has 10% of examinees shared by the two groups and common-item design has no overlapped examinee. Each of the examinee group in these designs has a standard deviation of 1. When the difference keeps constant between the means of ability for the two groups, the more common examinees are shared by the groups, the less variance exists in the sample.

The effect of the ability variance in each design is reflected in the slopes shown in Table 4.13: most of the slopes of "a" parameters for common-item equating are smaller (closer to 1) than the correspondent slope for common-group equating, since a slope closer to 1 indicates less difference between the variance of ability in the sample and that in the original data. The slope of the "b" parameter is the reverse to that of the corresponding "a" parameter (according to equations 5.2 and 5.3), again, the slopes of common-item equating b-parameters are most close to "1" among the three designs. The reason lies in that common-item design has the biggest variance in examinee ability levels.

On the other hand, the intercept of the b-parameter regression (coefficient B's) is related to the difference in the means of ability estimates (Equations 5.1 and 5.2). When coefficient A is close to 1, the intercept almost equals the difference between the means of the two groups' ability estimates. The results in the figures for the b-parameters reflect this rule in that for designs with bigger difficulty difference, the regression lines lie further from the origin. Because when the item difficulty difference increase, the ability difference between the two groups increases accordingly to the data sampling. During equating, one of the groups (the lower ability group in this study) was assigned as the reference group and its scale was kept unchanged. The averaged ability estimate of this group is 0.25, 0.5 or 0.75 unit lower than that of the examinees in the original data. This difference is reflected in the intercepts of the b-parameter regression lines.

B. Errors in parameter estimate of IRT equating

Even when coefficients A and B are determined, the relationship between "real" parameter and the equated parameter still cannot be expressed through an equation. The reason lies in that error exists in both IRT parameter estimates using the original data and in the parameter estimate during IRT equating. The error in the parameter estimates can be defined as the amount of variance around the true parameter value. In IRT equating, we look for designs that has smaller errors in parameter estimate.

A variety of factors can cause the error in parameter estimate of IRT equating. The first kind of factors come from IRT calibration process itself. Among these, four factors are often highlighted in the literature. First, because IRT make strong assumptions in modeling item functions, parameter estimate error is incurred when the assumptions are not met (Ackerman, 1992); second, estimation methods such as marginal maximum likelihood estimation (MMLE) or joint maximum likelihood estimation (JMLE) may not convert to the true values. Increased sample size and number of items may affect the accuracy of JMLE and

incorrect prior ability distribution specification may affect the result of MMLE (Baker, 1992; Seong, 1990). Third, model misfit can surely cause unstable item parameter estimation and fourth, practical limitations, such as small sample size or lack of variance in examinees' ability may cause increased error in parameter estimate of too hard or too easy items (Stocking, 1990).

Other than the errors that result from inaccuracies in the estimation of the parameters of the IRT model, the equating process does not perfectly transform item parameters to a common scale. Almost all the aspects of equating design can affect the translation of parameter estimates to a common scale such as the method of equating (single group or common-item, equivalent or non-equivalent group equating), the characteristics of the anchor test (in common-item equating), the characteristics of the two groups, the features of the two forms etc. Evaluating the errors in parameter estimate translation or ability estimation using the translated parameters is very important in evaluating the quality of certain equating design.

C. Evaluating errors in parameter estimation

Table 4.11 presents the correlation coefficients between the "real" and equated parameters: higher correlation coefficients indicate less discrepancy between the "real" and equated parameters. All the "real" parameters used here were obtained through the original data (130 items by 30,000 examinees). The first trend we can see in Table 4.11 is that the correlations of c-parameter estimates are obviously smaller than those of the correspondent a- and b-parameters'. c-parameters are not well estimated when the sample does not have enough low-ability cases. The second conclusion we can draw from the results presented in this table is that tests with fewer items tend to have lower correlations, and this trend is statistically significant for the "a" and "b" parameters. Since test reliability declines as the number of items decreases, error of estimate gets higher when the number of items gets

smaller. No obvious difference is seen between different equating designs or different form difficulty differences.

III. IRT ability estimate

In IRT equating, two sources of error exist in estimating ability. One is from the process of equating, the other comes from the process of IRT ability estimation itself. It is hard to separate the errors from the two resources. This study analyzes the error in ability estimate from two perspectives: first, comparing the ability scores obtained through the original data and those obtained through equating between the samples; second, computing the standard error of equating.

A. Scatter plots of IRT score estimate

In each equating design, the data set has about 5000-6000 examinees, scatter plots between equated scores and "real" scores of each examinee show strong relationship in the middle part of the ability range; while the dots are more scattered and the scores less related at the extreme values of ability (plots not shown). In the plots shown in Figures 4.10-12, "real score" is divided into intervals of 0.1 standard deviation, the corresponding equated scores for each interval are plotted. Compare with the scatter plots of the "real" vs. equated scores, Figures 4.10-12 provides a clearer relationship between the two score. Figures 4.10-12 shows that for ability levels from -2 to 2, a strong linear relationship is demonstrated between the "real" and equated scores. However, the two scores are not linearly correlated at extreme values. The reason is very likely because when an examinee has very high or very low ability, the exam does not have enough items to provide an accurate estimate at the examinee's ability level. Thus the ability estimate of such examinees is not consistent between the results obtained using the original data and those obtained through the equating. It is also noticeable that for the common-item designs, the ranges of "real" scores are usually broader than those of the common-common designs. The ranges are the narrowest for the

common-group designs. The fact that the samples for different designs are differ in their ability score variance has been discussed in the previous session (chapter 5, II. A).

The correlation coefficients summarized in Table 4.15 reflect no trend of the effect by test length or equating design (common-item, common-common or common-group). Longer tests are supposed to have higher reliability, and are thus expected to have higher correlation between the equated and "real" scores. However, in this case, all three lengths may have adequate reliability and the difference may not be large enough to be explicit in the scatter plots and the correlation coefficients. In Table 4.15, the averaged correlation coefficients by difficulty difference show that equated scores between forms that have bigger difficulty difference are more highly correlated with the "real" scores. And the difference is statistically significant. The reason probably lies in the sampling of the equating design for forms with bigger difficulty difference, more examinees with extreme ability levels are included in the sample, thus the ability at extreme levels are more accurately estimated. Another factor that contributes to the higher correlation is the bigger variance of examinees' ability, for we know when two variables are correlated, the bigger the variance of each variable, the higher the correlation coefficient. No obvious difference in scatter plots or correlation coefficient is seen between different equating designs (common-item, common-common or common-group).

B. Square Root of the Average squared difference between the "Real" and Equated Scores

In studies that using generated data to evaluate the quality of equating designs, squared difference between the true parameter and the parameters obtained from equating are calculated as a criteria for the evaluation (Hansen and Beguin, 2002). This study uses real data and the true values of examinees' ability or item parameters are not known, however, the parameters obtained from the original data can be considered as close to their true values. For each design, the average squared difference between the examinees' "real scores" and their adjusted equated scores is calculated. For equating designs with difficulty level differences of

0.5, 1.0 or 1.5, a value of 0.25, 0.5 or 0.75 were deducted from the equated scores to obtain the adjusted equated scores respectively. The reason of the deduction is because the scale indeterminacy introduced in the previous session (Chapter 5, II, A). By equating, a common scale is introduced, in which the lower ability group is assigned as reference group and its averaged ability is arbitrarily set as zero. However, while sampling the data, the averaged score of the lower ability group was set at -0.25, -0.5 and -0.75. Thus the equated scores were adjusted so that they are on the same scale with the ability scores obtained by the original data. Results of the average squared difference of each design are listed in table 4.16.

Unlike the correlation coefficients, the average squared differences show a trend that shorter exams have higher differences between the original and the equated scores, although the trend is not statistically significant. Combined with the plots of the "real" scores vs. the means of equated scores, the discrepancy between the "real" and the equated scores is mostly caused by unstable estimates of the ability with extreme values. It is likely that shorter tests has less items targeting examinees of very high or very low abilities, and thus are less reliable in measuring extreme abilities than longer tests. However, the overall reliability of shorter tests are big enough and thus the overall correlation between the "real" and equated scores show no difference across test lengths.

On average, the common-item design has smaller average squared differences than the common-common design, and the common-common design's average squared difference is lower than that of the common-group design. And the difference here is statistically significant. One of the explanations can be because the common-group design has the smallest variance in examinees' ability. Since fewer examinees score at the two extremities, the common-group equating cannot measure the abilities of these ranges as accurately as the other equating designs. Another possible explanation is that the common-group design itself is not as reliable as the common-item design. As very few studies on common-group

equating are available, no reference here can be quoted regarding the comparison between the two equating designs.

Another obvious trend seen in Table 4.16 indicates that designs for bigger difficulty difference between the test forms have higher average squared difference, especially for tests with fewer items. When the difficulty difference between the two forms is bigger, it is likely that IRT model fit becomes more difficult, and thus the score estimate through equating is less stable and accurate.

C. Standard Error of Equating

The standard error curves in Figure 4.13-4.21 reflect the effects of test length, difficulty difference and equating design from several perspectives. The averaged standard errors of equating between scores of -1.0 and 1.0 are listed in Table 4.17. Several conclusions can be drawn from the analysis of the standard errors. First, the standard error is lower (statistically not significant), and stays low for a wider range, when the test length is longer. When the two forms have a total of 120 items, most of the standard errors between ability level of -1 and +1 are lower than 0.2 of SD; however, when the forms have a total of 72 items, the standard error is almost never lower than 0.2 of SD. The second obvious trend is that when test length and difficulty difference between forms are kept the same, most of the time the standard error of common-item equating is the smallest, common-common is bigger and that of the common-group is the biggest among the three; however, the difference between the standard errors is very small, and sometimes the trend is not clear. Third, keeping test length the same, when the difference in item difficulty gets bigger, the standard error tends to be significantly higher.

The values of averaged standard errors in Table 4.17 agree with the average squared differences listed in Table 4.16. First, as test length increases, the standard error decreases; second, standard error increases with the difficulty difference; and third, common-group

equating has the biggest errors while common-item equating has the smallest errors. Possible explanations of this pattern are provided in the previous session (Chapter 5, III, B), the results of Table 4.14 support the discussions about the results of Table 4.16.

The standard error calculated here is the random error of equating. The standard error curves agree with the scatter plots shown in Figures 4.10 to 4.12: both reflect less error variance in the middle range of ability. However, the subtle difference between equating designs, form difficulty differences and test lengths are reflected through the standard error curves but not through the scatter plots. Part of the reason lies in that the sampling methods of the two analyses are different: in scatter plot analysis, the two groups have different ability levels; in standard error analysis, the two groups have similar ability levels (both were randomly drawn from the total sample).

IV. The effects of equating design, test length and difficulty difference

This dissertation study compares the effects on vertical equating of different equating designs, different test lengths and difference in averaged item difficulty between forms. The results of the analysis are summarized in the following sessions.

A. The effects of equating designs

In the analysis of item parameter estimates and examinee ability estimates, the samples were selected so that the difference between the two groups' abilities matches the difference between the two forms' test difficulty levels. Through the correlation between the "real" ability scores and the equated scores, no obvious difference between the common-item, common-common and common-group designs can be seen. However, the average squared differences between the two scores reflect that equating of the common-group designs may be less accurate than common-item and common-common designs. In the standard error of equating analysis, the samples were randomly selected from the original data and are considered equivalent. In standard error analysis, test forms are different in difficulty level,

while groups are equivalent in ability. The results of standard error agree with that of the average squared difference—common-item equating gives less error than common-group equating, although the difference in error is small and not significant.

B. The effects of test length

In general, longer test means a test has higher reliability, what is more, longer tests usually contains more items targeting different ability levels. Three different test lengths were chosen for this study so that in each equating, the numbers of unique items are 120, 96 and 72. The reliability of all the test forms are 0.85 and higher (results not shown), which satisfies the requirement of most achievement or ability tests in education; however the subtle difference in reliability may still affect the equating results. It is likely that longer tests have more items that accurately measure very high or very low ability, the average squared differences between "real" and equated scores for longer tests are lower than those for the shorter tests. However, the no obvious difference is seen in the correlation between the "real" and equated ability score estimation for different test lengths. For most of the examinees, different test lengths would not affect the effect of their ability estimation. The standard error is lower when test length is longer (not statistically significant), which indicates it is possible that standard error of the vertical equating is sensitive to test reliability.

C. The effects of form difficulty difference

The analysis results indicate that difference between the difficulty levels of two forms does not affect the item parameter estimation; however, in examinee ability estimation, the average squared difference is smaller for equating that has smaller difficulty difference between the forms. On the other hand, bigger difference in difficulty results in higher correlation between the "real" and equated scores; the reason for this may because of bigger ability variance instead of more accurate score estimation. Like the average squared difference, the standard error of equating is higher when item difficulty difference is bigger.

D. Future directions

This dissertation studies equating designs (common-group and common-common) that seldom explored in the practice of educational measurement. Although some of the issues such as fatigue, speededness may arise when using common-group or common-common designs, these designs can still serve as alternatives for the well-practiced common-item design, especially when common items are not available because of security or other issues. In vertical equating, especially when the testing data shows evidence of multidimensionality, common-item equating can be challenging. First, in vertical equating, the common items would be too advanced for examinees at the lower level while too elementary for examinees at the higher level. Higher level examinees may be careless in answering the questions, and lower level examinees may not be able to use their time efficiently (Kolen and Brennan, 2005). In common-item equating, when only a few such items are administered to both groups, the items may not function effectively as an anchor test. However, the common-group or common-common design may overcome this disadvantage. Second, when the test is multidimensional, it is ideal to design common-items that represent all the dimensions of the test; however, this may be impractical for some tests. The results of this study indicate that common-group or common-common designs, although they may not be superior to, are comparable in quality with common-item equating design in vertical equating of multidimensional data.

When common-group or common-common equating is applied in testing practice, it is suggested that examinees take both of the forms are recruited from different ability levels. To minimize the effect of speededness or fatigue, the sequence of the two forms should be arranged so that half of the examinees in the common group take form 1 first and form 2 second; the other half take form 2 first and form 1 second. If sample size allows, testing data obtained from the common-group can be calibrated separately, data of either forms can also

be calibrated separately, and then data from all the examinees can be calibrated using concurrent methods. The results can be compared in terms test structure, examinees' ability estimation etc. Such comparisons provide evidence of the validity for the scoring method. On the other hand, if the test is administered annually as in the case of many achievement exams, data from different administrations can be analyzed to check if the new equating design provides stable scoring over the years. Such analysis using longitudinal data can provide evidence for validity from different perspective.

Because the study on this topic is still preliminary, many directions can be explored. Based on the results obtained from this dissertation, following suggestions are made for future research. First, for the convenience of sampling, the ability levels of the common-group used for this study are centered around the middle range. For example, in common-group design where difficulty difference between the two test forms is 1.0, a sample of 1000 normally distributed cases with mean=0 and SD=0.5 were first selected from the original data that has 30000 cases. These 1000 cases were used as common-group who are administered all the items. Then a sample of 2000 cases were selected from the rest of the data (now has about 29000 cases), so that these cases, together with the 1000 cases selected previously, form a normal distribution of mean=-0.5 and SD=1. Only half of the items are administered to these cases. In the next step, another sample of 2000 cases were selected from the rest of the data (now has about 27000 cases), so that these cases, together with the 1000 cases selected before, is a 3000-case normal distribution of mean=0.5 and SD=1. Again, half of the items are administered to these cases. When common-item equating is applied, usually items from different difficulty levels are selected; thus common-group design may give better equating results when the common-group represents cases from different ability levels.

Second, the analyses here presented are based on data collected from real test. Although

the sample size is very big and normally distributed and the item model-fit is good, the results may not perfectly reflect the effects of the equating designs from pure theoretical perspective. To rule out the impact of some unexpected elements, it is suggested that generated data might be used to explore further about common-common or common-group design.

Appendix 1. MATLAB code (1)

--for selecting a normally distributed group from the original data

```
clear all; clc
%reset seeds for data generation%
rand('state',sum(100*clock));

% set the mean, SD and the number of subjects to be selected and
%the number of bins
u_demand=0
N_demand=1000
SD_demand=0.5
N_bin=40

% Start simulation
%data_30K.dat is the available data with 29935 theta values

data_30K=sort(data_30K);
[N_30K,X_30K] = hist(data_30K,N_bin);
N_30K=N_30K';
X_30K=X_30K';

%generate 1K normal distribution random numbers to define the bins
for i=1:N_demand
    data_3K(i,1) = randn*SD_demand + u_demand;
end

%check the mean and SD of the created data
mean(data_3K)
std(data_1K)

%set the center of each bin for 1K data equals the center of correspondent
%bin for the 30K data
```

```
N_1K=hist(data_1K,X_30K);

%compare the histogram of the original data and the target histogram
figure
hist(data_30K,N_bin)
title('Mother Data Set')
figure
hist(data_3K,N_bin)
title('Son Data Set')

%select the ability scores to fill the bins
for i=1:N_bin
    i
    N_large=N_30K(i)
    N_small=N_1K(i)

    if N_small~=0
      X_large=data_30K((sum(N_30K(1:(i-1)))+1):sum(N_30K(1:i)));

      %see to the attached "N_select_n.m"
      [X_small]=N_select_n(N_large,X_large,N_small);

      data_new_1K((sum(N_1K(1:(i-1)))+1):sum(N_1K(1:i)))=X_small;
    end
end

data_new_1K=data_new_1K';
%display the histogram of the selected cases
figure
hist(data_new_1K,N_bin)
title('Final Data Set')
%check the number, the mean and the SD of the selected cases
N_demand
size(data_new_3K)
mean(data_new_3K)
std(data_new_3K)
%in the resulted data contains 1K cases
save data_new_1K.dat data_new_1K -ascii
```

```
% label the 1K selected data in the original 30K data

data_30K(:,2)=0;

for i=1:1000
    i
    for j=1:size(data_30K,1)
      if data_30K(j,2)==0
        if data_new_1K(i,1)==data_30K(j,1)
            data_30K(j,2)=1;
        end
      end
    end
end

%the result data set contains 29935 cases with the 1000 cases labeled. the
%labled cases were screened out, the rest can be used to select group 1 or
%group 3.
save new_30K.dat data_30K -ascii

% 'N_select_n.m'
% To generate a small data set from a large data set

function [X_small]=N_select_n(N_large,X_large,N_small)

for i=1:N_small
    index(i)=round(rand*N_large+0.5);
% To compare with index selected for the small data set
    for j=1:(i-1)
% If two index are the same, select again until there are no two identical index
        while index(i) == index(j)
            index(i)=round(rand*N_large+0.5);
        end
    end

    X_small(i)=X_large(index(i));
```

End

Appendix 2. MATLAB code (2)

--for selecting 2000 cases for Group 1 in common-group equating

```
clear all; clc
%reset seeds for data generation%
rand('state',sum(100*clock));
%first generate a distribution that has 2000 cases, in which 1000 cases
%N{0, 0.5} are deleted from 3000 cases N{0.5, 1}
% Input
u_demand=0.5
N_bin=40
%generate 3K normal distribution random numbers to define the bins
for i=1:3000
    data_3K(i,1) = randn + u_demand;
end

%generate 1K normal distribution random number mean=0,std=0.5
for i=1:1000
    data_1K(i,1) = randn*0.5;
end
mean(data_3K)
std(data_3K)
mean (data_1K)
std(data_1K)

data_3K=sort(data_3K);
[N_3K,X_3K] = hist(data_3K,N_bin);
N_3K=N_3K';
X_3K=X_3K';

%set the center of each bin for 1K data equals the center of correspondent
%bin for the 3K data

N_1K=hist(data_1K,X_3K);
```

```
%compare the histogram of the original data and the target histogram
figure
hist(data_3K,N_bin)
title('Mother Data Set')
figure
hist(data_1K,N_bin)
title('Son Data Set')

for i=1:N_bin
    i
    N_large=N_3K(i)
    N_small=N_1K(i)

    if N_small~=0
      X_large=data_3K((sum(N_3K(1:(i-1)))+1):sum(N_3K(1:i)));

      %see to the attached "N_select_n.m"
      [X_small]=N_select_n(N_large,X_large,N_small);

      data_new_1K((sum(N_1K(1:(i-1)))+1):sum(N_1K(1:i)))=X_small;
    end
end

data_new_1K=data_new_1K';

figure
hist(data_new_1K,N_bin)
title('Final Data Set')

size(data_new_1K)
mean(data_new_1K)
std(data_new_1K)
save data_new_1K.dat data_new_1K -ascii
%the in the 3000-case group, the 1000 cases were labled
data_3K(:,2)=0;

for i=1:1000
```

```
    i
    for j=1:3000
      if data_3K(j,2)==0
        if data_new_1K(i,1)==data_3K(j,1)
            data_3K(j,2)=1;
        end
      end %data_3K(j,2)==0
    end
end

save data_3K.dat data_3K -ascii

%the 3000-case group were reorganized in Excel file and the 1000 cases were
%deleted from it, ending up with 2000 cases that have the target
%distribution, the data's name is gen_plus_2K.dat. data_29K is the data set
%that has 1000 cases {0, 0.5) deleted from the original 30K data.
clear all; clc

% Input
N_bin=40

% load data
load data_29K.dat
load gen_plus_2K.dat
data_29K=sort(data_29K);
[N_29K,X_29K] = hist(data_29K,N_bin);
N_29K=N_29K';
X_29K=X_29K';

N_2K=hist(gen_plus_2K,X_29K);

figure
hist(data_29K,N_bin)
title('Mother Data Set')
figure
hist(gen_plus_2K,N_bin)
title('Son Data Set')
```

```
for i=1:N_bin
    i
    N_large=N_29K(i)
    N_small=N_2K(i)

    if N_small~=0
      X_large=data_29K((sum(N_29K(1:(i-1)))+1):sum(N_29K(1:i)));

      [X_small]=N_select_n(N_large,X_large,N_small);

      data_new_2K((sum(N_2K(1:(i-1)))+1):sum(N_2K(1:i)))=X_small;
    end
end

data_new_2K=data_new_2K';

figure
hist(data_new_2K,N_bin)
title('Final Data Set')

size(data_new_2K)
mean(data_new_2K)
std(data_new_2K)

save data_new_2K.dat data_new_2K -ascii
```

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement.* 29, 67-91.

Baker, F. B. (1992). *Item response theory: parameter estimation techniques.* New York: Marcel Dekker, Inc.

Beguin, A.A., Hanson, B. A., & Glas, C. A. (2000). Effect of multidimensionality on separate and concurrent estimation in IRT equating. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Bock, R. D, Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika,* 37, 29-51.

Bogan, E.D., & Yen, W. M. (1983). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model.* Monterey, CA: CTB/McGraw-Hill. (ERIC Document Reproduction Service No. ED229450).

Bolt, D. M. (1999). Evaluating the Effects of Multidimensionality on IRT True-Score Equating. *Applied Measurement in Education* 12, 4, 383-407

Camilli, G. Wang, M.M., & Fesq, J. (1992).*The Effects of dimensionality on true score conversion tables for the Law School Admission Test.* LSAC Research Report Series. Newton, PA.

Camilli, G. Wang, M.M., & Fesq, J. (1995). The effect of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement,* 32, 79-96.

Cook, L. L. & Douglass, J. B. (1982). Analysis of fit and vertical equating with the three-parameter model. Paper presented at the Annual Meeting of the American Educational Research Association. New York, NY.

Cook, L. L., & Eignor, D. R. (1991). NCME Instructional Module: IRT Equating Methods. *Educational Measurement: Issues and Practice.* 10, 37-45.

Cook, L.L., Dorans, N.J., Eignor, D.R., & Petersen, N.S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating.* (Research Rep. No. RR-85-30). Princeton, NJ: Educational Testing Service.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, 15(1), 15-31.

Donoghue, J. R., & Hombo, C. M. (2001). The distribution of an item-fit measure for polytomous items. Paper presented at the Annual Meeting of the NCME, Seattle, WA.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, 22, 249-262.

Dorans, N. J. (1990). Equating Methods and Sampling Designs. *Applied Measurement in Education.* 3(1), 3-17.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum, Mahwah, NJ

Fraser, C. (1986). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [computer program], Center for Behavior Studies, The university of New England, Armidale, New South Wales, Australia.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Lease squares item factor analysis. *Multivariate Behavioral Research*, 23, 267-269.

Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology*, 42, 139-167.

Hambleton, R.K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing*, (pp. 31-49). New York: Academic

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

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

Hanson, B. A. & Beguin, A. A. (1999). *Separate versus Concurrent Estimation of IRT Item Parameters in the Common Item Equating Design.* American Coll. Testing Program, Iowa City, IA., ACT-RR-99-8

Hanson, B. A. & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement.* 26, 3-24.

Harris, D. J. (1991). A comparison of Angoff's Design I and Design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement*, 28(3), 221-235.

Jodoin M.G. and Davey, T. (2003). A multidimensional simulation approach to investigate the robustness of IRT common item equating. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Jodoin, M.G., Keller, L. A. & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *Journal of Experimental Education*, 71(3). 229-250

Johnson, J.S., Yamashiro, A.D. and **Yu. J** (2004). The role of cloze in a model of foreign language proficiency. Annual Conference of Language Testing Research Colloquium, Temecula, CA.

Kim, J. P. (2001). *Proximity measures and cluster analysis in multidimensional response theory.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

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

Kolen M. J. & Brennan R. L. (2004). *Test Equating, Scaling, and Linking (2nd edition).* Springer, New York, NY

Kolen, M. J., & Whitney, D.R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational Measurement.* 19, 279-293.

Li, Y. H., Griffith W. D. & Tam H. P. (1997). Equating Multiple Tests via an IRT Linking Design: Utilizing a Single Set of Anchor Items with Fixed Common Item Parameters during the Calibration Process. Paper presented at the Annual Meeting of the Psychometric Society. Knoxville, TN.

Linn, R. L. (1993). Linking Results of Distinct Assessments. *Applied Measurement in Education*, 6(10), 83-102.

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

Matlab [computer program]. Mathwork Inc.

McDonald, R. P. (1989). Future directions for item response theory. *International Journal of Educational Research* 13, 205-220

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied*

Psychological Measurement, 24, (2), 99-114

Mislevy, R. H, (1992). *Linking educational Assessments: Concepts, Issues, Methods, and Practice.* Princeton, NJ: Educational Testing Service Policy Information Center.

Noguchi, H. (1986). An equating method for latent trait scales using common subjects' item response patterns. *Japanese Journal of Educational Psychology*, 34, 315-323 (In Japanese with English abstract)

Noguchi, H. (1990). Marginal maximum likelihood estimation of the equating coefficients for two IRT scales using common subjects' design. *Bulletin of the Faculty of Education, Nagoya University (Educational Psychology)*, 37, 191-198. (In Japanese).

Ogasawara, H. (2001). Marginal maximum likelihood estimation of item response theory (IRT) equating coefficients for the common-examinee design. *Japanese Psychological Research.* 43, 72-82.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.

Pomplun, M., Omar, M. H. & Custer, M (2004). *Educational and Psychological Measurement.* 64, 600-616

Reckase, M. D. (1989). Controlling the Psychometric Snake: Or, How I Learned to Love Multidimensionality. Paper presented at the annual meeting of the American Psychological Association, New Orleans, LA.

Reckase, M. D. (1997). A Linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer-Verlag.

Reckase, M.D. (1998). *Investigating Assessment Instrument Parallelism in a High Dimensional Latent Space.* Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Woodcliff Lake, NJ.

Reckase, M. D., Ackerman, T. A. & Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25, 193-203.

Russell, M. (2000). *Using Expected Growth Size Estimates to Summarize Test Score Changes.* ERIC/AE Digest Series EDO-TM-00-04, University of Maryland, College Park

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameter to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, 14.11-20.

Sykes, R.C., Hou, L., Hanson, B. Wang, Z. (2002). Multidimensionality and the equating of a mixed-format math examination. Paper presented at the annual meeting of the

National Council on Measurement in Education. New Orleans, LA.

Snieckus, A.H., & Camilli, G. (1993). Equated score scale stability in the presence of a two-dimensional test structure. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Stocking, M. L., & Eignor, D. R. (1986). *The impact of different ability distributions on IRT pre-equating* (Research Rep. No. 86-49). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1990). Specifying optimum examinees for item parameter estimations in item response theory. *Psychometrika*, 3, 461-475.

Stone, C. A. (2000). Monte-Carlo based null distribution for an alternative fit statistic. *Journal of Educational Measurement*, 37, 58-75.

Swaminathan, J., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13-30). New York: Academic.

R. Tate (2003). A comparison of selected empirical methods fro assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159-203.

Toyota, H. (1986). An equating method of two latent ability scales by using subjects' estimated scale values and test information. *Japanese Journal of Educational Psychology*, 34, 163-167. (In Japanese with English abstract).

van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer-Verlag.

Wang, M. M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Yamashiro, A. D. and Yu. J. (2005). *The ECCE three –year technical report: 2001-2003.* Ann Arbor, MI: English Language Institute, University of Michigan.

Yamashiro, A. D. and Yu. J. (2005). *The ECPE three –year technical report: 2002-2004.* Ann Arbor, MI: English Language Institute, University of Michigan.

Yen W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied psychological Measurement*, 8, 125-145.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for

unidimensional item response theory. *Psychometrika*, 50(4), 399-410

Yen, W. M. (1986). The choice of scale for educational measurement: an IRT perspective. *Journal of Educational Measurement,* 23(4), 399-325.

Zimowski, M. F., Muraki, E., Mislevy, R.J., &Bock, R. D. (2003). *BILOG-MG* [computer program]. Chicago, IL: Scientific Software International.