1-1-1987

A study of item response theory equating with an anchor test design.

George A. Johanson
University of Massachusetts Amherst

A STUDY OF ITEM RESPONSE THEORY EQUATING

WITH AN ANCHOR TEST DESIGN

A Dissertation Presented

By

GEORGE A. JOHANSON

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

September 1987

Education

A STUDY OF ITEM RESPONSE THEORY EQUATING

WITH AN ANCHOR TEST DESIGN

A Dissertation Presented

By

GEORGE A. JOHANSON

Approved as to style and content by:

Hariharan Swaminathan, Chairperson

Ronald K. Hambleton, Member

Janice A. Gifford, Member

George Urch, Acting Dean
School of Education

For my father,

Arthur B. Johanson

ACKNOWLEDGMENTS

First, I would like to sincerely thank the members of my committee who have each, in their own way, made my efforts successful. My chairperson, Dr. H. Swaminathan, has been exceptionally flexible and patient in working with me under the sometimes inconvenient constraints of time and travel. I am especially grateful to him for freely giving of his time in the form of unscheduled, but much needed, tutorials. Both is classes and less formal teachings have led me to a way of working and thinking that I most appreciate. Dr. R. K. Hambleton seemed to know just the right time for giving a bit of encouragement or a spur-to-action. His classes were decidedly some of the most informative and pleasant that I have had the pleasure of taking. Finally, his efforts and flexibility to work within my occasionally choked schedule must also be acknowledged. The third member of my committee, Dr. J. Gifford, was freely giving of her time and energies as well. She was reassuring in a most timely fashion and, most importantly, I sincerely appreciate her attitude and concern. In addition, Dr. Gifford gave me the much needed, but often unheralded, "where-is-the-switch" beginnings.

Second, I feel I must confess that without the help of B. McDonald this dissertation would perhaps never have come to pass and, most certainly, not on schedule. I am most grateful to her for both the typing and for attending to details that, from a distance, would be most worrisome.

Last, but farthest from least, is my family. My wife, Susan, has perhaps contributed more than anyone towards the completion of this project. She has seemed to thrive upon strange schedules, child-raising, and an occasionally moody spouse. She has been my nourishment and I love her dearly. From my children, Jim and Katie, I have taken the one thing that will be hardest to replace, my time, but I will try.

ABSTRACT

A STUDY OF ITEM RESPONSE THEORY EQUATING

WITH AN ANCHOR TEST DESIGN

September 1987

George A. Johanson, B.S., Trenton State College

Ed.M., Rutgers University, M.S., Rutgers University

Ed.D., University of Massachusetts

Directed by: Professor H. Swaminathan

In the vertical equating of test scores, procedures based on item response theory used with an anchor test design have received wide acceptance. An issue of primary concern, however, is the length of the anchor test needed to provide an accurate equating of scores. While recent work has shown that very short anchor tests may give acceptable results, there is little information available concerning anchor test length. A further concern is the effect that differences in ability distributions have on the equating. Ability distributions may have an impact on both the choice of equating procedure and the length of the anchor test. In this study, the effects of such factors as length of anchor tests, of group ability differences, and equating methods on the accuracy of equating were investigated.

The data for this study were generated using the three-parameter logistic model. Parameters for three populations, each consisting of two groups of examinees, were estimated using the LOGIST program. Four anchor test lengths were studied with each combination of population and equating method. The design included an anchor test which spanned the difficulty range of the combined tests. The anchor tests were nested and the anchor item difficulties were uniformly distributed. The equating procedures studied were concurrent or simultaneous estimation, characteristic curve, mean and sigma, orthogonal least squares, and ordinary least squares.

The results indicated that the characteristic curve equating method was the most accurate of the equating methods studied using a criterion based upon the true item difficulties and the true equating constants. The characteristic curve method was the only method studied to give acceptable results with as few as four anchor test items. With longer anchor tests and smaller mean differences in ability between groups, all of the equating methods studied gave an acceptably accurate equating. When the mean ability differences were very large, the item parameters were poorly estimated and, as a result, the criterion was predictably affected by the increased variation in these parameters. The conclusion was that these parameter estimation errors would make it difficult to accurately equate tests that differ greatly in difficulty if the anchor test used was relatively short and a miniature of the combined tests.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

1.1 Introduction

The first step required in the equating of test scores is the selection of an equating design. The design of this study is the anchor-test design and a major concern of those using this design is the length of the anchor test required for an accurate equating of scores. Second, an equating method must be selected from either the classical or item response frameworks. Any equating method should meet certain conditions if the equating is to be both fair and accurate. The theoretical conditions for test equating are quite severe but test equating is often a necessity and, in many cases, the criteria for an accurate equating are more empirical than theoretical. As mentioned previously, an open question in equating with an anchor-test design is the length of the anchor test. While it is desireable to have as few anchor items as possible, the accuracy of the equating must not be compromised. An additonal factor in test equating is the degree to which the ability levels within the tested groups differ.

Equating scores between groups of differing mean abilities is referred to as vertical, as opposed to horizontal, test equating. The purpose of this study is to investigate the interactions of equating method, anchor test length, and mean ability differences in groups of examinees.

1.2 Equating Designs

There are only three designs that allow for test equating. Note that, in general, "two different tests administered to two different groups of examinees cannot be equated." (Hambleton & Swaminathan, 1985, p. 198). The three designs are (Cook & Eignor, 1983, p. 180; Hambleton & Swaminathan, 1985, p. 198):

1. Single-group design
2. Equivalent (or random) group design
3. Anchor-test design

In the single-group design, the same examinees take both tests to be equated and, thus, the relationship between abilities or scores may be determined without confronting the issue of group ability verses test difficulty. That is, any differences in difficulty level between the tests may be accounted for without adjusting for group ability differences. One difficulty with this design is the problem of finding a group of examinees willing to take several tests or test forms. Another difficulty is the sometimes conflicting effect of both practice and fatigue upon the examinees.

The equivalent-group design attempts to overcome the difficulties of the single-group design by using random samples of examinees.

However, it is very difficult to obtain populations with nearly identical ability distributions. In both designs, conventional or classical methods of test equating yield good results if the difficulty levels of the two tests are somewhat similar (Cook & Eignor, 1983, p. 180).

The third design is perhaps the most popular since it may be used with different (non-random) groups. The anchor-test design requires that a common subset of items (the anchor test) be administered to both groups. Using item responses theory, it is then possible to use the relationship between the common item parameters in the different groups of examinees to find the relationship between both the item parameters for the two tests and the abilities for the two groups of examinees.

1.3 Conditions for an Equating

In all of the following, x (or x_j) will represent an observed score on test X and y (or y_j) an observed score on test Y. Further, $y^*=x(y)$ is a y score transformed to the scale of test X. Lord (1980, p. 199) gives the following three requirements for the equating of test scores.

1. Equity: For every θ, the conditional frequency distribution of $x(y)$ given θ must be the same as the conditional frequency distribution of x.
2. Invariance across groups: $x(y)$ must be the same regardless of the population from which it is derived.

3. Symmetry: The equating must be the same no matter which test is labeled X and which is Y.

A critical observation is that all conventional approaches are group dependent and hence violate the invariance requirement. In addition, the simple regression approach is non-symmetric. However, conventional methods do give reasonable results in horizontal equatings (Harris & Kolen, 1986). In a vertical equating situation, these classical methods are unsatisfactory (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978, p. 499).

The equity requirement can be conceptualized as follows:

> If an equating of tests x and y is to be equitable to each applicant, it must be a matter of indifference to applicants at every given ability level θ whether they are to take test x or test y (Lord, 1980, p. 195).

Certainly, the tests must have equal variance at every ability level or the more capable examinee would choose the test with the smaller variance at his or her ability level. The less able individual would possibly prefer the less accurate measure. Actually, the restrictions imposed by the equity requirement are so severe as to prohibit practical test equating altogether:

> Theorem 13.3.1
>
> Under realistic regularity conditions, scores x and y on two tests cannot be equated unless either (1) both scores are perfectly reliable or (2) the two tests are strictly parallel (in which case $x(y) \equiv y$) (Lord, 1980, p. 198).

In practice, however, fallible tests must frequently be equated. The only reasonable solution seems to be empirical. That is, we must have a good fit between our data and our mathematical model and thus try to minimize the inherent inequities.

1.4 Equating Methods

Test scores may be equated either within a classical or item response frame of reference. In both cases, there are many equating methods possible. For this study, five IRT methods were selected to cover as wide a range as possible from the more common or more promising to the less common or easily dismissed. Among the most common are the simultaneous estimation procedure and the mean and sigma method. One of the most promising is the characteristic curve method. A less common approach to test equating is the method of orthogonal least squares. Perhaps the most easily dismissed method of test equating is ordinary least squares due to its obvious lack of symmetry and, hence, failure to meet the equity requirement.

With real data, a true equating is unknowable. With simulated data, however, the true equating is known and a criterion based upon the true values of the item parameters and the true equating may be developed. Such a criterion was employed in this study to identify the more accurate equating methods.

1.5 Statement of the Problem

Test equating is a procedure that attempts to make scores from different tests comparable. Traditional or classical test theory is not well-suited to equating scores between groups of examinees who differ substantially in their abilities or to equating test scores for examinees on two tests that differ substantially in difficulty. Equating in the above situations is referred to as vertical equating. Procedures based upon item response theory are more suitable for vertical test equating (Hambleton & Swaminathan, 1985). A frequently chosen design for the vertical equating ·f test scores is the anchor-test design. The item response theory model recommended is the three-parameter logistic model (Cook & Eignor, 1983). The problem of equating scores is complicated by the scaling or method of reporting scores. A simplifying assumption is that ability scores are acceptable.

A criterion was developed to determine the accuracy of an equating based upon the true parameter values. This measure is also able to judge the accuracy of the equating that results from a simultaneous estimation procedure.

The minimum length of an anchor test that allows an acceptably accurate equating has been the subject of two recent papers (Wingersky & Lord, 1984; Vale, 1986). Under certain circumstances, it appears that much shorter anchor tests than previoulsy thought may be acceptable. One facet of this study is to attempt to answer the following question:

1. Given a reasonable criterion, what length anchor test is required to produce an acceptably accurate equating of test scores?

Different equating methods will yield different criterion measures. A second aspect of this study is the following:

2. Given a reasonable criterion, which of a selected group of equating procedures results in the most accurate equating of test scores?

A third point of interest is the effect of the ability distributions of the groups of examinees on the equating. If the tests are at a difficulty level suitable for the abilities of the examinees, then as the difference between mean abilities becomes that is larger, an accurate equating may become more difficult to achieve. That is, differing ability distributions may have an adverse effect on the parameter estimates and thus could affect the accuracy of the equating. The third question to be answered is thus:

3. Given a reasonable criterion, how do different mean ability differences affect the accuracy of an equating of test scores?
4. The final concern of this study is the interaction of these three components.

1.6 Purposes

The purposes of this study were to attempt to address the previously stated problems in a very structured, but necessarily limited, fashion. The decision was made to use generated or artificial data in which it would be possible to know the true equating constants. A criterion was developed using these true constants as the basis for all comparisons. Given this criterion, the purposes were to attempt to answer the following questions:

1. Using anchor tests ranging in length from 25 items (standard) to 4 items (very short), which anchor test length will produce equatings that are acceptably accurate?

2. Using five equating techniques ranging from the most popular to those that are seldom used, which methods will result in acceptably accurate equatings?
3. Using three populations each of which contains two groups of examinees that differ in abilities such that the equatings range from vertical to extremely vertical, which populations will permit acceptably accurate equatings?
4. Which combinations of the above factors produce acceptably accurate equatings?

1.7 Significance of the Study

Since test equating with an anchor-test design is rather common, a very practical concern of test developers is the number of items required in the anchor test. While it is true that, in general, longer anchor tests yield a more accurate equating of test scores, for reasons of efficiency and test security, it is advisable to use as few anchor items as possible. In addition, the length of the anchor test may very well be affected by both the choice of equating method and the mean ability differences of the groups being tested.

Another practical concern of test developers and users is the choice of equating method. Certain methods are easily implemented while others are quite complex. The use of different evaluative measures in the research literature makes the choice even more difficult. Clearly, some of the most common and easily used equating procedures may be more or less accurate at some anchor lengths and with some mean ability differences.

A final concern must be the interaction of these components of an equating. If particular combinations of anchor test length, equating method, and mean ability difference prove to be exceptional in either direction, there would be obvious practical implications.

1.8 Organization of the Dissertation

This dissertation contains five chapters and two appendices. The first chapter is an introduction to IRT and a statement of the problem and purposes of the study. Chapter II introduces test equating and reviews the literature on equating. Chapter III contains the methodology and the review of the literature concerning methods of evaluation of an equating. Chapter IV presents the results of the study. The final Chapter, V, contains the conclusions of the study. The first appendix consists of scattergrams of the anchor item difficulties with the equating lines while the second appendix has the computer programs for data generation and the characteristic curve equating procedure.

CHAPTER II

REVIEW OF THE LITERATURE

2.1 Introduction

Since it is frequently necessary to administer several forms of a test, the horizontal equating of test scores is necessary if it is desirable to compare individual scores across test forms. On the other hand, if it is necessary to measure growth in some content domain, then it is necessary to equate test scores vertically across, say, grade levels. Clearly, such situations occur often and, therefore, either horizontal or vertical test equating is required in many testing circumstances. However, we have seen that there are theoretical requirements for an equating that are difficult or sometimes even impossible to meet. In short, test equating is a necessity and there is no theoretically clear path to a solution. To minimize the inequities and inaccuracies, careful attention must be paid to model fit, equating design, and equating method. The first decision to be made concerning the equating method is whether to use a classical or IRT approach.

2.2 Classical Equating

The problem: If we have two tests purporting to measure the same ability and, if these are administered to two different groups of individuals, may we compare or equate their scores?

If the tests are at similar levels of difficulty and the groups have nearly the same ability distributions, then we have a problem of horizontal equating. If both tests and groups are at different levels of difficulty and ability, respectively, then vertical equating is the result.

Classical or conventional equating methods include the following (Angoff, 1971; Hambleton & Swaminathan, 1985).

1. Equipercentile equating, in which scores from two tests are equated when they have the same percentile rank in their respective groups.
2. Linear methods, where a linear equating of scores X and Y by $y=Ax+B$ can be determined from the equations $\sigma_y=A\sigma_x$ and $\mu_y=A\,\mu_x+B$ (Hambleton & Swaminathan, 1985, p. 201).
3. Regression methods, in which either x or y may be predicted from the other by OLS regression or via some external criterion (Lord, 1980).

As mentioned in section 1.2, classical methods perform well in horizontal equating situations but, there is still the group-dependency issue to contend with.

In the classical test theory model, the parameters that characterize an item depend on the group of examinees to whom the test is administered. For example, the proportion of examinees who answer an item correctly, the item difficulty, is clearly group-specific and, as such, not only characterizes the item but, also the interaction between the item and the group of examinees (Hambleton & van der Linden, 1982). Hence, the item statistics would have to be recalculated for a group different than the norming group. In addition, an individual's test score will depend not only on the particular subset of items that he or she is confronted with but, also on his or her group membership. Thus, two examinees who take different tests cannot be compared directly. The classical route around these difficulties is the parallel test and an all-inclusive norming group. Unfortunately, parallel tests are difficult to construct and precision of measurement suffers when an individual takes a test of a difficulty level that is not matched to his or her ability level.

2.3 Item Response Theory Equating

In direct contrast to the group-dependence of the item parameters in classical test theory is the independence of the item parameters over groups in item response theory (IRT). To achieve this group-independence or, more accurately, to make the item parameters independent of the sample of examinees, it is necessary to estimate the item parameter values from the entire population of interest. Large and representative samples are required and estimation procedures are complex. However, once these parameters are determined, it is possible to compare the scores of any two or more individuals on any sub-collection of test items.

At the very heart of IRT is the item characteristic curve or item response function. The independent variable for this function is a single or unidimensional ability or trait measure. The dependent variable is the probability of success on a particular test item. This single-valued item-ability relationship allows the prediction of the probability of a correct response for an individual whose underlying ability in a particular content domain is given. The reverse, which has a more practical consequence, is also true: given the response to an item and the mathematical relationship, we may infer the examinee's latent ability in this content domain.

Currently, there are two functional forms in use for the item characterisitc curve.

The (three-parameter) normal ogive is given by:

$$P_i(\theta_j) = c_i + (1-c_i)\int_{t=-\infty}^{a_i(\theta_j-b_i)} (1/\sqrt{2\pi})e^{-t^2/2}dt \qquad [1.2.1]$$

The (three-parameter) logistic function is given by:

$$P_i(\theta j) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_j-b_i)}} \qquad [1.2.2]$$

In both functions, θ_j is the ability of the jth examinee, j=1,...,N. Ability is usually standardized or scaled to mean zero, standard deviation one. The item parameters are subscripted over items, i=1,...,n. a_i is the discriminating power, it is proportional to the maximum slope of the item response function. The item difficulty, b_i, is the value of θ_j at which a_i is achieved. That is, $P_i(b_i) = k_i a_i$, where $K_i = -1.7(c_i-1)/4$. A typical item characteristic curve is illustrated below. Note that the point of inflection occurs at b_i and that $P_i(b_i)$ is midway between c_i and 1.0. c_i is referred to as the guessing parameter or pseudo-chance level.

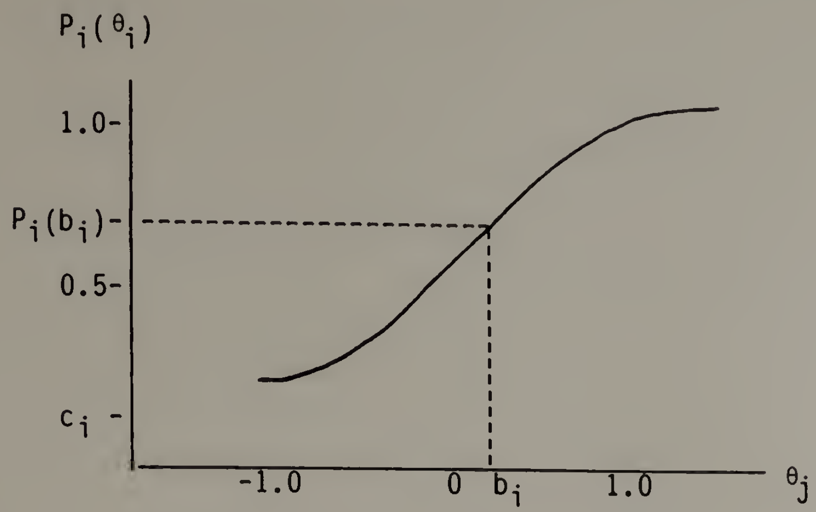


Figure 1.2.1. An item characteristic curve.

For many purposes, the choice of model (normal ogive or logistic) is less than critical since "the two models give very similar results for most practical work" (Lord, 1980, p. 14). The constant -1.7 is chosen to maximize the agreement between the models.

The three-parameter model may be modified by assigning fixed values to item parameters c_i or a_i and c_i. In particular, if c_i=0 the resulting function is referred to as the two-parameter model and assumes that guessing is not a factor. If c_i=0 and a_i=1, the resulting function is the one-parameter or Rasch model. The items are assumed to be of equal discriminating power in the one-parameter model.

The three-parameter logistic model appears to be the most flexible: "the results at present do seem to suggest, however, that the three-parameter logistic model offers a more viable alternative for the vertical equating of approximately unidimensional tests" (Cook & Eignor, 1983, p. 188). For this reason it is the model of choice for this study.

In classical test theory, the test and item parameters or statistics are always group-specific. In addition, examinee scores are test-specific and the accuracy or variability of these scores is assumed to be uniform over scores. Item response theory attempts to overcome these limitations by directly relating an underlying ability to the probability of success on an individual item. If the chosen model fits the data and the ability, θ, is unidimensional, then the item parameters will remain invariant across groups. If this were not the case, we could use these parameter differences to distinguish subgroups and, thus, would be measuring another dimension or ability contrary to our unidimensional assumption. The assumption of unidimensional ability is equivalent to the assumption that the responses of an individual to different test items are independent of one another if the items measure the same ability.

The invariance of an individual's ability measure across tests composed of subcollections of items from a pool of items measuring the same unidimensional ability is one of the key features of IRT. To cite but one example, it allows for <u>tailored testing</u> in which each of two individuals or groups of differing ability is tested at the appropriate difficulty level and, under certain circumstances, the ability scores are comparable. The "certain" circumstances require that the item and ability parameters be on the same scale. Recall that ability was standardized within each group. Putting these scores on the same scale is called test equating and is the subject of this study.

Suppose that two tests are constructed from a unidimensional item pool in which the IRT item parameters are known for all groups of interest. Further, if ability scores are reported, an equating is not even necessary since the exact same ability will result for an individual regardless of the test taken or group membership. The reality, however, is that item parameters are never known exactly in practice and must be estimated from the test data. If the estimates are made separately for each test/group, there is the additional problem that standard procedures arbitrarily set the mean and variance of θ at zero and one, respectively, for each group. When an anchor-test design is used, the result is that θ has been standardized or scaled differently for each group of examinees on the common items. The solution to the equating problem becomes one of finding the relationship between the ability scales on the anchor items across groups and using this same relationship for all items. Recall that in classical linear test equating we assumed that the relationship between observed scores was linear. According to IRT, if the same group of examinees takes both tests X and Y, then the difference between a particular individual's ability scores on the tests will be due solely to the scales of measurement and measurement error.

Therefore, the standardized ability scores will be identical. The relationship is necessarily linear:

$$(\theta_{x_j} - \mu_\theta)/\sigma_{\theta_x} = (\theta_{y_j} - \mu_{\theta_y})/\sigma_{\theta_j} \text{ for each } j=1,\ldots,N \qquad [2.2.1]$$

$$\text{or, } \theta_{x_j} = \theta_{y_j} = \alpha\theta_{x_j} + \beta \text{ for all } j \qquad [2.2.2]$$

(Hambleton & Swaminathan, 1985, p. 204).

Since they are on the same scale, we could have just as well have used the relationship between item difficulties as abilities. In fact, "...item difficulty estimates are typically used because they are the most stable of any of the IRT parameter estimates" (Cook & Eignor, 1983, p. 182). Omitting the subscripts for individuals, we find that if $\theta_y = \alpha\theta_x + \beta$, then $b_y = \alpha b_x + \beta$ and $a_y = a_x/\alpha$ while $c_y = c_x = c$ for each item, i=1,...,n (subscripts omitted). If we use the three-parameter logistic model, it is easy to see that $P_i(\theta_y) = P_i(\theta_x^*) = P_i(\theta_x)$ for all i. Consequently, $\xi_{y_j} = \xi_j^* = \sum P_i(\theta y_j) = \sum P_i(\theta_{x_j}) = \xi_{x_j}$ for all j where the sums are taken over the anchor items. More simply, the true scores on the common items will be identical.

Simultaneous Estimation of Parameters

Using the LOGIST program (Wood, Wingersky, & Lord, 1976), it is possible to simultaneously estimate all item and ability parameters by simply coding the unique items on each test as "not reached" by the examinees who took the other test. The coding will be discussed more fully in a later section, but the result is that all parameter estimates for both groups are automatically on the same scale. This is clearly a very attractive procedure if it is reasonable to apply.

Separate Estimations of Parameters

If the item and ability parameters are estimated with two separate LOGIST runs, then it is necessary to find the relationship between these sets of parameters. It was previously shown that the desired function is linear or, equivalently, that the only difference between the ability or difficulty estimates is the metric or scale of measurement and choice of origin. These will differ since the groups are different and LOGIST assigns $\mu_\theta = \theta$, $\sigma_\theta = 1$ within groups. If we consider the two ability estimates for each person from the anchor-test items, the plot should be a perfectly straight line, $\theta_y = \alpha\theta_{x_j} + \beta$ for all j. Of course, the usual errors of measurement will instead give us a scatter about a line. Our task is to estimate the best fitting line. Ordinary least squares (OLS) regression is not suitable since, as previously mentioned, it is not symmetric and hence would violate the equity requirement of an equating. An orthogonal least squares approach, which involves determining the major or principal axis (Ironson, 1983), while symmetric, "...is not suitably invariant under a change of scale" (Stocking & Lord, 1983, p. 202), since the eigenvalues and eigenvectors of a matrix are not invariant under linear transformations. For example, if the θ_x values are all halved, the resulting (or new) α should be twice the original α and this is not necessarily the case with orthogonal least squares.

Another approach to finding the best fitting line is the mean and sigma method. Since $b_{y_j} = \alpha b_{x_j} + \beta$, it follows that $\bar{b}_y = \alpha\bar{b}_x + \beta$ and $s_{b_j} = \alpha s_{b_j}$. Therefore, $\alpha = (s_b / s_b)$ and $\beta = \bar{b}_y - \alpha\bar{b}_x$. This method is symmetric, but "poorly estimated item difficulties may have a serious impact on the computation of sample moments..." (Stocking & Lord, 1983, p. 203).

More robust procedures have been developed (Stocking & Lord, 1983) to compensate for the effect of outliers and the varying standard errors of the estimates of the item difficulties. However, "a drawback to all of these 'mean and sigma' transformation procedures is that they are typically applied only to the estimated item difficulties" (Stocking & Lord, 1983, p. 203). That is, not all of the available information is being used.

The above approaches determine the line of best fit using only the item difficulty parameters. A group of procedures that attempts to use more than just the difficulty estimates is the characteristic curve methods. Since $P_i(\theta_{y_j}) = P_i(\theta_{x_j})$ for all i and each j, we may compare the item response functions and compute parameter estimates that minimize some aspect of their difference (Haebara, 1980; Divigi, 1980). Stocking and Lord (1983) propose that the mean of the squared differences in estimated and equated true scores over examinees be minimized. They compared this method with their robust mean and sigma method and concluded that "the robust mean and sigma method never provided a better fit to the estimated item difficulties and discriminations; in some cases it provided a worse fit" (Stocking & Lord, 1983, p. 206). Further, they claim that the characteristic curve method is "logically superior" to the robust mean and sigma method in that it makes use of all of the available information in the form of the item response function.

There may be situations where true scores must be equated. That is, instances in which it is inappropriate to report on the ability scale. Unfortunately, "the graph of ξ_x against ξ_y will be non-linear" (Hambleton & Swaminathan, 1985, p. 213). To retain the advantage of a linear relationship, Hambleton and Swaminathan recommend equating abilities and then graphically determining the corresponding, but non-linear, relationship between true scores using a plot of ability verses true scores. An alternative procedure is to use raw scores to equate the tests.

Since the expected value of an observed or raw score, r, is a true score, it may seem reasonable to use the true score procedure described above to equate raw scores. But, recall that $\xi_x = \sum P_i(\theta_x) = \sum (c_i + \ldots)$, or ξ_x is bounded below by $\sum c_i$ while corresponding raw scores are bounded below only by zero. Raw scores and true scores are not simply interchangeable. Be that as it may, "...most IRT users presently equate their tests using estimated true scores and then proceed to use their equated scores table with observed test scores (Hambleton & Swaminathan, 1985, p. 218). A more appropriate procedure is to generate the theoretical observed score distributions and from these obtain the marginal observed score distributions. These are then equated using an equipercentile procedure. This approach to the equating problem seems to yield results very similar to the true-score procedure (Lord & Wingersky, 1983).

CHAPTER III

THE METHOD OF THE STUDY

3.1 Introduction

The objective of this study was to investigate the results of vertical IRT equatings using an anchor test design, generated data, and subject to the following conditions:

1. Anchor Size: The lengths of the anchor tests will be 25, 13, 7, and 4 items. The individual tests will each have 60 items.
2. Group Ability Distributions: Each group to be equated will consist of 500 examinees with normally distributed abilities. Three populations, of two groups each, with ability overlaps of 10%, 30%, and 50% will be equated.
3. Equating Methods: The methods selected were a concurrent LOGIST, characteristic curve, mean and sigma, orthogonal least squares, and ordinary least squares.

Since artificial data permits the true equating constants to be known, it was possible to develop a criterion for comparisons based upon these true values.

3.2 Data Generation

While there are many criteria available for evaluating an equating (see Section 3.4), the only certain way to judge the accuracy of a particular equating is to know the true equating and this information is only available when data are generated. Monte Carlo studies also offer such benefits as perfect fit to the mathematical model and content independence. When real data are used, these factors become confounding issues in determining the accuracy of an equating. Precisely defined and relatively narrow questions would seem to lend themselves to constructed data sets because some confounding issues may then be contained. Of course, results from Monte Carlo studies cannot be casually extended to real data sets.

A data generation program was written in PASCAL using the three-parameter logistic model. The probability of success of person j on item i, $P_i(\theta_j)$, was calculated for each combination of ability and item. A random number between zero and one was then generated (RANDOM, a pseudo-random number generator used in PASCAL 6000, University of Minnesota, 1978) for each such combination. Whenever $P_i(\theta_j)$ was greater than or equal the corresponding random number the item was said to have been answered correctly by that person. If $P_i(\theta_j)$ was less than the random number, the item was coded as incorrect. In this way, dichotomous data was created for each group on the appropriate test. Each group had 500 examinees and each test had 60 items exclusive of the anchor items.

In all, three data sets were created each with 85,000 dichotomous responses (2 testsX500 examineesX(60+25) items). These sizes represent a compromise between accuracy and practicality. The data were generated under the following assumptions or conditions:

1. The abilities, θ_j, were normally distributed within each group with mean-ability differences of 3.30, 2.08, and 1.34 for each of the three sets of data. Standard deviations were all 1.0.
2. The mean item difficulty for each of the six tests was set at the corresponding group mean ability. All difficulties were uniformly distributed with a span of 1.5 units.
3. The mean item discriminations ranged from 0.8 to 1.0 and had spans from 0.8 to 1.2. The test assignments were random and the distributions peaked in the sense that the less discriminating items were those with the more extreme difficulties.
4. The mean pseudo-chance level for each item was set at 0.2 for all tests. The distribution was uniform with range 0.15 to 0.25 for all tests.
5. Anchor items were duplicates of selected items on particular tests. Anchor lengths of 25, 13, 7, and 4 were used.

Each of the three data sets consisted of two groups of examinees and two anchored tests. The group of lesser ability is referred to as group A, the more able group is B. The corresponding tests are X and Y. The populations or abilities were normally distributed. However, within each data set, the combined ability distribution is bimodal due to the rather large mean ability differences. These differences of 3.30, 2.08, and 1.34 resulted in populations with overlapping abilities. The percentages of overlap were 10, 30, and 50, respectively. These ability differences were sufficiently large to enable all equatings to be considered genuinely vertical. Mean item difficulty was set at the group mean ability to make each test most suited to the abilities of the population being tested. Originally, a span of difficulties larger than 1.5 units was employed, but due to the large mean differences in ability, it became very difficult to generate data in which the easiest and most difficult anchor items had realistic parameter estimates. While a larger span might be more usual (Hambleton & Swaminathan, 1985, p. 36), it was not possible with these large mean ability differences. A uniform distribution of difficulties seemed reasonable and is common in the literature, for example, Vale (1986), Skaggs & Lissitz (1986), or Hambleton & Rovinelli (1986). It is equally common to have the discrimination distribution uniform. However, in an effort to construct a good test, it seemed justifable to slightly favor the items with difficulties near the mean ability by assigning to them a better or larger discrimination. The peaked discrimination distribution does precisely this. Discrimination means and spans were consistent with the current literature. The pseudo-chance parameter values were randomly assigned.

Petersen, Marco, and Stewart (1982, p. 134) concluded from their study that "An anchor test constructed to be a miniature of the total tests gives the best equating results." Table 3.2.1 shows the selection rule for the anchor items for each of the three data sets.

Table 3.2.1. The Composition of the Anchor Tests

Anchor Item Number	Test Item Number	Identical to Item Number/Test
1	61	1/X
2	62	41/X
3	63	21/Y
4	64	60/Y
5	65	21/X
6	66	1/Y
7	67	41/Y
8	68	11/X
9	69	31/X
10	70	51/X
11	71	11/Y
12	72	31/Y
13	73	51/Y
14	74	6/X
15	75	16/X
16	76	26/X
17	77	36/X
18	78	46/X
19	79	56/X
20	80	6/Y
21	81	16/Y
22	82	26/Y
23	83	36/Y
24	84	46/Y
25	85	56/Y

The four length anchor consists of anchor items one through four, the seven length anchor of items one through seven, the thirteen length anchor of items one through thirteen, and the twenty-five length anchor of items one through twenty-five. Thus, the four anchor tests are nested. Since the twenty-five anchor items were arranged in order of difficulty, the shorter anchors could be obtained by deleting every other item starting with the second item at each stage.

Within each of the six tests, the items are in increasing order of difficulty. Therefore, within each of the three data sets, the first item on test X was the easiest of the combined 120 items and the last item on test Y was the most difficult. Each anchor test contains both of these items and thus spans the difficulty range of the combined tests for each data set. Skaggs and Lissitz (1986) used an anchor in which the difficulties only spanned the overlap in difficulties of the two tests being equated. However, they concluded that "better results might have been achieved with a wider range of difficulty on the anchor test items" (p. 315). The remaining anchor items in each anchor test were chosen in such a way that the item difficulties within each anchor test were nearly uniformly distributed. Each anchor test was thus constructed to resemble the combined tests as closely as possible.

For reasons of time and economy as well as security, it is frequently desirable to use as small an anchor test as possible. The 'rule of thumb' is the larger of twenty items or twenty percent of the total number of test items (Budescu, 1985, p. 15). Using this rule, all but the longest of the anchor tests in this study are too short.

However, more recent studies by Wingersky and Lord (1984) and Vale (1986) suggest that anchor tests of as few as two good items may permit adequate linking of test scores. Tables 3.2.2 and 3.2.3 show the minimum, maximum, and mean values for both the item and ability parameters.

As a partial verification of the accuracy of the data generation program, checks were run on the ability distribution. Means, standard deviations and normalcy were as desired. The means and standard deviations of the raw and true scores were calculated to verify model fit. These results are summarized in table 3.2.4. Raw scores and true scores within each group were, as desired, nearly identical.

For each combination of anchor length (25, 13, 7, 4) and data set or ability overlap (10%, 30%, 50%), item and ability parameters had to be estimated from the dichotomous data. These estimations were carried out for group A on test X and group B on test Y. In addition, the combined group of examinees, AB, in each data set, was treated as if they had taken all of the items from both tests A and B plus the anchor items. This new, combined test, XY, is discussed more completely in the next section. All parameter estimations were done using LOGIST (Wood, Wingersky, and Lord, 1976). A total of 36 LOGIST runs were required (4X3X3) to estimate all of the combinations of anchor length, population, and group. The maximum number of stages for convergence was set at 40 and the other options were set to the default values. In both groups A and B, the number of subjects was 500. In the combined group, AB, the number was 1000. The total test

Table 3.2.2. Item and Population Parameters Used for Data Generation in Group A/Test X.

	Ability Overlap		
	10%	30%	50%
Minimum a/Maximum a	0.5/1.5	0.5/1.3	0.2/1.4
Mean a	1.0	0.9	0.8
Minimum b/Maximum b	-2.5/-1.0	-2.0/-0.5	-1.74/-0.24
Mean b	-1.75	-1.25	-0.99
Minimum c/Maximum c	0.15/0.25	0.15/0.25	0.15/0.25
Mean c	0.2	0.2	0.2
Mean θ	-1.75	-1.25	-0.99
Standard Deviation θ	1.0	1.0	1.0

Table 3.2.3. Item and Population Parameters Used for Data Generation in Group B/Text Y.

	Ability Overlap		
	10%	30%	50%
Minimum a/Maximum a	0.4/1.2	0.5/1.5	0.4/1.4
Mean a	0.8	1.0	0.9
Minimum b/Maximum b	0.8/2.3	0.08/1.58	-0.4/1.1
Mean b	1.55	0.83	0.35
Minimum c/Maximum c	0.15/0.25	0.15/0.25	0.15/0.25
Mean c	0.2	0.2	0.2
Mean θ	1.55	0.83	0.35
Standard Deviation θ	1.0	1.0	1.0

Table 3.2.4. The Means and Standard Deviations of Raw and True Scores in Groups and Within Populations.

Population	Raw Scores		True Scores	
	Mean	Standard Deviation	Mean	Standard Deviation
10%				
Group A	43.3540	17.1421	43.5665	16.6771
Group B	53.5180	15.7783	53.4743	15.2927
30%				
Group A	47.1780	16.3286	47.1257	16.0229
Group B	55.0080	16.7355	55.0091	16.3546
50%				
Group A	48.3220	15.7537	48.2148	15.3218
Group B	53.9120	16.4395	53.7949	16.0494

lengths ranged from 64 with the shortest anchor (4) to 85 with the longest anchor (25). The combined test, XY, had from 124 to 145 items.

While the total number of test items was reasonable for the purposes of parameter estimation, the number of examinees required for the three-parameter logistic model was barely sufficient to provide good estimates of the parameters (Hulin, Lissak, and Drasgow, 1982). In those combinations where convergence was not possible with a 40 stage maximum, the pseudo-chance level, c_i, and occasionally the discrimination, a_i, were not estimated completely. In particular, the iterative procedure did not converge because the sample size was too small. Tables 3.2.5 to 3.2.7 show the stages to convergence. In all cases, however, the item difficulties, b_i, had at least stabilized. Difficulty estimates may be adversely affected by poorly estimated discrimination and pseudo-chance parameters (Thissen & Wainer, 1982), but only the characteristic curve equatings will be directly affected by the discrimination and pseudo-chance parameter estimates.

It was necessary to try various seeds for the random number generator before a data set could be had without either an item being answered correctly by all of the examinees in a group or missed by all of the examinees in a group. Recall that the groups are quite diverse and the anchor test, in particular, spans the entire range of difficulties. That is, there were instances of some very able individuals answering some very easy questions and vice-versa. There were two instances in which difficulty parameters were estimated very poorly (outliers) for no apparent reason. The items were not

Table 3.2.5. Stages Required for LOGIST Convergence, Group A.

Anchor Length	Ability Overlap		
	10%	30%	50%
25	26	19	22
13	30	20	21
7	40*	23	22
4	19	22	22

*Maximum Stages Allowed/Terminated

Table 3.2.6. Stages Required for LOGIST Convergence, Group B.

Anchor Length	Ability Overlap		
	10%	30%	50%
25	20	25	20
13	18	19	20
7	19	21	19
4	18	22	21

Table 3.2.7. Stages Required for LOGIST Convergence with Combined Groups (Concurrent).

Anchor Length	Ability Overlap 10%	30%	50%
25	40*	37	33
13	40*	39	39
7	40*	40*	40*
4	40*	40*	40*

*Maximum Stages Allowed/Terminated

exceptional in any noticeable way and were estimated without difficulty in other groups. Skaggs and Lissitz (1986) report an almost identical situation under very similar circumstances. In this study, equating was done both with and without the outliers. Results with the outliers removed were similar to what might be expected. Leaving such extreme outliers in the data set completely distorted the equatings. See the chapter on Results for a further discussion of outliers.

3.3 Equating Methods

Five equating methods were selected for this study:

1. a simultaneous estimation method
2. a characteristic curve method
3. mean and sigma method
4. orthogonal least squares method
5. ordinary least squares method.

This selection includes some of the more common methods of equating and some uncommon methods. The rationale for these choices is included in the following discussion.

Simultaneous Estimation Method

A very popular and relatively easy method of vertical equating with an anchor test design is to use a single LOGIST run on a combined data collection that is cleverly coded. That is, X and Y are the tests to be equated on groups A and B, respectively. Let W be the anchor test. Consider the total population (A+B) as having taken the test composed of all items (X+Y+W). For the examinees in group A, code the items in test Y as unreached and the examinees in group B as having not reached the items in test X. The resulting LOGIST run will place all of the ability and item parameters on a common scale. If we are content to report scores on the ability metric, then the equating is complete. Note that if N examinees answer the items in test X and M respond to test Y, then N+M will have scores on the anchor test, W. The anchor items, therefore, play a major role in the parameter estimation procedure. In many studies concerning true or raw score equating, the underlying equating is done with this concurrent or simultaneous LOGIST process.

Characteristic Curve Method

Recall the intuitive appeal of these methods in that they use all of the available information from the imposed IRT structure (Hambleton & Swaminathan, 1985, p. 210). The approach of Stocking and Lord (1983) is to minimize the mean squared differences of true scores. More precisely,

$$F = N^{-1} \sum_{j=1}^{N} (\xi_j - \xi_j^*)^2 \qquad [3.3.1]$$

is minimized with respect to α and β where $\theta_x^* = \alpha\theta_x + \beta$, N is the number of examinees, and ξ_j^* is the transformed true score of the j^{th} examinee on the common items.

To minimize F with respect to α and β, set the partial derivatives to zero:

$$\partial F/\partial\alpha = -2N^{-1} \sum_{j=1}^{N} (\xi_j - \xi_j^*) (\mu\xi_j^*/\partial\alpha)=0 \qquad [3.3.2]$$

$$\partial F/\partial\beta = -2N^{-1} \sum_{j=1}^{N} (\xi_j - \xi_j^*) (\partial\xi_j^*/\partial\beta)=0 \qquad [3.3.3]$$

Note that $b_{Y_i}^* = b_{Y_i} + \beta$. That is, $b_{Y_i}^*$ is the transformed i^{th} item difficulty from test Y to the scale of test X. $a_{Y_i}^* = a_{y_i} / \alpha$. Also, $\xi_j^* = \sum P_i^*(\theta_j)$ where, $P_i^*(\theta_j) = P_i(\theta_j, a_i^*, b_i^*, c_i)$ and the sum is over i. Therefore,

$$\partial\xi_j^*/\partial\alpha = \sum_{i=1}^{n} (b_{Y_i} \frac{\partial P_i^*(\theta_j)}{\partial b_{Y_i}^*} - a_{Y}i \ \alpha^{-2} \frac{\partial P_i^*(\theta_j)}{\partial a_{Y_i}^*}) \qquad [3.3.4]$$

$$\partial\xi_j^*/ \partial\alpha = \sum_{i=1}^{n} \partial P_i^*(\theta_j)/\partial\, b_{Y}^*i \qquad [3.3.5]$$

The partial derivatives of $P_i^*(\theta_j)$ from the three-parameter logistic model are substitued into equations 3.3.4 and 3.3.5 which are then substituted into 3.3.2 and 3.3.3. This system is then solved iteratively for α and β. A PASCAL program was written using the Fletcher, Powell (1963) method of solution suggested by Stocking and Lord (1983).

Haebara (1980) and Divigi (1980) have suggested minimizing other functions, but the approach of Stocking and Lord has been shown (1983) to be at least as accurate as their robust mean and sigma method. Divigi (1985) has recently proposed a mathematically simpler method that minimizes a chi-square statistic for item bias. It has the intuitive appeal of the Stocking and Lord approach but the function being minimized is quadratic and thus the derivative is linear and may be solved directly without rather complicated iterative procedures. Preliminary results show that this is a comparable method to the characteristic curve procedure. Vale (1986) states, "To date, there has been little evidence that any of the complex procedures are superior to simple mean and standard deviation transformations." The complex procedures to which Vale refers are the characteristic curve method of Stocking and Lord and Divigi's chi-square.

Mean and Sigma Method

A mean and standard deviation approach to test equating using IRT is very similar to the classical linear equating approach in which standardized raw scores are equated. In the IRT framework, standardized abilities are equated. "While the similarity is clear, the linear relationship that exists between θ_x and θ_y is a consequence of the theory, whereas in the linear equating procedure, this relationship is assumed" (Hambleton & Swaminathan, 1985, p. 204). Of course, all standardization takes place on the common items.

Since the item difficulties, b_i's, are on the same scale as the abilities, θ_j's, it is possible to use the common item difficulties rather than the abilities. The mean and sigma method of vertical equating has been extended to more robust procedures as previously discussed. Both robust and non-robust methods are popular since they are well-known and easy to apply.

For the purposes of this study, the non-robust method was chosen. In particular,

$$\hat{\alpha} = S_b / S_b \qquad [3.3.6]$$

$$\beta = \bar{b}_y - \bar{b}_x \qquad [3.3.7]$$

where S_{b_y} and S_{b_x} represent the standard deviations of the common item difficulties in test Y and X, respectively. $\bar{b}_y$ and $\bar{b}_x$ are the means of the common item difficulties in tests Y and X, respectively.

A non-robust procedure was chosen for the following reasons:

1. simplicity
2. stocking and Lord found their robust procedure yielded results very similar to their characteristic curve method
3. popularity, for example, CTB/McGraw-Hill (1982, p. 95).

Ordinary Least Squares Method

The method of ordinary least squares is a simple method for determining a line of best fit and is most commonly used outside of test equating. However, as pointed out earlier, it is not symmetric with respect to the tests. It is soley included as a bench-mark for the symmetric methods.

Orthogonal Least Squares Method

When the ordinary least squares is dismissed due to an obvious lack of symmetry, the solution that seems unassailable is an orthogonal least squares or first principal component or major axis approach. The theoretical flaw in this approach has been previously discussed, but it is clear that any approach to equating imperfect tests will fail the test of theory. The test of interest then, is the more empirical one. Little interest seems to have been paid to this rather straight-forward method and so it is included in this study, theoretical warts and all.

The major or principal axis of the set of anchor test difficulties is determined by the eigenvector corresponding to the largest eigenvalue of the (real, symmetric) variance-covariance matrix of these difficulties:

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \qquad [3.3.8]$$

The eigenvector $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and the corresponding eigenvalue, λ, are solutions to the equation:

$$\Sigma \underline{x} = \lambda \underline{x} \qquad [3.3.9]$$

or, equivalently,

$$\begin{bmatrix} \Sigma & \lambda I \end{bmatrix} \underline{x} = \underline{0} \qquad [3.3.10]$$

This system of equations has a non-trivial solution if, and only if:

$$\begin{bmatrix} \Sigma & \lambda I \end{bmatrix} = \begin{bmatrix} \sigma_x^2 \quad \lambda & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 - \lambda \end{bmatrix} = 0 \qquad [3.3.11]$$

Therefore,

$$\lambda 2 - \lambda (\sigma_x^2 + \sigma_y^2) + (\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2) = 0 \qquad [3.3.12]$$

are the larger eigenvalue is given by:

$$\lambda = 1/2 \left[\sigma_x^2 + \sigma_y^2 + ([\sigma_x^2 + \sigma_y^2]^2 - 4[\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2])^{1/2} \right] \qquad [3.3.13]$$

Substituting this numerical value back into [3.3.9] permits the calculation of

$$\hat{\alpha} = \frac{x_2}{x_1} = \frac{\lambda - \sigma_x^2}{\sigma_{xy}} = \frac{\sigma_{xy}}{\lambda - \sigma_y^2} \qquad [3.3.14]$$

while,

$$\hat{\beta} = b_x - \hat{\alpha}\, b_y \qquad [3.3.15]$$

3.4 Method of Evaluation

In an anchor test design, the common item difficulties are theoretically identical except for the mean and unit of measure. That is, the standardized common item difficulties are the same for test A and test B. However, the ability distributions overlap by 10%, 30%, and 50%. The resulting mean differences in ability and difficulty are 3.30, 2.08, and 1.34, respectively. Since the standard deviations are 1.0, the true equating constants are known to be:

$\alpha = 1.0$, $\beta = 3.30$ for the 10% overlap in abilities

$\alpha = 1.0$, $\beta = 2.08$ for the 30% overlap in abilities

$\alpha = 1.0$, $\beta = 1.34$ for the 50% overlap in abilities

The difference between the estimated equating constants and these true values is one criterion for judging the accuracy of one equating method/anchor length/ability overlap combination as compared to another such combination. Of course, such direct comparisons are not without problems. For example, it often is the case that the slope estimate for one combination is more accurate than for another combination while, at the same time, the intercept estimate is less accurate. In addition, one of the more popular equating procedures, the simultaneous estimation method, does not result in equating constants and, thus, could not be compared with the other methods of equating using this criterion. To overcome these limitations, another method of comparison was developed.

Let b_{iX} and b_{iY} represent the true item difficulties on tests X and Y, respectively, for i=1 to 60. Using the true equating constants, define

$$b_{iX}^{*} = \alpha b_{iY} + \beta \qquad [3.4.1]$$

Similarly, let $\hat{b}_{iX}$ and $\hat{b}_{iY}$ be the estimated item difficulties on test X and Y, respectively, for i=1 to 60. Using the estimated equating constants from one of the equating methods, define

$$\hat{b}_{iX}^{*} = \hat{\alpha}\, \hat{b}_{iY} + \hat{\beta} \qquad [3.4.2]$$

Now, the composite sets of difficulties $\{b_j\} = \{b_{iX}, b_{iX}^{*}\}$ and $\{\hat{b}_j\} = \{\hat{b}_{iX}, \hat{b}_{iX}^{*}\}$ for j=1 to 120 are each on a common scale. However, the scales will not be the same. To evaluate the equating method, it is reasonable to measure the strength of this unknown linear relationship. A correlation coefficient, γ, is suitably symmetric, but a linear transform of the correlation is more intuitive. In particular, if $\{z_j\}$ and $\{\hat{z}_j\}$ represent the standardized $\{b_j\}$ and $\{\hat{b}_j\}$, respectively, then define

$$\text{MSE} = \text{mean squared error} = E((z_j - \hat{z}_j)^2) \qquad [3.4.3]$$

where E is the expectation. Now, $E((z_j-\hat{z}_j)^2) = E(z_j^2 + \hat{z}_j^2 - 2z_j\hat{z}_j) = 1 + 1 - 2\gamma = 2(1-\gamma)$. As the strength of the linear relationship increases, the correlation will increase and the corresponding MSE will decrease. To summarize

$$MSE = 2(1-\gamma) \quad \text{and} \quad \gamma = 1-(MSE/2) \qquad [3.4.4]$$

Since MSE is a measure of a difference in z-scores, it is possible to have some feeling for its magnitude. Certainly MSE is bounded above by 2 and below by 0. [3.4.2] would indicate that MSE is composed of both parameter estimation errors and the error in estimated equating constants. If, however, the true equating constants are used with both the true difficulties and with the estimated difficulties, then the MSE would reflect the parameter estimation errors alone. That is, let [3.4.2] be replaced by

$$\hat{b}_{iX} = \alpha\hat{b}_{iY} + \beta \qquad [3.4.5]$$

and denote the corresponding MSE by PEE, parameter estimation error. While there is not a strictly additive relationship, PEE will provide a baseline measure for comparable MSEs. That is, a MSE that is nearly the same as the corresponding PEE will indicate that the estimated equating constants are performing nearly as well as the true equating constants. In short, MSE, as defined, yields both an absolute and relative measure of the accuracy of estimated equating constants based upon the true values of these constants.

Perhaps the primary reason for using MSE is that it will permit the comparison of a simultaneous estimation procedure with both the true equating on the true difficulties and with the separate estimation procedures that result in estimated equating constants. In particular, a simultaneous estimation procedure such as the concurrent LOGIST method used in this study will have all of the estimated item difficulties on a common, but undetermined, scale. This is a comparable situation to the equated difficulty estimates in [3.4.2]. Standardizing within this set of difficulty estimates will yield a comparable set of 120 estimated z-scores which may then be compared to the standardized true difficulties equated with the true equating constants.

Somewhat similar, but more complex, MSE measures were used by Marco (1983), Vale (1986), Petersen, Cook, and Stocking (1983), Skaggs and Lissitz (1986), and Lord (1982). In the Skaggs and Lissitz study, the equating coefficient estimates were not available since the concurrent LOGIST method of equating was used. The MSE was on the actual and equated raw scores. In the Vale study, a RMSE was used on the actual and equated difficulties. Again, a concurrent equating method was employed and, thus, the equating coefficients were unavailable. Vale notes that:

> RMSE is an index often used in evaluations of calibration and linking. It is useful, however, only if the scale onto which the parameters are linked is the same as the true scale. In simultaneous calibrations, the scale is defined to have a mean of 0 and a variance of 1, the parameters of the true distributions used in the simulations. In separate calibrations, the scale of one administration is typically expressed on the scale of the other. This makes RMSE comparisons with true parameters meaningless. RMSE was thus not computed for the separate calibration cells (p. 340).

The method used in this study avoids this problem by measuring correlation. That is, the error is not between estimated and true difficulties. Rather, the error is of the theoretically linear relationship between estimated and true difficulties.

The Petersen, Cook and Stocking study used a weighted MSE on raw scores. In this study, the equating coefficients were available from a characteristic curve and other equatings, but the data was from the Scholastic Aptitude Test (SAT) and, therefore, the true equating coefficients were not known.

Lord (1982) derives a formula for the standard error of a true-score equating. He uses this as a criterion in comparing several equating methods using real (SAT) data.

Stocking and Lord (1983) and Divigi (1985) use scatterplots of discriminations and difficulties from the separate calibrations or estimations and then insert the equating line. Better equatings will nearly bisect the point set. See Appendix A for similar scatterplots of difficulties.

Kolen (1984) creates a cross-validation statistic for evaluating and equating. He selects a sample of examinees and performs the equatings, then he selects a second distinct cross-validation sample and constructs a "mean-squared error in the proportion-correct score metric" (p. 33).

With real data, especially, it has been common to see equatings evaluated by comparing the equating to that of a well-established procedure, e.g. equipercentile in the horizontal case or concurrent LOGIST in the vertical case. Scale drift is another technique used with real data. An equating is judged to have drifted little if the direct equating of test A to B is similar to the results of equating A to T_1 and T_1 to T_2 and so on until T_n is equated with test B. That is, if the chain of equatings gives a result like the direct equating then there is little scale drift and the equating is judged accurate in this sense.

As a final measure with real data, a test may be equated with itself. An acceptable equating method should produce the identity transform when random sample of examinees is equated to another random sample of examinees all having taken the same test. Since equating is a lengthy and costly process, there is usually only one replication in this and other approaches to finding a suitable criterion. Phillips (1985) has shown that "single-replication error estimates may provide misleading assessments of the errors associated with equating a test to itself" (p. 59).

Since all equatings are theoretically flawed, empirical results must be the deciding factor. Or, as Divigi states, "There are not theoretical criteria for choosing among different methods, or for evaluating the quality of a particular method" (1985, p. 415).

Many studies use either real data or a concurrent equating method or both. Therefore, it is usually the case that both the true equating constants and the estimated equating constants are not available for comparison. It is also rather common to not report scores on the ability metric and thus to require some sort of true-score or raw-score equating. By using the MSE statistic described, this study permits the concurrent equating method to be compared to other methods that are in turn comparable to the true constants.

CHAPTER IV

RESULTS

4.1 Introduction

The criterion used in this study to judge the accuracy of an equating was defined by equation 3.4.3:

$$\mathrm{MSE} = E((z_j-\hat{z}_j)2)$$

Recall that the z_j were the standard scores for the set of true item difficulties from both tests X and Y put onto a common scale using the true equating and the $\hat{z}_j$ were the standard scores for the set of estimated item difficulties from both tests X and Y put onto a common scale using the estimated equating from one of the equating methods investigated. The $\hat{z}_j$ could also be the standard scores for the set of estimated item difficulties from the simultaneous estimation procedure. It was shown that MSE is a linear transformation of the correlation, γ, and is given by equation 3.4.4:

$$\mathrm{MSE} = 2(1-\gamma) \text{ or } \gamma = 1-(\mathrm{MSE}/2)$$

Still another way to conceptualize MSE is to note that the set of estimated and equated item difficulties, $\{\hat{b}_j\}$, will be on a common scale and so will the set of true and equated item difficulties, $\{b_j\}$. Except for measurement error, these equated sets of estimated and true item difficulties will differ only in origin and unit of measure. Therefore, the linear relationship between the sets may be found by equating standard scores:

$$\frac{b_j - \bar{b}_j}{\sigma b_j} = \frac{\hat{b}_j - \bar{\hat{b}}_j}{\sigma \hat{b}_j} \quad \text{or } b_j = a\hat{b}_j + b$$

where, $a = \sigma b_j / \sigma \hat{b}_j$ and $b = \bar{b}_j - a\bar{\hat{b}}_j$.

The MSE may thus be thought of as a lack-of-fit measure to this line.

The parameter estimation error, PEE, was the same MSE measure as above with one exception: the estimated equating used with the estimated item difficulties was replaced by the true equating. The result is a somewhat better measure of the error component due to parameter estimation procedures since PEE does not contain the equating error component. Equating methods that produce MSE criterion measures nearer the corresponding PEE will be judged more accurate in an absolute sense as opposed to being simply more accurate than another equating method.

Section 4.2 includes comparisons of equating methods for a fixed anchor test length within a particular overlap of ability and comparisons of anchor test length for a fixed equating method within a particular overlap of ability. The former comparisons are reasonable since MSE allows separate and simultaneous equating procedures an equal opportunity to match the true equating. The latter comparisons are reasonable because the anchor test items are nested, uniformly distributed, and all span the item difficulties of the combined tests.

Section 4.3 includes comparisons of anchor test length for a fixed overlap of ability within a particular equating method and comparisons of ability overlap for a fixed anchor test length within a particular equating method. Section 4.4 includes comparisons of equating methods for a fixed overlap of ability within a particular anchor test length and comparisons of ability overlap for a fixed equating method within a particular anchor test length. Both sections 4.3 and 4.4 present a problem not encountered in section 4.2 where all comparisons were done within a single ability overlap. The problem is due to the increased variability of the parameter estimates in instances where the differences in mean ability (or difficulty) are large. The greater variability of difficulty estimates, in particular, is attributable to both the minimal number of examinees and the full difficulty span of the anchor tests which require, in the most extreme cases, examinees with a mean ability of $\overline{\theta}$ to respond to items with difficulties of $\overline{\theta} \pm 4.05$. Such extreme mismatches of ability and item difficulty will cause the least appropriate items to have difficulty estimates that approach outlier status. The increased variability of the item difficulty estimates impacts in turn upon the correlation of estimated and true item difficulties and, thus, upon the MSE.

There are many possible solutions to the problem of poorly estimated item parameters. The fallible items may be rewritten, or replaced, the sample of examinees may be increased or broadened, new items may be added, or any combination of adjustments made. If, however, the test is beyond the development stage and the final data collected, then the only choices are to either remove or not remove the offending items. In this study, item number 67 on test Y in the 10% ability overlap with anchor test lengths of 13 and 25 was judged to have been extreme and removed from further computations. In addition, item number 82 on test Y in the 10% ability overlap with anchor test length 25 was also removed. These two items, 67 and 82, were anchor items and hence identical to items whose parameters were more accurately estimated within the more appropriate group. Also, rather surprisingly, item number 67 in the 10% ability overlap on test Y was reasonably estimated in the anchor test with 7 items. As previously mentioned, Skaggs and Lissitz (1986) reported a very similar situation in which seemingly innocent items achieved outlier status.

Since the difficulties of the extreme items discussed above were estimated to be more than 100, there was no thought of retaining the estimates for further calculations. In general, however, the decision to omit anchor items with less extreme parameter estimates is not easy. To be more specific:

1. In an equating situation with a short anchor test, each data point has proportionally greater importance than it might have were there more anchor items.

2. With as few as 4 anchor items, it is not always clear which item is the outlier. Figure A.4 illustrates this point.

3. The process of judging outlier status is arbitrary by its very nature if there is sampling or measurement error present.

4. As reported, the outlier status of an item may change when only the number of items in the anchor test is altered.

For these reasons, anchor test items whose parameter estimates were only moderately outlying were retained.

Returning to the discussion of MSE, recall that the increased variability of the estimated anchor item difficulties will affect the correlation and MSE. However, a decrease in correlation, or attenuation due to restriction of range (Allen & Yen, 1979, p. 34) may be compensated for by using the appropriate attenuation formula and then the corrected correlation may be used to compute the MSE that might be expected were the variability unchanged. In particular, the corrected correlation may be obtained from:

$$\rho_u^2 = \rho_r^2 k^2 / (1 + \rho_r^2 k^2 - \rho_r^2) \qquad [4.1.1]$$

where ρ_u is the correlation with the unrestricted variable, ρ_r is the correlation with the restricted variable, and k is the ratio of unrestricted standard deviation of the variable to the restricted standard deviation (Hopkins, et al., 1987, p. 86).

It is possible to shorten the MSE correction process described above by combining Equations 3.4.4 and 4.1.1:

$$(1-MSE_u/2)^2 = \frac{(1-MSE_r/2)^2 k^2}{1+(1-MSE_r/2)^2(k^2-1)}$$

or,

$$1+MSE_u^2/4-MSE_u = \frac{k^2(1+MSE_r^2/4 - MSE_r)}{1+(1+MSE_r^2/4-MSE_r)(k^2-1)}$$

If the relatively small second-order MSE_r terms are dropped, the result is

$$MSE_u \doteq 1 - \frac{k^2(1-MSE_r)}{1+(1-MSE_r)(k^2-1)} = \frac{MSE_r}{k^2+MSE_r(1-k^2)}$$

But, the second term of the denominator is also very small when compared to the first term and, thus,

$$MSE_u \doteq \frac{MSE_r}{k^2} \qquad [4.1.2]$$

This approximation has proven accurate for numbers in the range of this study and will be used in sections 4.3 and 4.4.

To complete the MSE or correlation correction, it is only necessary to observe that the ratio of standard deviations will be equal to the ratio of spans if a variable is uniformly distributed. This will, again, be an approximation in the case of estimated item difficulty parameters since the estimated distribution is only approximately uniform.

4.2 Anchor Length by Equating Method

4.2.1 Ten Percent Overlap in Abilities

Table 4.2.1 shows the mean squared error, MSE, and the parameter estimation error, PEE, for the population or groups of examinees with a 10% overlap in ability distributions. Note that two outliers (items number 67 and 82) were removed in the 25 item anchor test and one outlier (item number 67) removed in the 13 item anchor test. Table 4.2.2 contains the anchor item difficulty estimates within the 10% overlap in abilities. First, results will be discussed for a fixed anchor test length.

With the 25 item anchor test, the MSEs for all of the separate equating methods were acceptably accurate in the sense that the MSEs were very close to the corresponding PEE. That is, the largest of the MSEs for the separate equating procedures was .0076 while the PEE at this level was .0067. This represents an increase of approximately 13% of the PEE for the MSE of the mean and sigma equating method. The MSE of the simultaneous estimation procedure, however, was .0100 which represents a 49% increase over the PEE. Arbitrarily, increases larger than 25% were judged unacceptable.

With the 13 item anchor test, the results were similar. The largest of the MSEs for the separate equating procedures was .0068 which represents an increase of only 5% over the PEE of .0065. The MSE of the simultaneous estimation procedure was .0155 and this was 138% increase over the corresponding PEE. Certainly, the separate procedures performed better than the simultaneous estimation method

Table 4.2.1. Mean Squared Error for Equating - thod Versus Anchor Length in Populations with a 10% Ability Overlap

Anchor Length	Con-current	Charac-teristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares	Parameter Estimation Error
25(23)	.0100	.0062	.0076	.0071	.0064	.0067
13(12)	.0155	.0068	.0057	.0062	.0067	.0065
7	.0193	.0061	.0090	.0102	.0105	.0065
4	.0256	.0066	.0122	.0113	.0166	.0066

Table 4.2.2. LOGIST Estimates of Anchor Item Difficulties in the 10% Ability Overlap Population

Item Number	Length of Anchor Test							
	25		13		7		4	
	X	Y	X	Y	X	Y	X	Y
61	-3.7050	-1.6130	-3.8630	-1.6030	-4.3600	-1.6040	-3.9490	-1.5950
62	-4.3520	0.1550	-5.2250	0.1950	-5.2270	0.2000	-5.2610	0.1820
63	-0.1460	2.2030	-0.0830	2.1740	-0.0970	2.2250	-0.0950	2.2220
64	0.4550	4.0180	0.5300	2.6830	0.5300	2.6180	0.5330	2.7030
65	-3.2820	-0.3750	-3.7030	-0.3220	-3.8650	-0.3110		
66	-0.8060	4.0980	-0.7310	2.0090	-0.7590	2.0310		
67	0.0480	120.5150	0.1060	130.0700	0.1020	3.3050		
68	-6.0450	-0.5670	-6.6470	-0.5150				
69	-4.1820	0.0290	-4.5900	0.0190				
70	-3.1440	0.4450	-3.2570	0.4800				
71	-1.0650	3.4700	-1.0040	3.6400				
72	-0.0560	11.2970	-0.0050	6.8270				
73	0.7320	2.6730	0.7770	2.8710				
74	-4.8570	-0.9280						
75	-4.0340	-0.3950						
76	-7.6860	-0.3790						
77	-2.9420	-0.0700						
78	-3.8810	0.1450						
79	-2.7320	0.7160						
80	-0.8720	2.4650						
81	-0.4840	6.5070						
82	-0.3010	147.4010						
83	0.2660	4.6640						
84	0.3950	2.6980						
85	0.5220	3.8190						

for this most vertical equating situation when the anchor tests were of a somewhat traditional length.

With the 7 item anchor test, only the characteristic curve equating could be judged accurate. The mean and sigma method was next best but had a MSE of .0090 which represents an increase of 38% over the corresponding PEE of .0065. The simultaneous estimation procedure was the least accurate of all methods being off by 197% of the PEE.

With the 4 item anchor test, again, only the characteristic curve equating could be judged accurate. The method of orthogonal least squares was next with an increase of 71% over the PEE and the simultaneous estimation procedure was again least accurate with a MSE of .0256 or an increase of 288% of the PEE.

With the shorter anchor tests, the only acceptably accurate method of test equating was the characteristic curve method. The least accurate method in this extremely vertical equating process was the simultaneous estimation procedure. Furthermore, the percentage increase in error over the PEE was larger with the shorter anchor tests while the most accurate equating method, the characteristic curve method, had uniformly small MSEs over all anchor test lengths.

Tables 4.2.3-4.2.6 contain the estimated equating constants for the four separate equating methods. It is interesting to note that the characteristic curve method consistently underestimated both of the equating constants for all anchor test lengths.

Table 4.2.3. Estimated Equating Constants in the 10% Ability Overlap with a 25 Item Anchor Test, $\alpha = 1.0$, $\beta = 3.30$

Constant	Equating Method Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	0.8708	1.2624	0.7269	0.9149
$\alpha - \hat{\alpha}$	0.1292	-0.2624	0.2731	0.0851
$\hat{\beta}$	3.1708	4.7699	3.5780	3.9965
$\beta - \hat{\beta}$	0.1292	-1.4699	-0.2780	-0.6965

Table 4.2.4. Estimated Equating Constants in the 10% Ability Overlap with a 13 Item Anchor Test, $\alpha = 1.0$, $\beta = 3.30$

Constant	Equating Method: Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	0.7162	0.9276	1.1030	0.7110
$\alpha - \hat{\alpha}$	0.2838	0.0724	-1.1030	0.2890
$\hat{\beta}$	3.0162	3.6639	4.0660	3.1675
$\beta - \hat{\beta}$	0.2838	-0.3639	-0.7660	0.1325

Table 4.2.5. Estimated Equating Constants in the 10% Ability Overlap with a 7 Item Anchor Test, $\alpha = 1.0$, $\beta = 3.30$

	Equating Method			
Constant	Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	0.8179	0.7424	1.3903	0.6690
$\alpha - \hat{\alpha}$	0.1821	0.2576	-0.3903	0.3310
$\hat{\beta}$	3.1179	2.6303	3.8705	2.4897
$\beta - \hat{\beta}$	0.1821	0.6697	-0.5705	0.8103

Table 4.2.6. Estimated Equating Constants in the 10% Ability Overlap with a 4 Item Anchor Test, $\alpha = 1.0$, $\beta = 3.30$

Constant	Equating Method Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	0.7414	0.6992	1.5225	0.5904
$\alpha - \hat{\alpha}$	0.2586	0.3008	-0.5225	0.4096
$\hat{\beta}$	3.0414	2.3813	4.1514	2.1474
$\beta - \hat{\beta}$	0.2586	0.9187	-0.8514	1.1526

Note also, that the parameter estimation error, PEE, in Table 4.2.1 is nearly uniform across anchor test lengths. PEE is simply the MSE calculation using the true equating constants and is a measure of the parameter estimation error. In each of the three populations, 10%, 30%, and 50%, the anchor tests were nested and had identical spans and distributions. The purpose of this structure was to attempt to control these estimation errors within populations. The uniformity of the PEEs confirms the success of the design and allows the comparisons within each population or ability distribution overlap.

As a final observation, note the apparent reversal of MSEs with the mean and sigma equating method for the 25 and 13 item anchor tests. This pattern is unexpected since the errors should tend to get smaller with the larger anchor test lengths. The pattern with the simultanious estimation procedure and ordinary least squares method was as expected. The method of orthogonal least squares also seems to have the same reversal as the mean and sigma method. A possible explanation for this behavior can be had from Appendix A, Figures A.1, A.2, A.3, and A.5. Notice that the first item with a potential for outlier status is item number 72 in Table 4.2.2. This item has an estimated difficulty of nearly 7 in the anchor test of length 13 (actually, 12). Notice further, that the difficulty estimate of item 72 increases to more than 11 when estimated in the anchor test of length 25 (actually, 23). It appears to be the case that the mean and sigma and orthogonal least squares methods are rather sensitive to the presence of outliers. By way of contrast, the characteristic curve equating method would seem rather robust against such outlying values and the simultaneous estimation method perhaps the least influenced by outliers.

Now, consider the results for each equating method across anchor test length. The simultaneous estimation or concurrent procedure was inaccurate at all anchor test lengths but, least accurate with the shortest anchors. The characteristic curve method was acceptably accurate at all anchor test lengths and the errors were relatively constant. The mean and sigma method performed in a very similar manner to the orthogonal least squares and ordinary least squares methods in that the errors were acceptably small for the two longer anchor test lengths but, the errors were too large to be judged acceptable for the two shorter anchor test lengths.

4.2.2 Thirty Percent Overlap in Abilities

Table 4.2.7 contains the MSEs and PEEs for the population or examinee groups with a 30% ability overlap. No outliers were removed from this data set. Table 4.2.8 contains the anchor item difficulty estimates for this population. Results will be first discussed for a fixed anchor test length.

With the 25 item anchor test, the MSE was smallest for the characteristic curve method, but acceptabley small for the method of ordinary least squares as well. The MSE for the simultaneous

Table 4.2.7. Mean Squared Error for Equating Method Versus Anchor Length in Populations with a 30% Ability Overlap

Anchor Length	Con-current	Charac-teristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares	Parameter Estimation Error
25	.0159	.0118	.0181	.0178	.0130	.0120
13	.0176	.0122	.0214	.0230	.0144	.0118
7	.0232	.0162	.0363	.0399	.0241	.0117
4	.0367	.0145	.0239	.0174	.0222	.0119

Table 4.2.8. LOGIST Estimates of Anchor Item Difficulties in the 30% Ability Overlap Population

Item Number	Length of Anchor Test 25 X	25 Y	13 X	13 Y	7 X	7 Y	4 X	4 Y
61	-2.2290	-1.3850	-2.1010	-1.4020	-2.0840	-1.4220	-2.2320	-1.4380
62	-1.9810	0.2070	-1.8810	0.2020	-1.8590	0.2060	-2.0390	0.1670
63	-0.1700	1.9580	-0.1630	1.9760	-0.1610	2.0150	-0.1670	2.0560
64	0.6660	2.7000	0.6240	2.8390	0.6100	2.8550	0.6310	2.8980
65	-2.4990	-0.2830	-2.4990	-0.2810	-2.4770	-0.2760		
66	-0.7860	1.4820	-0.7880	1.5170	-0.7750	1.5260		
67	0.1240	6.8180	0.1130	6.6670	0.1130	7.2310		
68	-3.3390	-0.5270	-3.1150	-0.5240				
69	-2.0450	0.0690	-2.0160	0.0630				
70	-1.7380	0.4610	-1.6850	0.4570				
71	-0.8880	2.0220	-0.8800	2.1810				
72	-0.0030	2.2760	-0.0050	2.3140				
73	0.7750	2.1190	0.7600	2.1750				
74	-3.5960	-0.8310						
75	-2.5350	-0.3890						
76	-2.5220	-0.3760						
77	-2.3930	-0.0330						
78	-2.0520	0.1920						
79	-1.8690	0.7830						
80	-0.7070	1.1370						
81	-0.4260	1.6890						
82	-0.2280	5.6550						
83	0.2900	2.5390						
84	0.3820	1.9590						
85	0.6130	2.5780						

estimation procedure was the next best, but judged unacceptable showing a 33% increase over the corresponding PEE. The methods of orthogonal least squares and mean and sigma were both in the 50% increase range.

With the 13 item anchor test, the results were similar. The characteristic curve equating method was the most accurate but, the ordinary least squares procedure was also judged acceptable with an increase in error over the corresponding PEE of 22%. None of the remaining three equating procedures was acceptably accurate.

With the 7 item anchor test, all the methods were unacceptable. The best method was the characteristic curve method, once again. However, this time the MSE of .0162 represented an increase of 37% over the PEE of .0117. The largest errors were with the mean and sigma and orthogonal least squares methods. In both of these cases the percentage of increase in MSE over PEE was in excess of 200%. To explore this more fully, graphs with the true equating line and each of the estimated equating lines were constructed for the anchor test of length seven in the 30% ability overlap. These appear as Figures 4.2.1-4.2.4. The slopes and intercepts are from Table 4.2.9. The simultaneous estimation or concurrent procedure could not be included since it does not result in estimated equating constants. In Figure 4.2.1, the characteristic curve equating line is seen to have responded to the presence of the outlier, item number 72 (Table 4.2.8). In Figure 4.2.2, the mean and sigma equating line is seen to have been pulled far from the true equating line. In Figure 4.2.3,

Table 4.2.9. Estimated Equating Constants in the 30% Ability Overlap with a 25 Item Anchor Test, $\alpha = 1.0$, $\beta = 2.08$

Constant	Equating Method: Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	0.9670	1.4815	0.6055	1.1422
$\alpha - \hat{\alpha}$	0.0330	-0.4815	0.3945	-0.1422
$\hat{\beta}$	2.0470	3.0017	2.0030	2.6149
$\beta - \hat{\beta}$	0.0330	-0.9217	0.0770	-0.5349

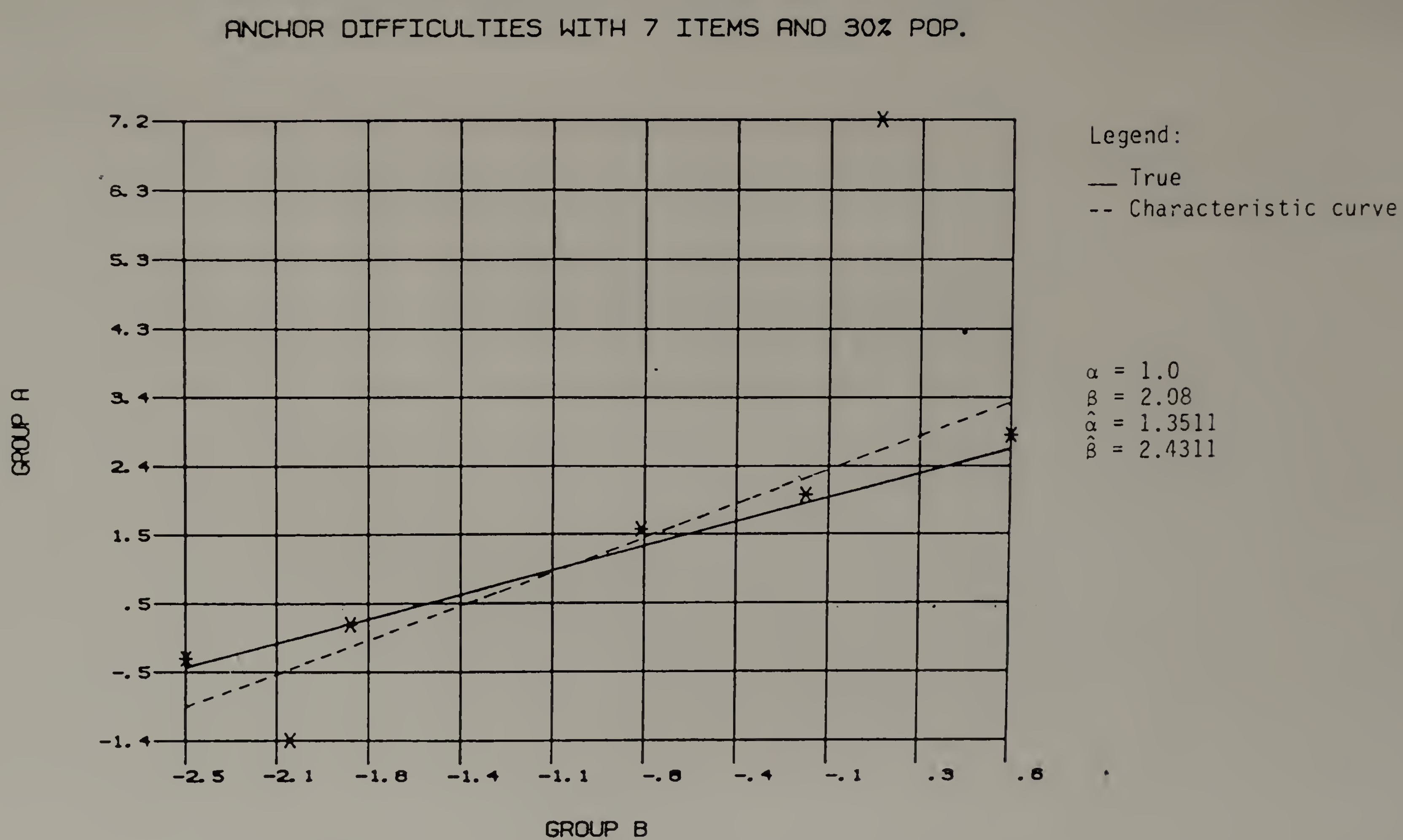


Figure 4.2.1. True equating line verses characteristic curve equating line.

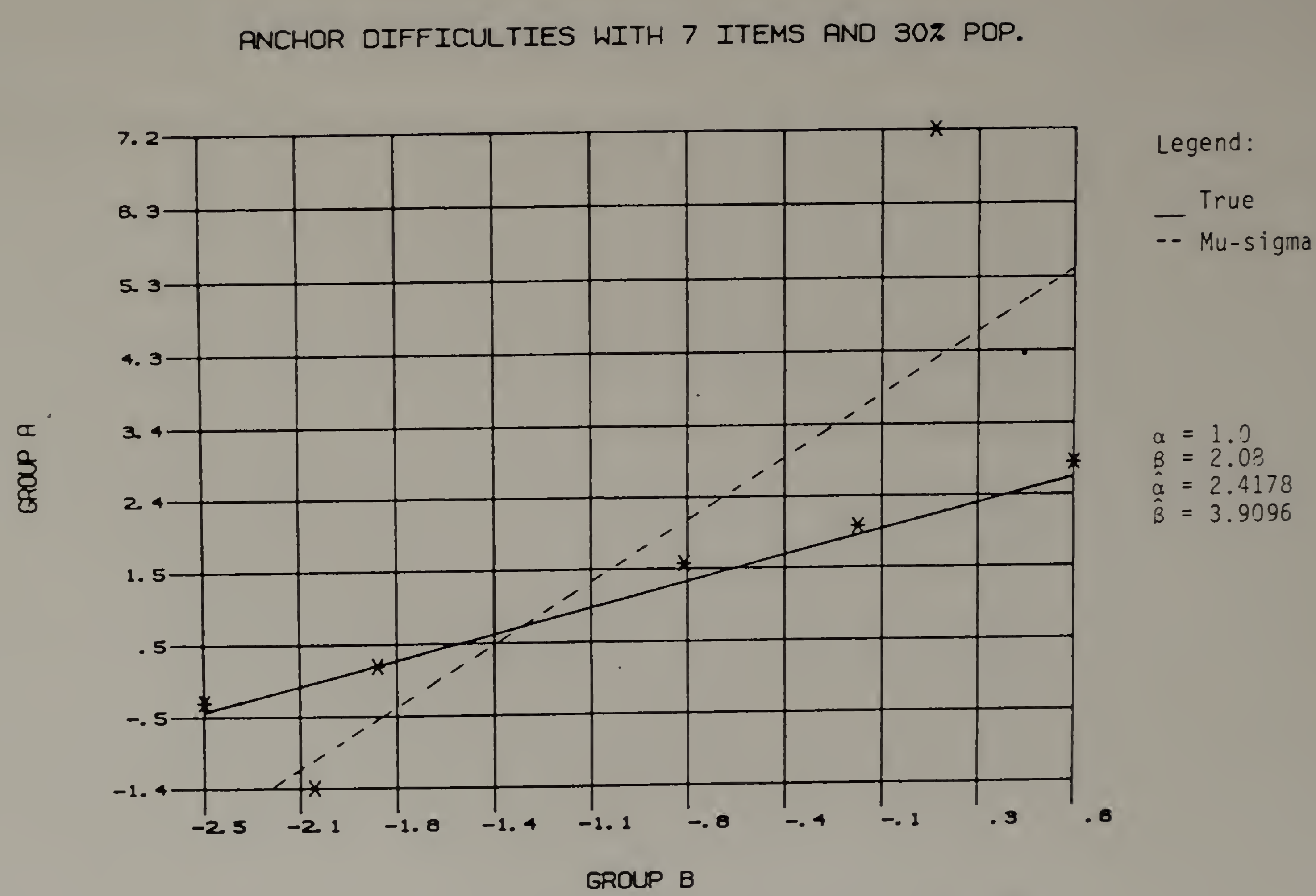
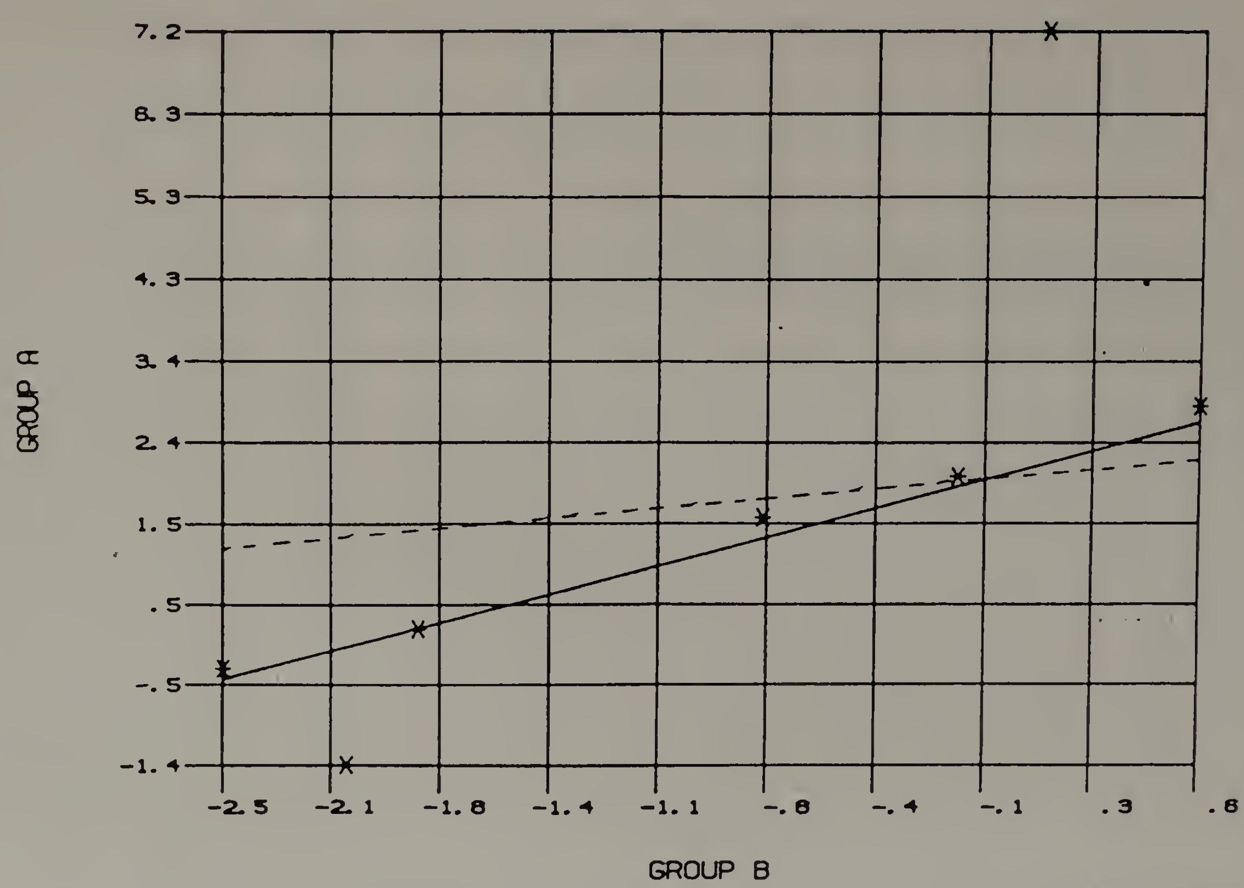


Figure 4.2.2. True equating line verses mean and sigma equating line.

ANCHOR DIFFICULTIES WITH 7 ITEMS AND 30% POP.



Legend:
__ True
-- First principal component

$\alpha = 1.0$
$\beta = 2.08$
$\hat{\alpha} = 0.3391$
$\hat{\beta} = 2.0388$

Figure 4.2.3. True equating line verses orthogonal least squares equating line.

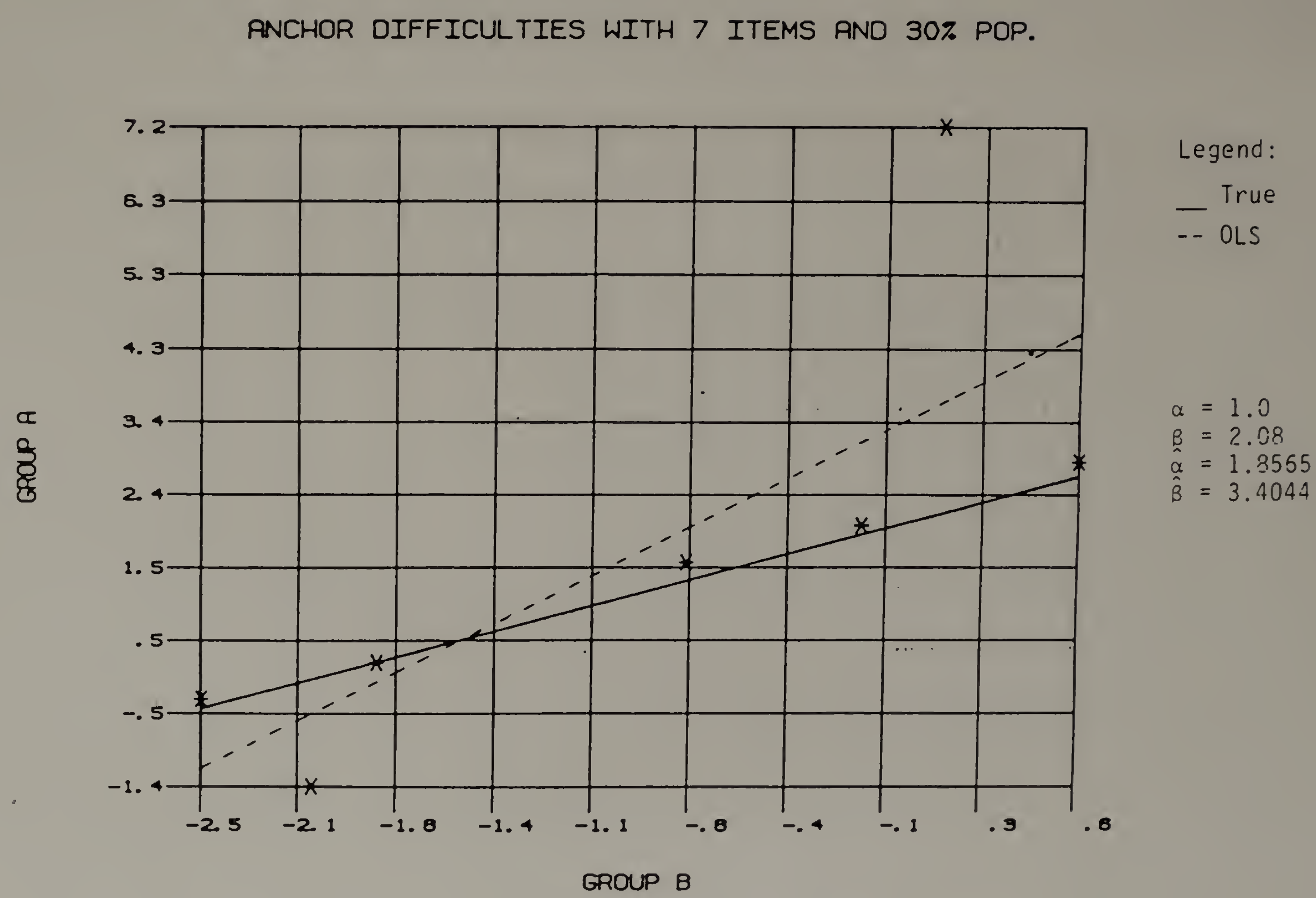


Figure 4.2.4. True equating line verses ordinary least squares equating line.

the major axis or equating line from the method of orthogonal least squares is seen to have been pushed from the true equating line in an effort to minimize the sum of perpendicular or orthogonal distances from the outlier. In Figure 4.2.4, the OLS equating line has attempted to minimize the sum of vertical distances to the outlier. It is in a position that would be between the characteristic curve equating line and the mean and sigma equating line. Clearly, the mean and sigma equating was more affected by the presence of item number 72 than the OLS or characteristic curve equatings. As for the concurrent estimation procedure, it is unique in that it is the only one of the equatings studied in which the MSE does not decrease in going from the 7 item anchor test to the 4 item anchor test. That is, the presence of the outlier is not noticable from the MSEs for the simultaneous estimation procedure. In addition, none of the MSEs for the simultaneous estimation method are at an acceptable level when compared with the corresponding PEEs. This was also the case in the 10% ability overlap.

With the 4 item anchor test, only the characteristic curve equating method provided an acceptably accurate equating of test scores with a MSE of .0145 which was a 22% increase over the PEE of .0119.

In summary, only the characteristic curve equating method and the method of ordinary least squares provided acceptably accurate equatings with the longer anchor tests of 25 and 13 items. The best equating procedure with the shorter anchor tests (7 and 4 items) was the characteristic curve method. However, this was only judged acceptable with the 4 item anchor test due to the presence of an exceptional value in the 7 item anchor test.

Tables 4.2.9-4.2.12 contain the estimated equating constants for the four separate equating methods. It is interesting to note that the characteristic curve equating method overestimated both of the equating constants for all but the 25 item anchor test. Again, the PEE in Table 4.2.7 is nearly uniform across anchor test lengths as desired. Finally, note that the outlier in this data set affects the equating methods in the same manner that the outlier did in the previous data set and, hence, tends to confirm the conjectures concerning the impact of outlying values on the various equating methods.

Considering the results for each equating procedure across anchor test length, the simultaneous estimation method was, again, unacceptably accurate at all anchor test lengths and least accurate with the shortest anchors. The characteristic curve method was acceptable at all anchor test lengths except 7 where, even though the most accurate of the methods studied, the presence of the outlying value was sufficient to produce a MSE that was 38% more than the corresponding PEE. The mean and sigma and orthogonal least squares methods were similar in that neither produced an acceptably accurate equating at any anchor length and the outlier with the 7 item anchor test caused a reversal of the MSEs for the 4 and 7 item anchor tests.

Table 4.2.10. Estimated Equating Constants in the 30% Ability Overlap with a 13 Item Anchor Test, $\alpha = 1.0$, $\beta = 2.08$

	Equating Method			
Constant	Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	1.0569	1.6701	0.5225	1.2859
$\alpha - \hat{\alpha}$	-0.0569	-0.6701	0.4775	-0.2859
$\hat{\beta}$	2.1369	3.0946	1.9294	2.7045
$\beta - \hat{\beta}$	-0.0569	-1.0146	0.1506	-0.6245

Table 4.2.11. Estimated Equating Constants in the 30% Ability Overlap with a 7 Item Anchor Test, $\alpha = 1.0$, $\beta = 3.30$

Constant	Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
	Equating Method			
$\hat{\alpha}$	1.3511	2.4178	0.3391	1.8565
$\alpha - \hat{\alpha}$	-0.3511	-1.4178	0.6609	-0.8565
$\hat{\beta}$	2.4311	3.9096	0.0412	-1.3244
$\beta - \hat{\beta}$	-0.3511	-1.8296	0.0412	-1.3244

Table 4.2.12. Estimated Equating Constants in the 30% Ability Overlap with a 4 Item Anchor Test, $\alpha = 1.0$, $\beta = 2.08$

Constant	Equating Method: Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	1.2024	1.3992	0.7050	1.3425
$\alpha - \hat{\alpha}$	-0.2024	-0.3992	0.2950	-0.3425
$\hat{\beta}$	2.2824	2.2150	1.5728	2.1625
$\beta - \hat{\beta}$	-0.2024	-0.1350	0.5072	-0.0825

The OLS equating method was acceptably accurate with the two longer anchors only and suffered the same reversal in response to the presence of the outlier as the mean and sigma and orthogonal least squares methods of equating test scores.

4.2.3 Fifty Percent Overlap in Abilities

Table 4.2.13 shows the MSEs and PEEs for the 50% overlap in abilities. No outliers were removed from this data set. Table 4.2.14 contains the anchor item difficulty estimates for this population. Results will again first be discussed for the fixed anchor test length.

With the 25 item anchor test, all of the equating methods studied were acceptably accurate.

With the 13 item anchor test, all of the equating methods were again acceptably accurate. It would appear that in this least vertical situation and with reasonably long anchor tests, the choice of equating method is less than critical.

With the 7 item anchor test, the only acceptably accurate equating methods were the simultaneous estimation and characteristic curve procedures. Since the concurrent or simultaneous method faired so poorly at all anchor test lengths in the more vertical populations, it must be the case that this method is rather sensitive to the mean ability differences in the groups under investigation. The characteristic curve method of test equating did not show this tendency at all.

Table 4.2.13. Mean Squared Error for Equating Method Versus Anchor Length in Populations with a 50% Ability Overlap

Anchor Length	Concurrent	Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares	Parameter Estimation Error
25	.0378	.0372	.0345	.0346	.0367	.0368
13	.0372	.0394	.0353	.0355	.0361	.0398
7	.0381	.0388	.0516	.0656	.0478	.0387
4	.0601	.0389	.0738	.1186	.0730	.0394

Table 4.2.14. LOGIST Estimates of Anchor Item Difficulties in the 50% Ability Overlap Populations

Item Number	Length of Anchor Test							
	25		13		7		4	
	X	Y	X	Y	X	Y	X	Y
61	-2.5280	-2.7750	-2.5630	-2.3760	-2.4588	-2.4520	-2.4400	-2.3690
62	-1.3350	0.2410	-1.3360	0.2440	-1.3190	0.2150	-1.3340	0.1930
63	-0.1990	1.1320	-0.2000	1.2600	-0.2070	1.2460	-0.2190	1.2200
64	0.4850	1.4750	0.4740	1.4280	0.4800	1.3980	0.4450	1.3560
65	-1.6530	-0.2540	-1.6800	-0.2460	-1.6250	-0.2820		
66	-0.8500	0.6850	-0.8780	0.7220	-0.8630	0.6950		
67	0.0730	2.1160	0.0660	2.2700	0.0680	2.3440		
68	-3.2310	-0.5150	-3.3040	-0.5200				
69	-1.2790	0.0880	-1.3030	0.0860				
70	-1.0370	0.4430	-1.0270	0.4480				
71	-0.9028	0.9150	-0.9170	0.9180				
72	-0.0370	1.5000	-0.0380	1.5060				
73	0.7880	1.5980	0.7950	1.6560				
75	-3.6180	-0.9700						
76	-1.4830	-0.4230						
77	-1.5730	-0.3370						
78	-1.4260	0.0100						
79	-1.1240	0.2300						
80	-1.0870	0.7270						
81	-0.7900	0.5440						
82	-0.4440	0.8540						
83	-0.2830	1.2590						
84	0.2720	1.2280						
85	0.3630	1.6470						
86	0.5350	1.7810						

With the 4 item anchor test, the only acceptably accurate equating procedure was the characteristic curve approach. With the simultaneous estimation method, the MSE was 53% larger than the corresponding PEE. Since the MSEs did consistently increase with decreasing anchor test length for the simultaneous estimation method, it must be the case that this method is also affected by the number of items on the anchor test. Again, the characteristic curve equating method did not show this tendency.

Tables 4.2.15-4.2.18 contain the estimated equating constants for the four separate equating methods. In this population, the characteristic curve estimated equating constants behaved precisely as in the 30% overlap in ability population. That is, the estimates of both constants were consistently greater than the true values for all but the 25 item anchor test.

Considering the results for each equating method across anchor test length, the simultaneous estimation procedure was acceptably accurate for all but the 4 item anchor test. The characteristic curve equating method was acceptably accurate at all anchor test lengths. The mean and sigma and orthogonal least squares were again similar in that the equatings were acceptably accurate for the two longer anchor tests and inaccurate for the shorter test lengths. The method of ordinary least squares was nearly the same as the mean and sigma and orthogonal least squares but, the MSE for the 7 item anchor test was marginally (24% increase over the PEE) acceptable.

Table 4.2.15. Estimated Equating Constants in the 50% Ability Overlap with a 25 Item Anchor Test, $\alpha = 1.0$, $\beta = 1.34$

Constant	Equating Method: Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	0.9598	0.9734	1.0317	0.8405
$\alpha - \hat{\alpha}$	0.0402	0.0266	-0.0317	0.1595
$\hat{\beta}$	1.2998	1.3768	1.4277	1.2609
$\beta - \hat{\beta}$	0.0402	-0.0368	-0.0877	0.0791

Table 4.2.16. Estimated Equating Constants in the 50% Ability Overlap with a 13 Item Anchor Test, α = 1.0, β= 1.34

Constant	Equating Method: Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
$\hat{\alpha}$	1.0595	1.0528	0.9416	0.9001
$\alpha - \hat{\alpha}$	-0.0595	-0.0528	0.0584	0.0999
$\hat{\beta}$	1.3995	1.5245	1.4236	1.3860
$\beta - \hat{\beta}$	-0.0595	-0.1845	-0.0836	-0.0460

Table 4.2.17. Estimated Equating Constants in the 50% Ability Overlap with a 7 Item Anchor Test, $\alpha = 1.0$, $\beta = 1.34$

Constant	Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
	Equating Method			
$\hat{\alpha}$	1.2059	1.5540	0.6233	1.4424
$\alpha - \hat{\alpha}$	-0.2059	-0.5540	0.3767	-0.4424
$\hat{\beta}$	1.5459	1.7619	0.9774	1.6678
$\beta - \hat{\beta}$	-0.2059	-0.4219	0.3626	-0.3278

Table 4.2.18. Estimated Equating Constants in the 50% Ability Overlap with a 4 Item Anchor Test, $\alpha = 1.0$, $\beta = 1.34$

Constant	Characteristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares
	Equating Method			
$\hat{\alpha}$	1.0622	1.3957	0.7045	1.3260
$\alpha - \hat{\alpha}$	-0.0622	-0.3957	0.2955	-0.3260
$\hat{\beta}$	1.4022	1.3212	0.7165	1.2603
$\beta - \hat{\beta}$	-0.0622	0.0188	0.6235	0.0797

4.3 Anchor Length by Ability Overlap

4.3.1 Concurrent

Table 4.3.1 contains the MSEs for the concurrent or simultaneous estimation equating method. At first glance, the pattern seems reversed in that larger errors might well be expected with the more vertical equating situation at 10% ability overlap than with either of the less vertical, 30% or 50%, ability overlaps. However, recall that in Section 4.1 it was pointed out that in order to compare MSEs across differing ability overlaps it will be necessary to correct for attenuation since the MSE is simply a linear transformation of the correlation between the entire set of equated true anchor item difficulties and the entire set of estimated equated anchor item difficulties. Equation 4.1.2 supplies the approximation of the corrected or predicted MSE, $MSE_u \doteq \frac{MSE_r}{k^2}$. Recall that k may, in turn, be approximated by the ratio of spans. The subscripts u and r indicate unrestricted (larger) and restricted (smaller) variances, respectively.

To illustrate, calculate the estimated anchor item difficulty spans in the case of 25 item anchor test.

10% ability overlap: 11.2970-(-7.6860)=18.9830

30% ability overlap: 6.8180-(-3.5960)=10.4140

50% ability overlap: 2.1160-(-3.6180)= 5.7340

Table 4.3.1. Mean Squared Error for Anchor Length Verses Ability Overlap with a Concurrent Equating

Anchor Length	Ability Overlap 10%	30%	50%
25	.0100	.0159	.0378
13	.0155	.0176	.0372
7	.0193	.0232	.0381
4	.0256	.0367	.0601

Use the true equating constants to adjust the spans for the PEE correction:

10% ability overlap: 18.9830-3.30=15.683

30% ability overlap: 10.4140-2.08= 8.334

50% ability overlap: 5.7340-1.34= 4.394

The ratios of these spans approximate k:

15.683/8.334 = 1.8818 to predict MSE at 10% from 30% ability overlap

8.334/4.394 = 1.8967 to predict MSE at 30% from 50% ability overlap

Therefore, the predicted MSEs are:

$$\text{predicted } MSE_{10\%} = \frac{MSE_{30\%}}{k^2} = \frac{.0120}{1.8818^2} = .0034$$

$$\text{predicted } MSE_{30\%} = \frac{MSE_{50\%}}{k^2} = \frac{.0368}{1.8967^2} = .0102$$

While the predicted MSE of .0102 compares rather favorably to the actual MSE of .0120 at this level, the .0034 prediction is rather far from the actual MSE of .0067. To account for this imprecision, notice that the estimated difficulty of item 72 in test Y for the 10% ability overlap population is 11.2970. The decision was made to retain such items but, if the calculations were done with this one item removed, the result would be a span of 14.1930 - 3.30 = 10.893 for a k of 10.893/8.334 = 1.3071. The resulting prediction would be:

$$\text{predicted } MSE_{10\%} = \frac{.0120}{1.3071^2} = .0070$$

This predicted .0070 compares favorably with the actual .0067.

Very similar results may be obtained with other combinations of equating constants and anchor lengths. The results with the true constants and the 13 item anchor test, for example, predict a MSE of .0068 at the 10% ability overlap while the actual MSE is .0065 and a predicted MSE of .0120 at the 30% ability overlap compared to an actual MSE of .0118.

While it is certainly the case that the shorter anchor tests and less accurate equating methods do not yield such close predictions, it is nontheless rather clear that, once corrected for, MSEs at different ability overlaps are reasonably uniform. This is really not too surprising since the difficulty involved with the most vertical equating is parameter estimation. The correction to the MSE brings these estimates to a more uniform variability. Therefore, with the correction for attenuation in place, there will be little or no difference in PEEs due to mean ability differences. It is, however, the case that different equating procedures are affected in different ways by these mean ability differences.

Returning to the concurrent or simultaneous estimation procedure, Table 4.3.1 indicates that this method of equating has increasing MSEs for decreasing numbers of anchor test items and, thus, simultaneous estimation will not be as accurate with the shortest anchor test lengths. In addition, the actual MSEs were greater than the predicted MSEs when corrected for attenuation. This would indicate that the method of simultaneous estimation does not equate scores as accurately as desired when there are large differences in the mean ability levels of the groups. For example, the predicted MSEs at the 10% and 30% ability overlaps with an anchor test of length 25 are .0045 and .0105, respectively. The actual corresponding MSEs are .0100 and .0159. For the 13 item anchor test, the predicted MSEs are .0101 and .0112 for the 10% and 30% ability overlaps, respectively, while the corresponding actual MSEs are .0155 and .0176.

The method of concurrent or simultaneous estimation gave acceptably accurate equatings only in the population with a 50% overlap in abilities and, even there, not with the 4 item anchor test. Note that LOGIST only converged in the less vertical situations and with the longer anchor tests (Table 3.2.7) and required a minimum of 33 stages overall. In the separate parameter estimates, LOGIST only failed to converge once and never required more than 30 stages when it did converge (Tables 3.2.5 and 3.2.6).

The concurrent estimation procedure was relatively immune to the influence of outlying values in the sense that the MSEs showed a consistent (albeit inaccurate) pattern whether an outlier was present or not.

4.3.2 Characteristic Curve

Table 4.3.2 contains the MSEs for the characteristic curve equating procedure. With one exception, the equatings were all acceptably accurate and predictable when corrected for attenuation.

Table 4.3.2. Mean Squared Error for Anchor Length verses Ability Overlap with a Characteristic Curve Equating

Anchor Length	Ability Overlap 10%	30%	50%
25	.0062	.0118	.0372
13	.0068	.0122	.0394
7	.0061	.0162	.0388
4	.0066	.0145	.0389

The exception was in the case of the 7 item anchor test in the 30% ability overlap population. The outlying value did have an impact on this procedure in this case but, it was less of an impact than with the remaining separate equating procedures. The predicability would indicate that this approach to test equating is relatively robust to differences in mean ability and may be a preferred method in the most vertical equating situations.

In addition, the MSEs were relatively uniform over the different lengths of anchor test. In every other equating procedure studied, the expected pattern was seen: an increase in MSE as the number of anchor items decreased. The characteristic curve method was the only method of those studied that could be considered for use with exceptionally short anchor tests.

4.3.3 Mean and Sigma

Table 4.3.3 contains the MSEs for the mean and sigma equating method. Were it not for the 30% ability overlap population, the results would be clear: with longer anchor tests (25 and 13 items) the mean and sigma method was accurate but, with the shorter anchor tests (7 and 4 items) the method was not accurate. The 30% overlap in abilities was unique, however, in that it retained a relatively extreme outlier. As previously discussed, the mean and sigma method is sensitive to these outlying values and this is no doubt the reason for the exception.

Table 4.3.3. Mean Squared Error for Anchor Length verses Ability Overlap with a Mean and Sigma Equating

Anchor Length	Ability Overlap 10%	30%	50%
25	.0076	.0181	.0345
13	.0057	.0214	.0353
7	.0090	.0363	.0516
4	.0122	.0239	.0738

When the MSEs were predicted using the formula to correct for attenuation, the results were mixed but, the actual MSEs wei for the most part larger than the corresponding predicted MSEs. That is, the method of mean and sigma test equating does seem to be affected by differing mean ability differences, but not to the extent of the simultaneous estimation procedures. This method is perhaps most affected by the presence or absence of outliers and the length of the anchor test.

4.3.4. Orthogonal Least Squares

Table 4.3.4 contains the MSEs for the orthogonal least squares equating method. The results for this relatively unused approach to test equating are nearly identical to the results for the mean and sigma method, one of the most popular equating methods. Both methods are sensitive to outliers, even though they tend to react or compensate differently. Both methods were acceptably accurate with the longer anchor tests and inaccurate with the shorter anchor tests when outliers were not present. Again, the major axis approach had MSEs that were only somewhat predictable after correction for attenuation and thus was also a bit sensitive to differences in mean ability.

Table 4.3.4. Mean Squared Error for Anchor Length Verses Ability Overlap with an Orthogonal Least Squares Equating

Anchor Length	Ability Overlap 10%	30%	50%
25	.0071	.0178	.0346
13	.0062	.0230	.0355
7	.0102	.0399	.0656
4	.0113	.0174	.01186

4.3.5 Ordinary Least Squares

Table 4.3.5 contains the MSEs for the OLS equating method. The results for this procedure are clear: with the two longer anchor tests, the equatings were acceptably accurate and with the two shorter anchor tests, the equatings were not acceptably accurate. Recall that OLS was less affected by the presence of the outlier in the 7 item anchor, 30% ability overlap equating situation and this no doubt accounted for the acceptably accurate MSEs as compared with the unacceptably accurate MSEs from the mean and sigma and orthogonal least squares methods.

As with the previous two equating methods, the predictability of the MSEs was mixed. Of course, the lack of symmetry and, hence, equity would preclude the actual use of OLS in a real test equating situation. As a benchmark, however, it does tend to put into perspective the other methods of test equating.

4.4 Equating Method by Ability Overlap

4.4.1 25 Item Anchor Test

Table 4.4.1 contains the MSEs for the 25 item anchor test. In the least vertical, 50% ability overlap, population, all of the equating methods produced acceptably accurate equatings. In the most vertical, 10% ability overlap, population, all but the simultaneous estimation procedures produced acceptably accurate equatings. Due to

Table 4.3.5. Mean Squared Error for Anchor Length Verses Ability Overlap with an Ordinary Least Squares Equating

Anchor Length	Ability Overlap 10%	30%	50%
25	.0064	.0130	.0367
13	.0067	.0144	.0361
7	.0105	.0241	.0478
4	.0166	.0222	.0730

Table 4.4.1. Mean Squared Error for Equating Method versus Ability Overlap with an Anchor Length of 25

	Equating Method					
Ability Overlap	Con-current	Charac-teristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares	Parameter Estimation Error
10%	.0100	.0062	.0076	.0071	.0064	.0067
30%	.0159	.0118	.0181	.0178	.0130	.0120
50%	.0378	.0372	.0345	.0346	.0367	.0368

the presence of a moderate outlier in the 30% ability overlap population, only the characteristic curve and OLS equating methods were acceptably accurate. Note that simultaneous estimation method was more affected by mean ability differences and that the mean and sigma and orthogonal least squares methods were more affected by the presence of a moderate outlier. Both the characteristic curve and OLS methods were predictably and acceptably accurate at all levels of mean ability difference for the 25 item anchor test.

4.4.2 13 Item Anchor Test

Table 4.4.2 contains the MSEs for the 13 item anchor test. Precisely the same results hold for this anchor test length as held for the 25 item anchor test.

4.4.3 7 Item Anchor Test

Table 4.4.3 contains the MSEs for the 7 item anchor test. In the least vertical, 50% ability overlap, population, all but two of the equating methods were acceptably accurate. The two inaccurate methods of test equating were mean and sigma and orthogonal least squares. OLS was barely acceptable. Clearly, these methods are more affected by the length of the anchor test than the other methods. In the most vertical, 10% ability overlap, population, only the characteristic curve equating method was able to overcome the combination of large mean ability differences and relatively short anchor test. The

Table 4.4.2. Mean Squared Error for Equating Method verses Ability Overlap with an Anchor Length of 13

		Equating Method				
Ability Overlap	Con-current	Charac-teristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares	Parameter Estimation Error
10%	.0155	.0068	.0057	.0062	.0067	.0065
30%	.0176	.0122	.0214	.0230	.0144	.0118
50%	.0372	.0394	.0353	.0355	.0361	.0398

Table 4.4.3. Mean Squared Error for Equating Method verses Ability Overlap with an Anchor Length of 7

	Equating Method					
Ability Overlap	Con-current	Charac-teristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares	Parameter Estimation Error
10%	.0193	.0061	.0090	.0102	.0105	.0065
30%	.0232	.0162	.0363	.0399	.0241	.0117
50%	.0381	.0388	.0516	.0656	.0478	.0387

outlier in the 30% ability overlap was most pronounced in the 7 item anchor test. The presence of this outstanding value was sufficient to make every single equating unacceptably accurate.

4.4.4 4 Item Anchor Test

Table 4.4.4 contains the MSEs for the 4 item anchor test. The characteristic curve equating method was acceptably accurate at all ability overlaps and it was the only acceptably accurate equating method at any ability overlap.

To briefly summarize the results:

1. The characteristic curve equating method was the most accurate of all the procedures studied, being inaccurate in only one instance where a sufficiently large outlier skewed all of the equatings.
2. The simultaneous estimation procedure was not able to accurately deal with the combination of small sample sizes, short anchor tests, and diverse abilities.
3. With the smaller mean differences in ability and the longer anchor tests, all methods of equating were reasonably accurate, although some were more sensitive to outlying values than others.
4. The correction for attentuation helped explain some facets of the data for this study. It was necessary because the criterion, MSE, was tied so closely to a correlation. The MSE may not be the most reeasonable criterion for evaluating

Table 4.4.4. Mean Squared Error for Equating Method verses Ability Overlap with an Anchor Length of 4

	Equating Method					
Ability Overlap	Con-current	Charac-teristic Curve	Mean and Sigma	Orthogonal Least Squares	Ordinary Least Squares	Parameter Estimation Error
10%	.0256	.0066	.0122	.0113	.0166	.0066
30%	.0367	.0145	.0239	.0174	.0222	.0119
50%	.0601	.0389	.0738	.1186	.0730	.0394

the accuracy of equatings across mean ability differences. That is, a major problem in extreme vertical equating is getting good parameter estimates. The MSE criterion is such that poorly estimated difficulties may increase the difficulty span which may in turn increase the correlation and, hence, decrease the error. The most extreme parameter estimations may thus yield the smallest MSEs. This seems unreasonable.

5. MSE does seem to be a reasonable criterion to use for comparing anchor test length and equating methods. It is a criterion based upon the true equating and one which is able to compare the simultaneous estimation procedure with separate equating methods.

6. Table 4.4.5 contains all of the PEEs. The uniformity within ability overlaps confirms the nested, full span, uniformly distributed anchor test design. That is, differences in MSEs at different anchor test lengths within the same ability overlap may be attributed solely to the length of the anchor test, as desired. The seemingly reversed pattern of MSEs and PEEs across ability overlaps was adequately explained by correcting the error measures for attenuation due to restriction of range.

Table 4.4.5. Parameter Estimation Error (PEE) for Anchor Length verses Population Ability Overlap

Anchor Length	Ability Overlap 10%	30%	50%
25	.0067	.0120	.0368
13	.0065	.0118	.0398
7	.0065	.0117	.0387
4	.0066	.0119	.0394

CHAPTER V

CONCLUSIONS

Briefly, the purposes of this study were to investigate the effects of the following on the accuracy of an equating of test scores:

1. length of the anchor test
2. equating method
3. group mean ability differences

In particular, it was the intent of this study to determine which combinations of the above factors would produce an acceptably accurate equating and, more generally, how the various factors interact.

Concerning the length of the anchor test, the results make the following conclusion inescapable:

> Acceptably accurate equatings are more likely to result when longer anchor tests are used. However, under particular combinations of method and mean ability difference, even the shortest anchor test was able to produce an acceptably accurate equating of test scores.

As for the equating method, the conclusions must be carefully conditioned:

This study involved relatively small sample sizes and relatively large group mean ability differences. Test equating under these circumstances is difficult. The simultaneous estimation procedure was most affected in that the method was sensitive to both large group mean ability differences and short anchor tests. That is, convergence of the maximum likelihood parameter estimation procedure was less likely under these conditions. The simultaneous estimation procedure was relatively unaffected by the presence of moderate outliers.

The characteristic curve method of test equating was able to accurately equate scores under even the most extreme combinations of anchor test length and mean ability difference. It was clearly the method of choice for such difficult equating.

The mean and sigma, orthogonal least squares, and ordinary least squares methods were somewhat comparable. With the longer anchor test lengths and less diverse abilities, these methods would all produce acceptably accurate equatings of test scores. They did not perform well with the shortest anchor tests and they were affected adversely by the presence of moderate outliers.

Differences in the mean ability between groups were large and resulted in the following conclusions:

The simultaneous estimation procedure was most affected. In direct contrast, the characteristic curve method was unaffected when the MSEs were corrected for attenuation. The other methods were somewhat affected.

As might have been expected, certain combinations of factors performed at very different levels of acceptability:

The simultaneous estimation procedure gave acceptable results only in the least vertical situation and never with the shortest anchor test. The characteristic curve method was unacceptable only in the presence of a most extreme outlier. All other methods failed here as well. With anchor tests of a more traditional length and in less vertical situations, any of the methods studied should give reasonable results.

These conclusions lead to the following recommendations which must also be conditioned by the limitations of the study:

1. Use as long as anchor test as possible but, be aware that as few as 4 anchor test items will suffice under certain circumstances.
2. The characteristic curve method or an equivalent is to be preferred for short anchor and highly vertical test equating.
3. While both commonly and easily used, mean and sigma and simultaneous estimation procedures are not recommended for short anchor and highly vertical test equating.
4. Anchor test items whose parameter estimates are outlying should be removed. If it is determined to leave moderate

outliers in the data set, then equating methods least affected by outliers should be used, namely, simultaneous estimation if possible or the characteristic curve method.

5. As large a range of difficulty as possible should be used for the anchor items but, parameter estimation will then become more difficult and outliers will appear.

The equating of test scores using an anchor test design would seem to require further study. In particular, it would be informative to increase the sample size to see if this is the major cause of the difficulty with the simultaneous estimation procedure. The use of ability overlaps in the 70% to 80% range might also impact upon a number of the conclusions of this study. A robust mean and sigma method would be a natural choice to compete with the characteristic curve approach. Anchor tests with a fixed span of difficulties would prohibit certain comparisons, but enhance others. Even shorter anchor tests could be investigated.

APPENDIX A

SCATTERGRAMS OF ANCHOR ITEM DIFFICULTIES

IBMI

87/07/28

HP 7550A

4399170

87/07/28 19.29.41

1 user generated page.

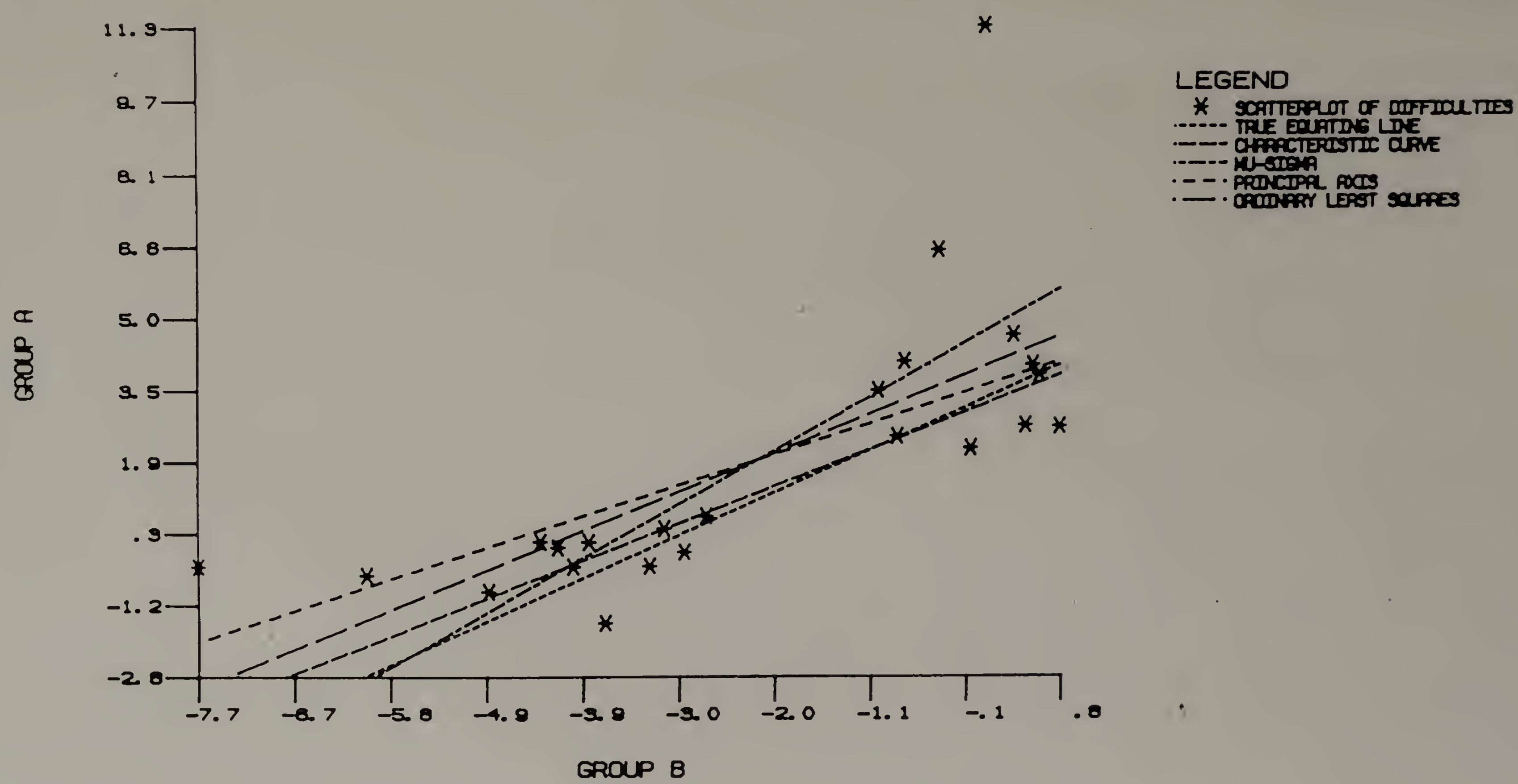


Figure A.1. Anchor difficulties with 25 (23) items and a 10% overlap.

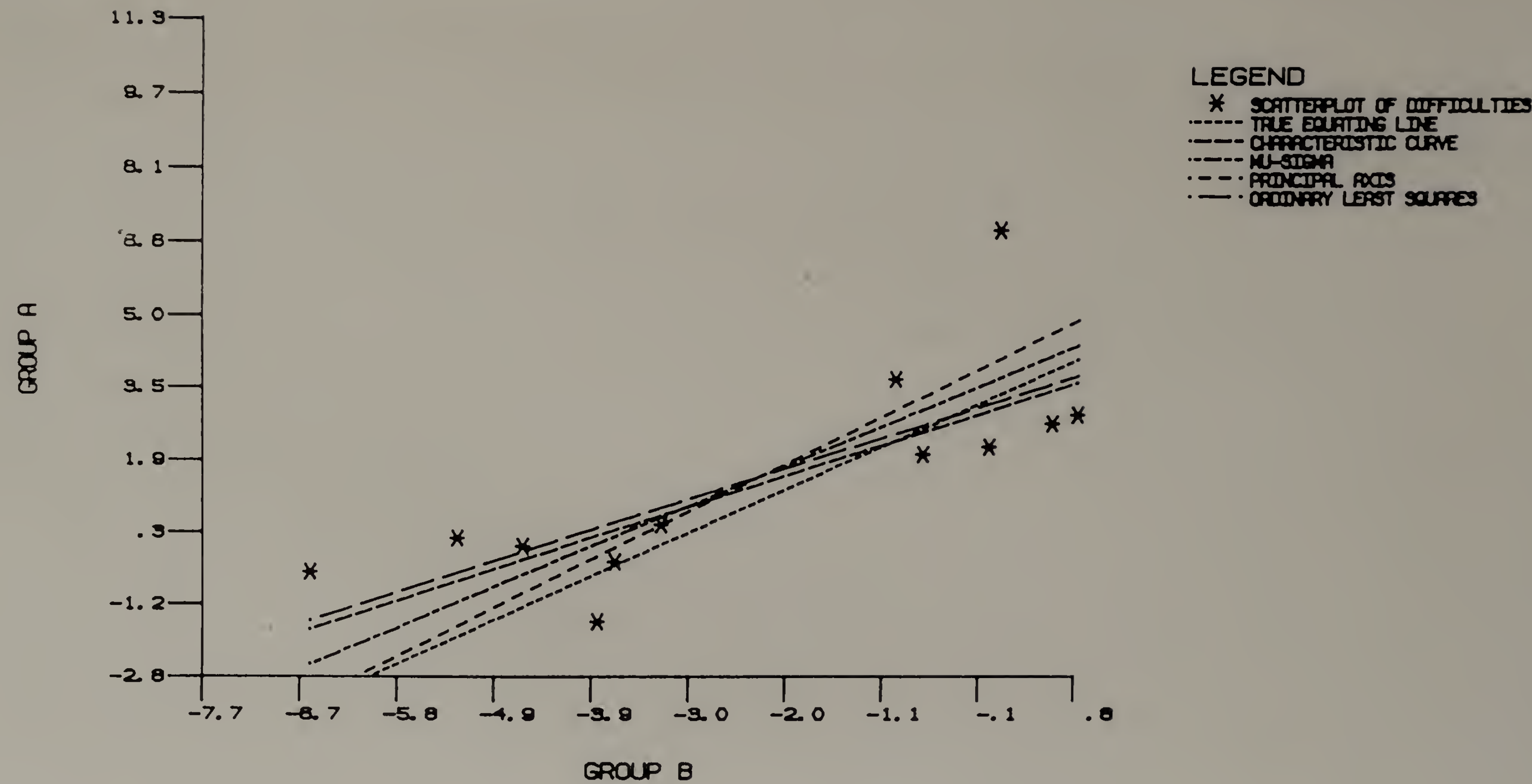


Figure A.2. Anchor difficulties with 13 (12) items and a 10% overlap.

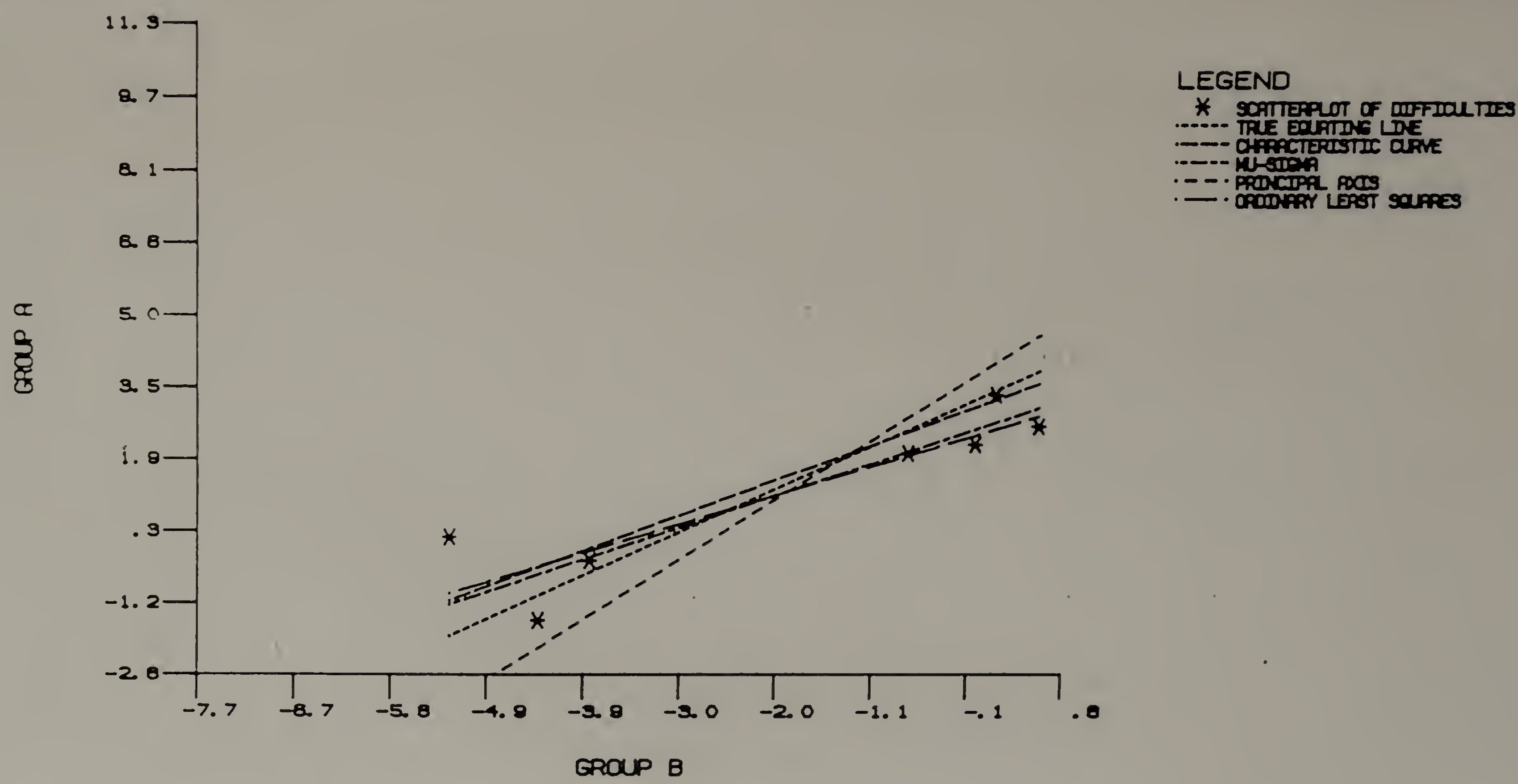


Figure A.3. Anchor difficulties with 7 items and a 10% overlap.

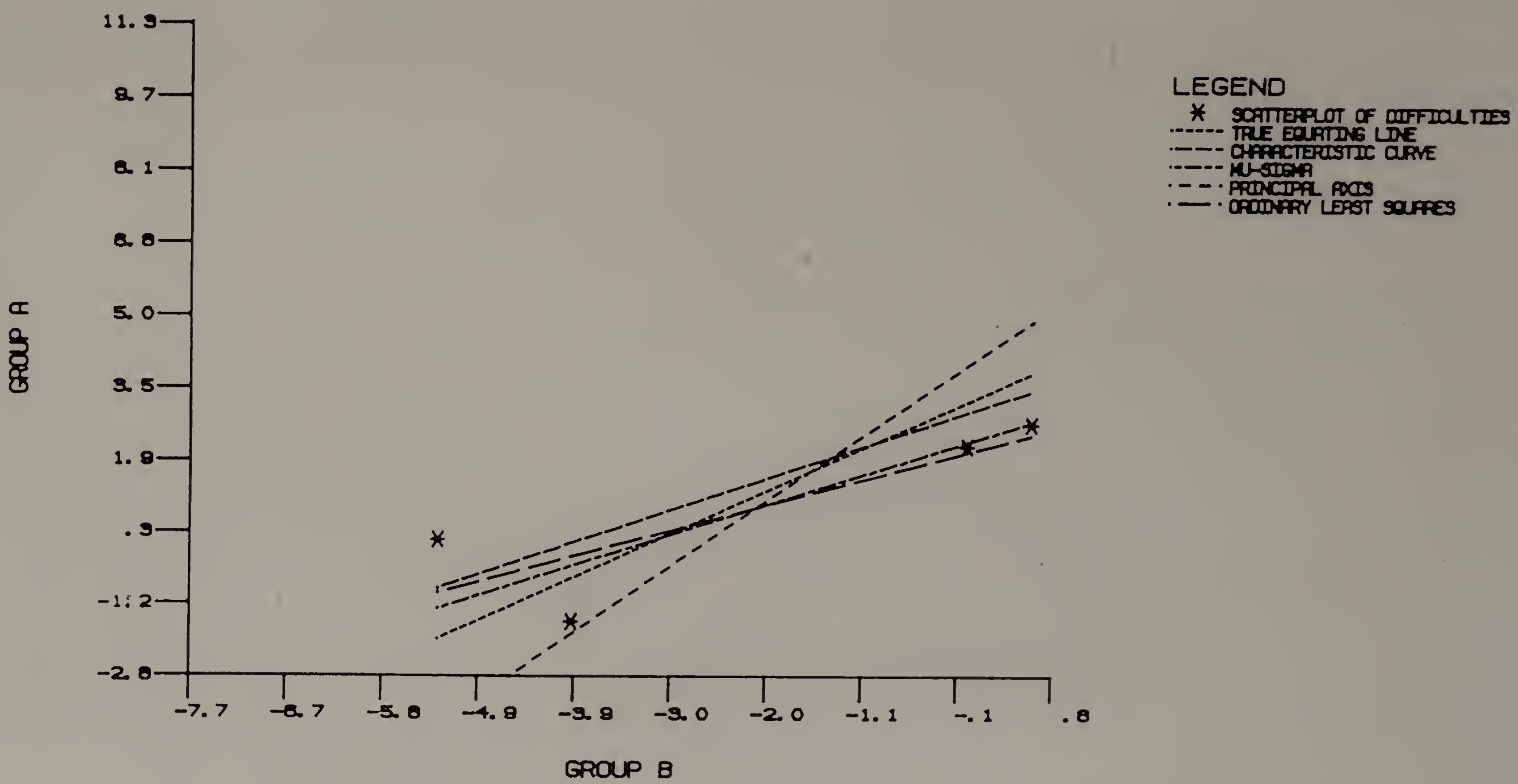


Figure A.4. Anchor difficulties with 4 items and a 10% overlap.

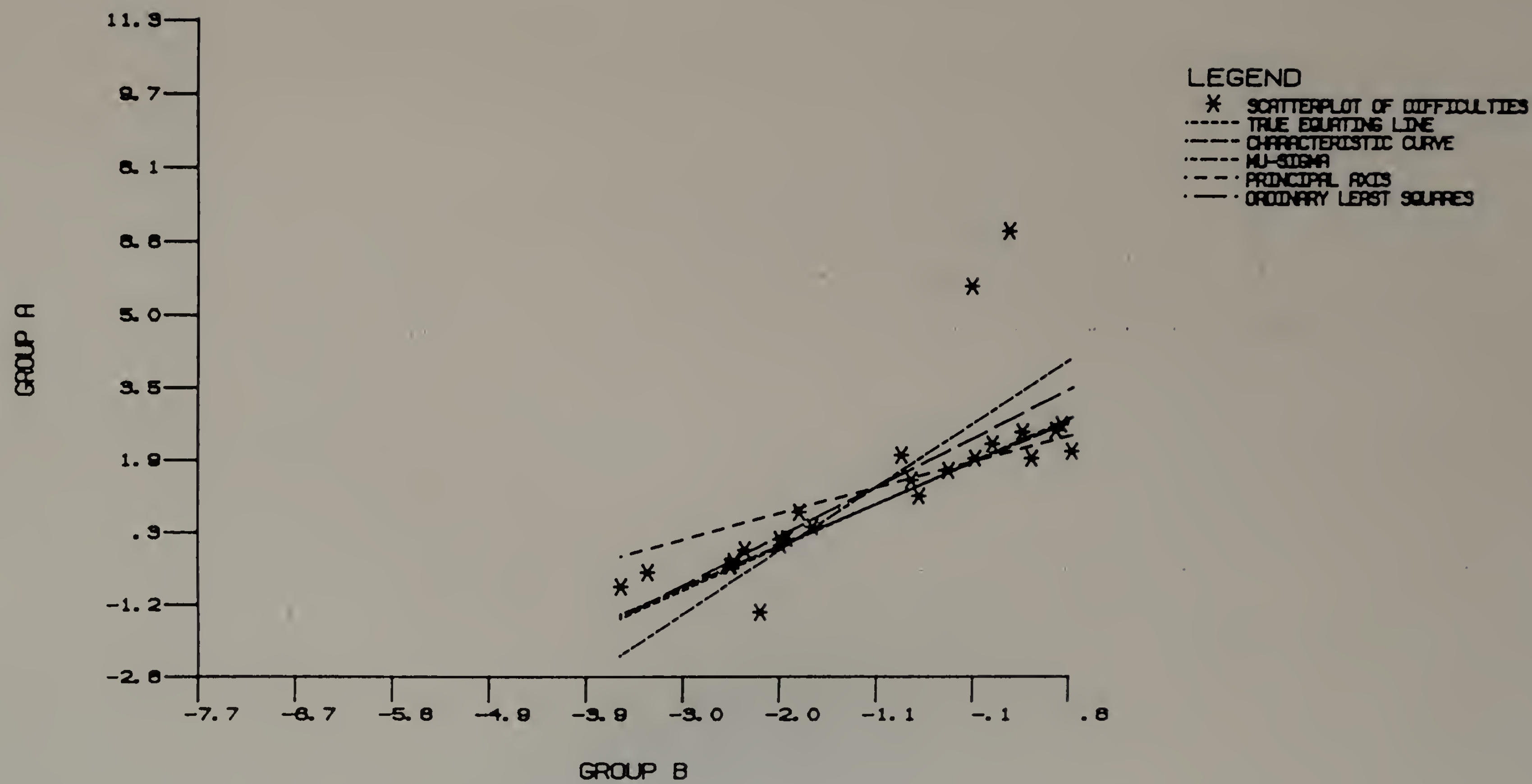


Figure A.5. Anchor difficulties with 25 items and a 30% overlap.

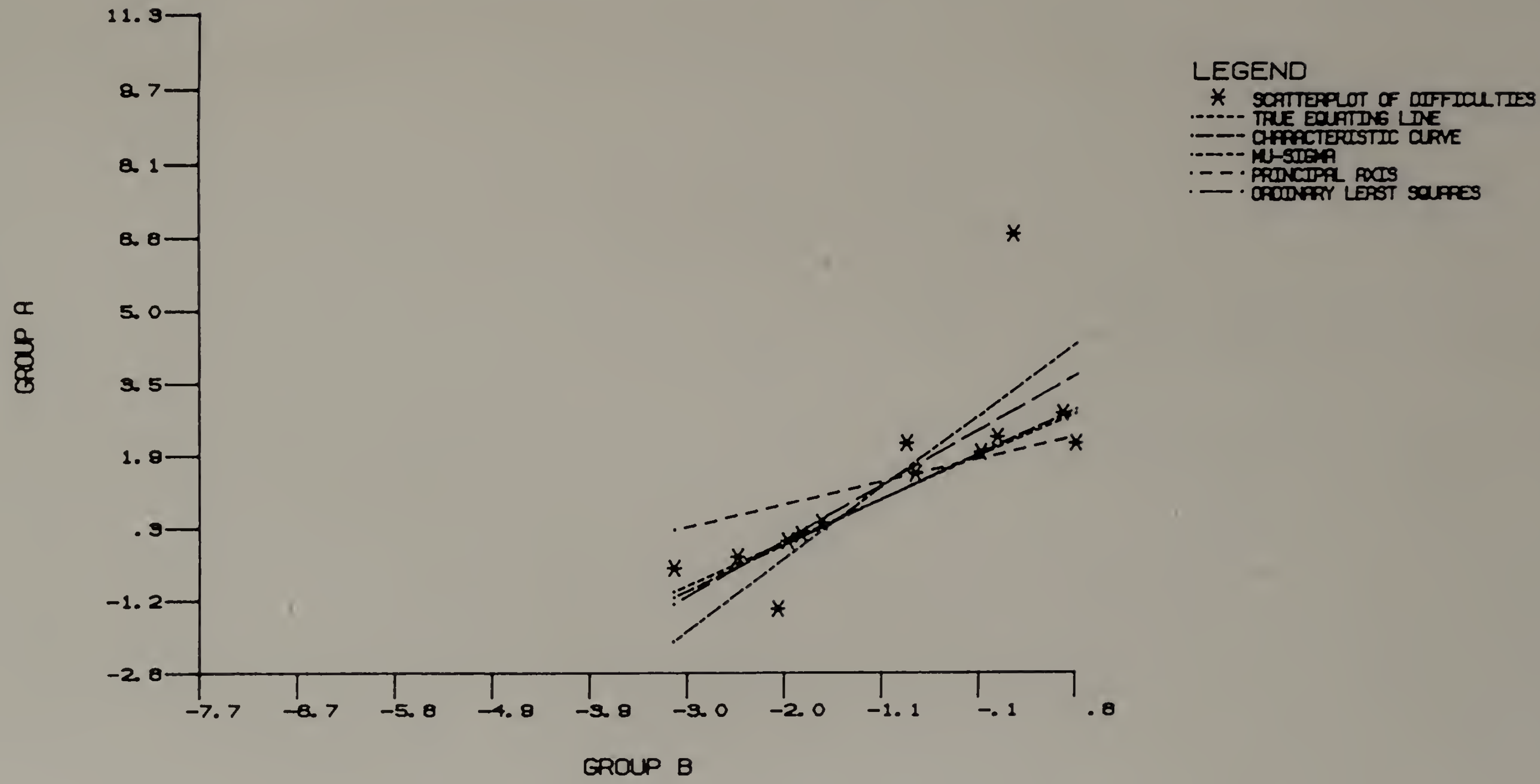


Figure A.6. Anchor difficulties with 13 items and a 30% overlap.

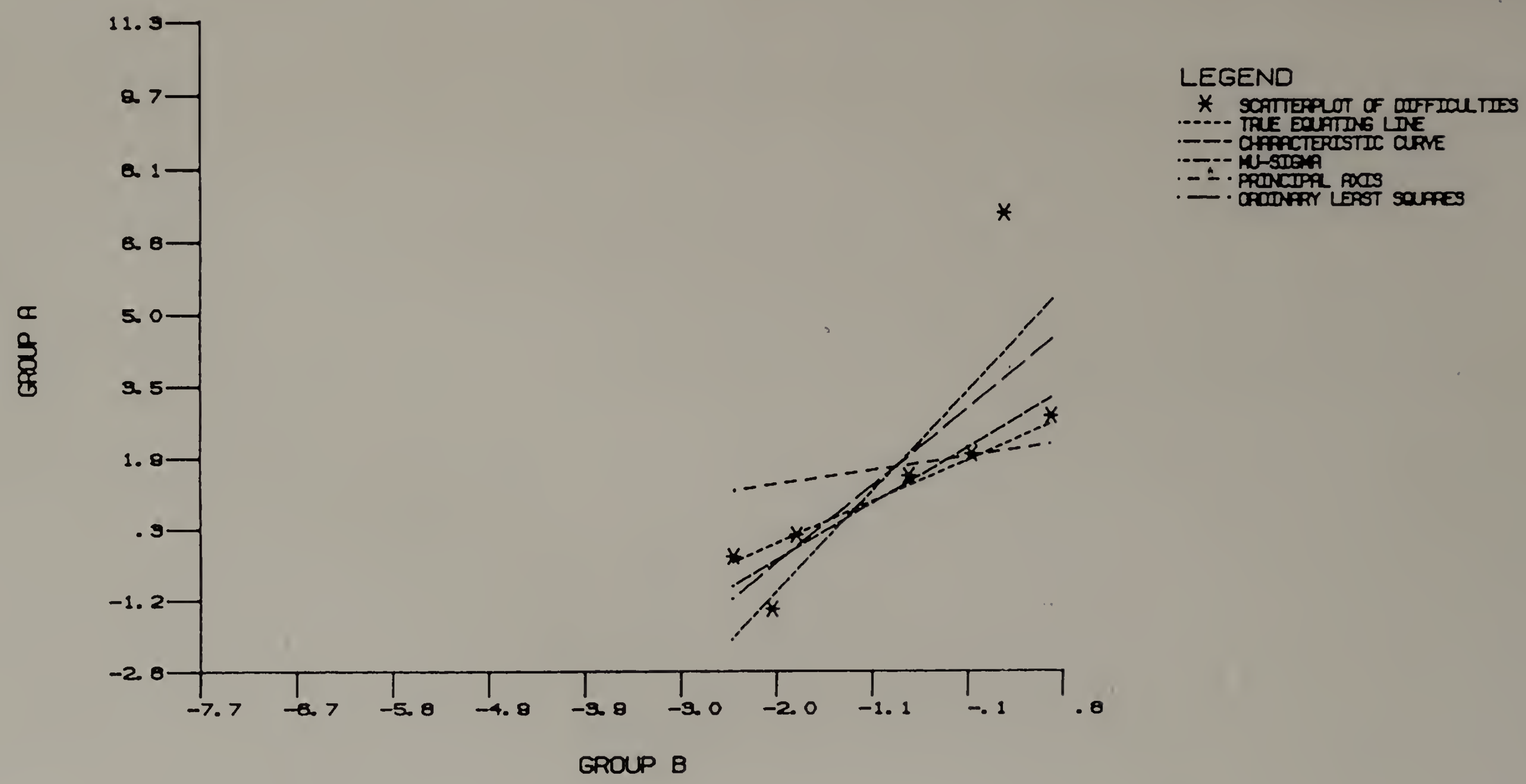


Figure A.7. Anchor difficulties with 7 items and a 30% overlap.

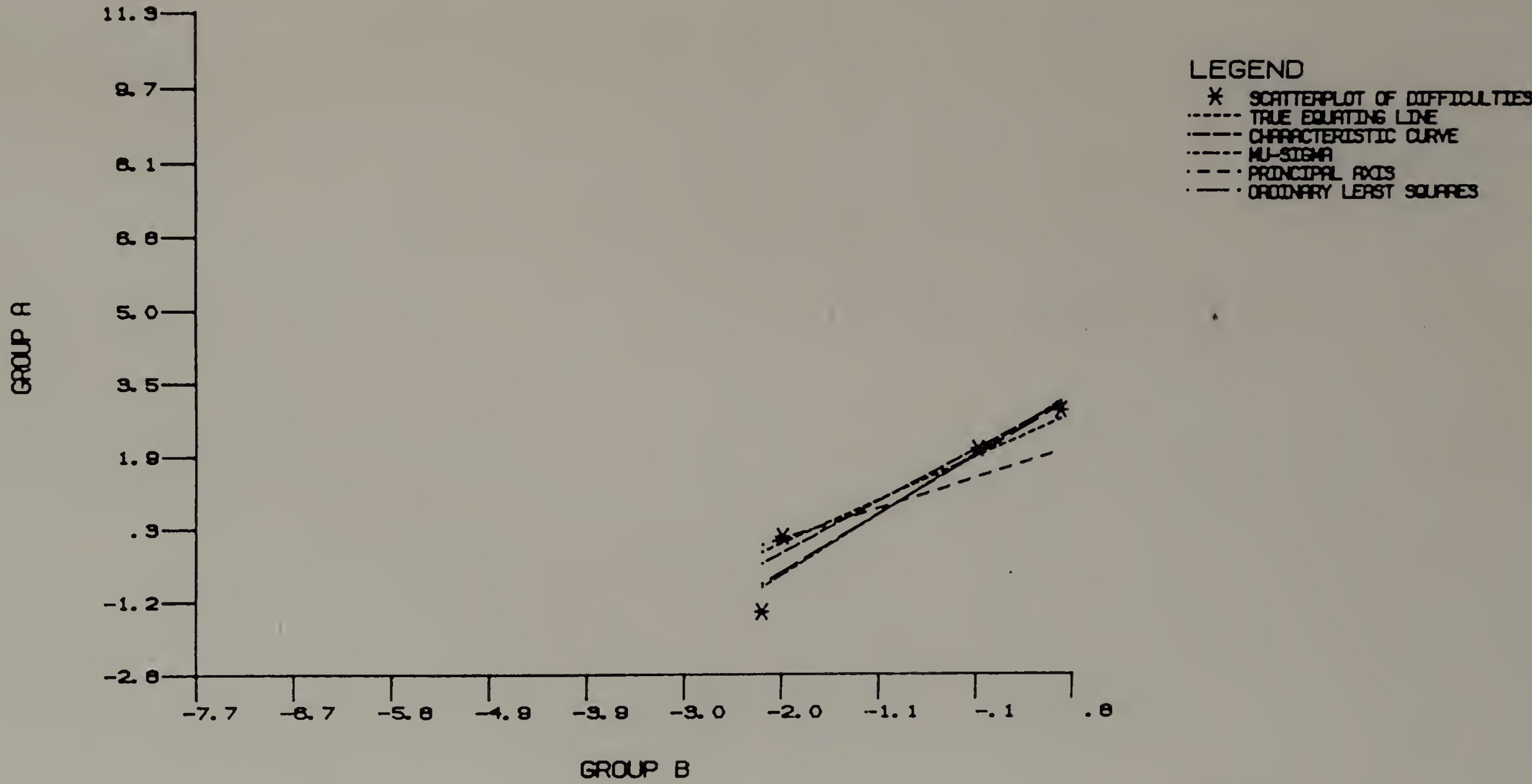


Figure A.8. Anchor difficulties with 4 items and a 30% overlap.

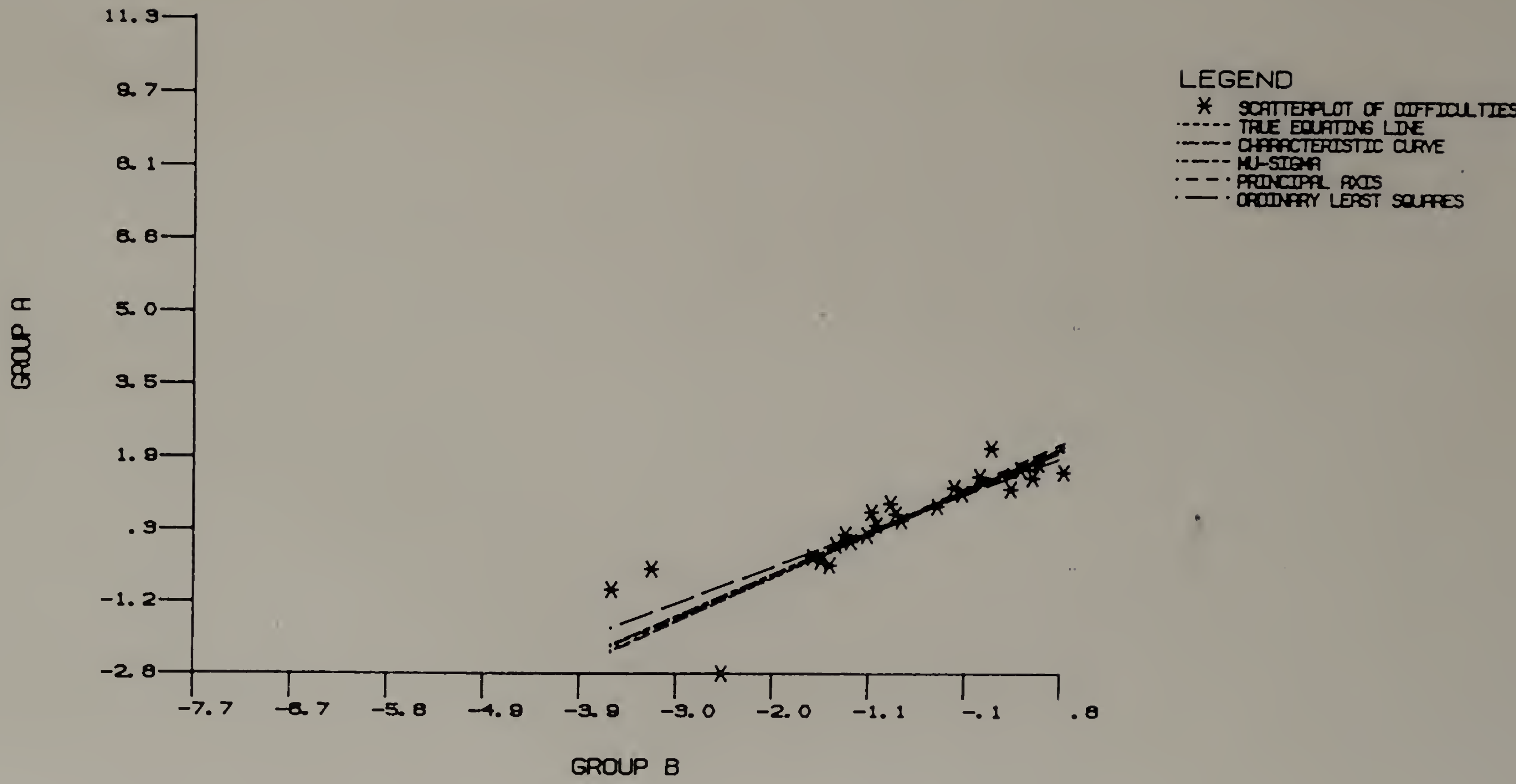


Figure A.9. Anchor difficulties with 25 items and a 50% overlap.

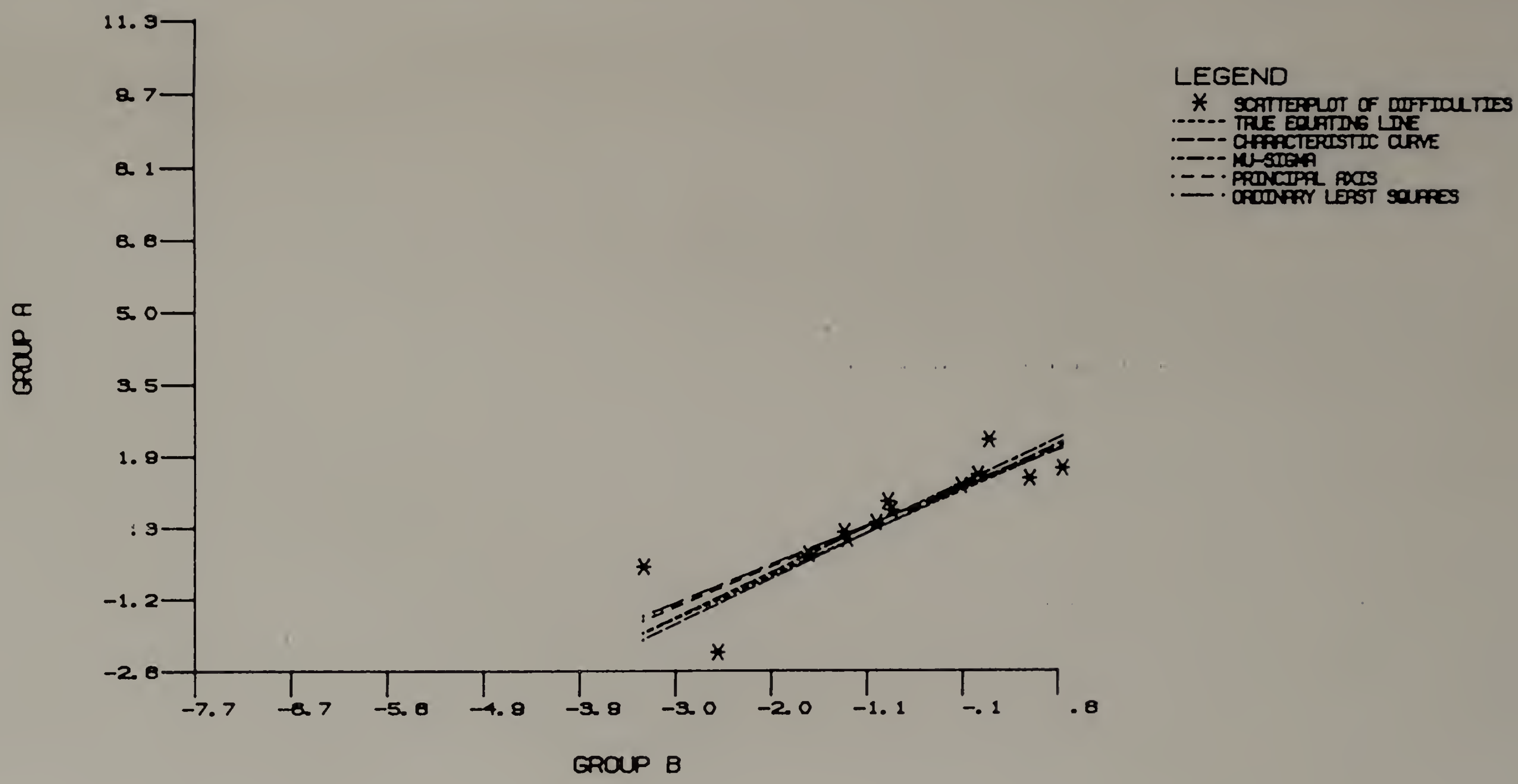


Figure A.10. Anchor difficulties with 13 items and a 50% overlap.

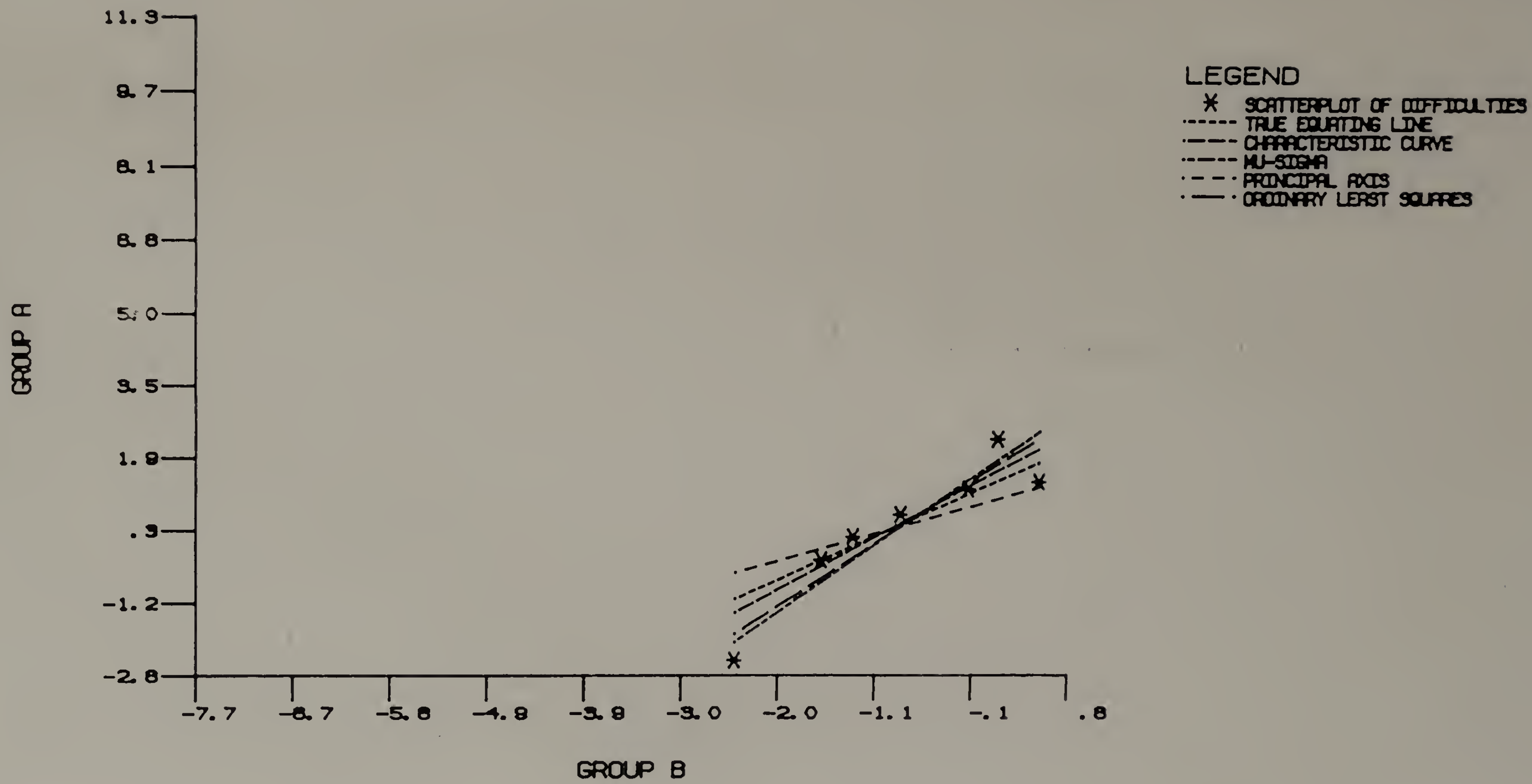


Figure A.11. Anchor difficulties with 7 items and a 50% overlap.

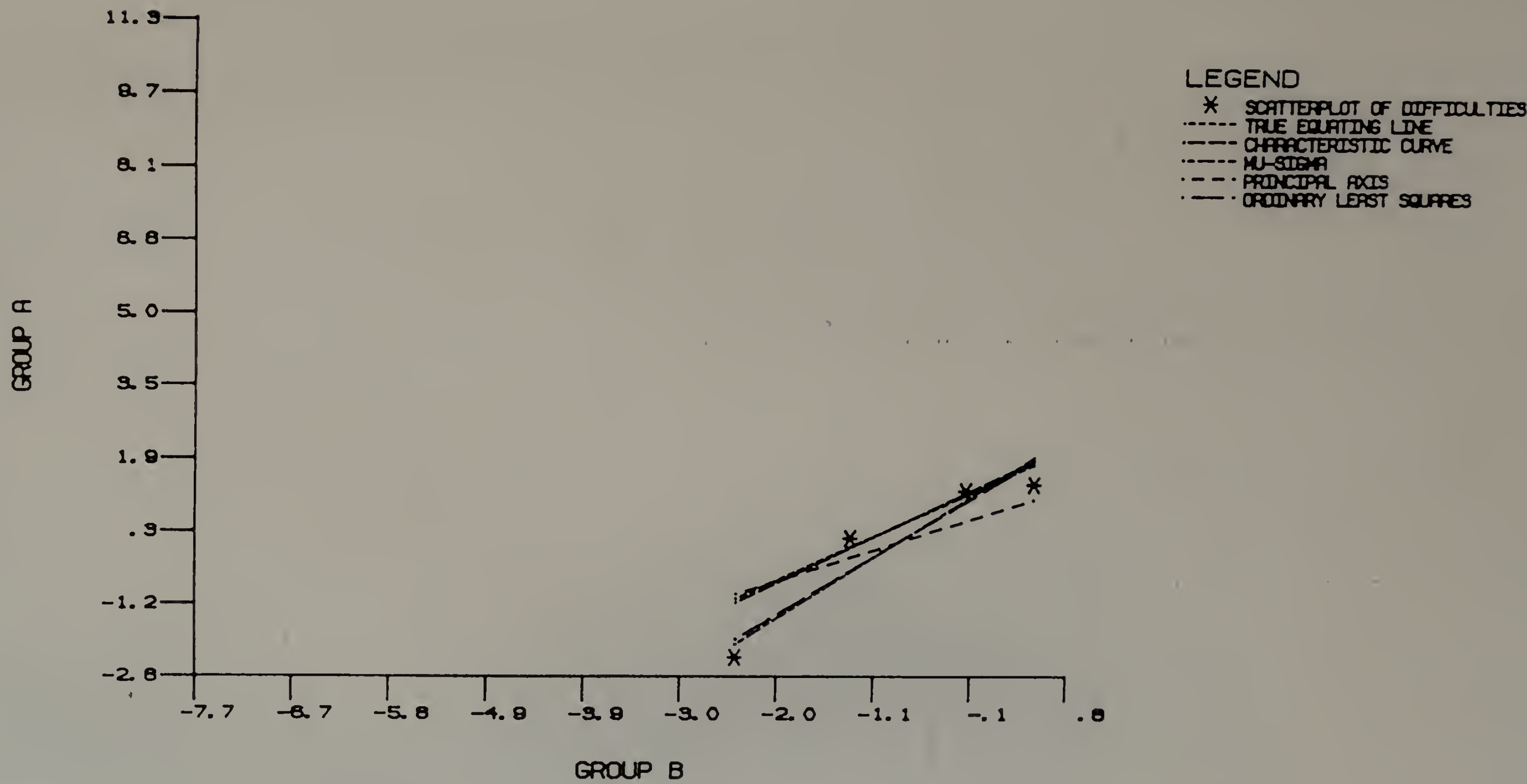


Figure A.12. Anchor difficulties with 4 items and a 50% overlap.

End of plot for 4399170

APPENDIX B

DATA GENERATION AND CHARACTERISTIC CURVE PROGRAMS

WMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMW
MWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWM

```
PROGRAM IRTDATA(INPUT/,OUTPUT,THETA5,TRUE5,RAW5,SCORE5,PARAM5);
(*$I'RANDOM' NUMBER GENERATOR DECLARATIONS*)

CONST

      A = 25;     (*ANCHOR*)
      B = 2;      (*TEST/GROUP*)
      N = 60;     (*ITEMS W/O ANCHOR*)
      R = 3;      (*ITEM PARAMETERS*)
      NA = 85;    (*ITEMS PLUS ANCHOR*)
      PI = 3.14159;
      M = 500;

TYPE

     IDXN = 1..N;
     IDXR = 1..R;
     IDXB = 1..B;
     IDXM = 1..M;
     IDXNA = 1..NA;
     TESTPARAM = ARRAY[IDXN] OF REAL;
     THETAS = ARRAY[IDXM] OF REAL;
     XTHETAS = ARRAY[IDXM,IDXB] OF REAL;
     MATRIX = ARRAY[IDXN,IDXR] OF REAL;
     XMATRIX = ARRAY[IDXN,IDXR,IDXB] OF REAL;
     AMATRIX = ARRAY[IDXNA,IDXR] OF REAL;
     XAMATRIX = ARRAY[IDXNA,IDXR,IDXB] OF REAL;
     PARAM = ARRAY[IDXB] OF REAL;

VAR

     E: IDXB;
     I: IDXN;
     K: IDXM;
     J: IDXR;
     IA: IDXNA;

     SEED1, SEED2: INTEGER;
     MINA,MAXA,MINB,MAXB,MINC,MAXC,MTHETA,SDTHETA: REAL;
     XMINA,XMAXA,XMINB,XMAXB: PARAM;
     XMINC,XMAXC,XMTHETA,XSDTHETA: PARAM;

     AI, BI, CI: TESTPARAM;
     T: THETAS;
     XT: XTHETAS;
     Q: MATRIX;
     QA: AMATRIX;
     XQ: XMATRIX;
     XQA: XAMATRIX;
     SCORE5,THETA5,RAW5,TRUE5,PARAM5: TEXT;
```

```
PROCEDURE CREATEP(MIN,MAX: REAL;
                  VAR  P:TESTPARAM);

VAR
     SPAN, DELTA: REAL;
     IDX: IDXN;

BEGIN
   WRITELN('** MAX/MIN',MAX,MIN);
     SPAN := MAX - MIN;
     DELTA := SPAN/(N-1);
     FOR IDX:=1 TO N DO
          P[IDX] := (IDX -1)*DELTA + MIN;
END;



PROCEDURE DOITEMS1(VAR  PA,PB,PC:TESTPARAM;
                   VAR  XQUES:XMATRIX);

VAR
     S: IDXN;
     T: IDXR;
     U: REAL;
     SU: 1..N;

BEGIN
     FOR S:=1 TO N DO
        FOR T:=1 TO R DO
           BEGIN
              IF T=1
                 THEN
                    BEGIN
                       IF S<31
                          THEN
                             XQUES[S,T,1] := PA[2*S - 1]
                          ELSE
                             XQUES[S,T,1] := PA[122 - (2*S)];
                    END;
              IF T=2
                 THEN
                    XQUES[S,T,1] := PB[S];
              IF T=3
                 THEN
                    BEGIN
                       U := RANDOM;
                       SU := TRUNC(U*60.0) +1;
                       XQUES[S,T,1] := PC[SU];
                    END;
           END;
END;
```

```
PROCEDURE DOITEMS2(VAR  PA,PB,PC:TESTPARAM;
                   VAR  XQUES:XMATRIX);

VAR
     S: IDXN;
     T: IDXR;
     U: REAL;
     SU: 1..N;

BEGIN
     FOR S:=1 TO N DO
        FOR T:=1 TO R DO
           BEGIN
              IF T=1
                 THEN
                    BEGIN
                       IF S<31
                          THEN
                             XQUES[S,T,2] := PA[2*S - 1]
                          ELSE
                             XQUES[S,T,2] := PA[122 - (2*S)];
                    END;
              IF T=2
                 THEN
                    XQUES[S,T,2] := PB[S];
              IF T=3
                 THEN
                    BEGIN
                       U := RANDOM;
                       SU := TRUNC(U*60.0) + 1;
                       XQUES[S,T,2] := PC[SU];
                    END;
           END;
END;
```

```
PROCEDURE ANCHOR(VAR  XQUES: XMATRIX;
                 VAR  XQUESA: XAMATRIX);


VAR
     S: IDXN;
     T: IDXR;
     F: IDXB;
     SA: IDXNA;
     W,Z: INTEGER;

BEGIN
   FOR F:=1 TO B DO
      BEGIN
         FOR S:=1 TO N DO
            BEGIN
               FOR T:=1 TO R DO
                  BEGIN
                     SA := S;
                     XQUESA[SA,T,F] := XQUES[S,T,F];
                  END;
            END;
         SA := N;
         FOR Z:=1 TO A DO
            BEGIN
               SA := SA + 1;
               IF Z<5
                  THEN
                     W := 8*Z - 7
                  ELSE
                     IF Z<8
                        THEN
                           W := 8*Z -35
                        ELSE
                           IF Z<14
                              THEN
                               W := 4*Z - 29
                              ELSE
                               W := 2*Z - 26;
               IF W<13
                  THEN
                     FOR T:=1 TO R DO
                        XQUESA[SA,T,F] := XQUES[5*W-4,T,1]
                     ELSE
                        BEGIN
                           FOR T:= 1 TO R DO
                              BEGIN
                               IF W<25
                                THEN
                                 XQUESA[SA,T,F] := XQUES[5*W-64,T,2]
                                ELSE
                                 XQUESA[SA,T,F] := XQUES[60,T,2];
                              END;
                        END;
            END;
      END;
END;
```

```
PROCEDURE PRINTPARAMS(VAR XQUESA: XAMATRIX);

VAR
     SA: IDXNA;
     T: IDXR;
     F: IDXB;

BEGIN
     FOR F:= 1 TO B DO
          FOR SA:=1 TO NA DO
               FOR T:=1 TO R DO
                  BEGIN
                     WRITELN('TEST',F:2,'ITEM':12,SA:3,'PARAMETER':16,T:2,
                             'EQUALS':14,XQUESA[SA,T,F]);
                     WRITELN(PARAM5,F,SA,T,XQUESA[SA,T,F]);
                  END;
END;



PROCEDURE DOTHETA5(VAR MT,SDT:REAL;
                   VAR  XPT:XTHETAS);

VAR
     Q: IDXM;
     X, Z: THETAS;
     H: 1..50;
     SD,SUM,MEAN,S,U,V: REAL;

BEGIN
     FOR Q:=1 TO M DO
        BEGIN
           V := 0;
           FOR H:=1 TO 50 DO
              BEGIN
                 U := RANDOM;
                 V := V+U;
              END;
           X[Q] := V/50;
        END;
     BEGIN
        S := 0;
        FOR Q:=1 TO M DO
           S := S + X[Q];
    END;
     BEGIN
        SUM := 0;
        MEAN := S/M;
        FOR Q:= 1 TO M DO
           SUM := SUM + SQR(X[Q] - MEAN);
     END;
     BEGIN
        SD := SQRT(SUM/M);
        FOR Q:=1 TO M DO
           Z[Q] := (X[Q] - MEAN)/SD;
     END;
     BEGIN
        FOR Q:=1 TO M DO
           XPT[Q,1] := Z[Q]*SDT + MT;
```

```
PROCEDURE DOTHETA2(VAR MT,SDT:REAL;
                   VAR  XPT:XTHETAS);

VAR
     Q: IDXM;
     H: 1..50;
     X, Z: THETAS;
     SD,SUM,MEAN,S,U,V: REAL;

BEGIN
     FOR Q:=1 TO M DO
        BEGIN
           V := 0;
           FOR H:=1 TO 50 DO
              BEGIN
                 U := RANDOM;
                 V := V+U;
              END;
           X[Q] := V/50;
        END;
     BEGIN
        S := 0;
        FOR Q:=1 TO M DO
           S := S + X[Q];
    END;
     BEGIN
        SUM := 0;
        MEAN := S/M;
        FOR Q:= 1 TO M DO
           SUM := SUM + SQR(X[Q] - MEAN);
     END;
     BEGIN
        SD := SQRT(SUM/M);
        FOR Q:=1 TO M DO
           Z[Q] := (X[Q] - MEAN)/SD;
     END;
     BEGIN
        FOR Q:=1 TO M DO
           XPT[Q,2] := Z[Q]*SDT + MT;
     END;
END;
```

```
PROCEDURE PRINTTHETAS(VAR XPT: XTHETAS);
VAR
     Q: IDXM;
     IDX,K: INTEGER;
     ABC: REAL;
     F: IDXB;

BEGIN
     FOR F:=1 TO B DO
        BEGIN
           WRITELN;
           WRITELN('THE ABILITY DISTRIBUTION FOR GROUP',F);
           FOR Q:=1 TO M DO
              WRITELN(THETA5,XPT[Q,F]);
           FOR IDX:= 0 TO N DO
              BEGIN
                 K := 0;
                 FOR Q:=1 TO M DO
                    BEGIN
                       ABC := 3-(0.1*IDX);
                       IF XPT[Q,F]>ABC
                          THEN
                             K := K+1;
                    END;
                 WRITELN(K,'INDIVIDUALS (G',F,') HAVE THETAS GREATER THAN',ABC);
              END;
        END;
END;
PROCEDURE LOGPROB(VAR XQUESA:XAMATRIX;
                  VAR  XPT:XTHETAS);
VAR
     SA: IDXNA;
     Q: IDXM;
     T: IDXR;
     D,P1,TS,U: REAL;
     F: IDXB;
     K: 0..1;
     RS: INTEGER;

BEGIN
     FOR F:=1 TO B DO
          FOR Q:=1 TO M DO
             BEGIN
                TS := 0;
                RS := 0;
                FOR SA:=1 TO NA DO
                   BEGIN
                      D := 1+EXP(-1.7*XQUESA[SA,1,F]*(XPT[Q,F]-XQUESA[SA,2,F]));
                      P1 := (1-XQUESA[SA,3,F])/D+XQUESA[SA,3,F];
                      U := RANDOM;
                      IF P1>=U
                         THEN
                            K := 1
                         ELSE
                            K := 0;
                      RS := RS+K;
                      TS := TS+P1;
                      WRITELN(SCORE5,K:2);
```

```
BEGIN  (*MAIN PROGRAM*)
     REWRITE(THETA5);
     REWRITE(TRUE5);
     REWRITE(RAW5);
     REWRITE(SCORE5);
     REWRITE(PARAM5);
     WRITELN('SEE DOCUMENTATION BEFORE USING IRTDATA');
     WRITELN;
     WRITELN('ENTER THE TWO INTEGRAL SEEDS');
     WRITELN;
     READLN;
     READ(SEED1,SEED2);
     SETRANDOM(SEED1,SEED2);
     FOR E:=1 TO B DO
          BEGIN
             WRITELN;
             WRITELN;
             WRITELN('ENTER THE PARAMETER CONSTRAINTS FOR ');
             WRITELN('GROUP/TEST',E,'IN THE FOLLOWING ORDER:');
             WRITELN('MINA,MAXA,MINB,MAXB,MINC,MAXC,MT,SDT.');
             READLN;
             READ(XMINA[E],XMAXA[E],XMINB[E],XMAXB[E],
                  XMINC[E],XMAXC[E],XMTHETA[E],XSDTHETA[E]);
          END;
          MINA := XMINA[1];
          MAXA := XMAXA[1];
          CREATEP(MINA,MAXA,AI);
          MINB := XMINB[1];
          MAXB := XMAXB[1];
          CREATEP(MINB,MAXB,BI);
          MINC := XMINC[1];
          MAXC := XMAXC[1];
          CREATEP(MINC,MAXC,CI);
          DOITEMS1(AI,BI,CI,XQ);
          MINA := XMINA[2];
          MAXA := XMAXA[2];
          CREATEP(MINA,MAXA,AI);
          MINB := XMINB[2];
          MAXB := XMAXB[2];
          CREATEP(MINB,MAXB,BI);
          MINC := XMINC[2];
          MAXC := XMAXC[2];
          CREATEP(MINC,MAXC,CI);
          DOITEMS2(AI,BI,CI,XQ);
          ANCHOR(XQ,XQA);
          PRINTPARAMS(XQA);
          MTHETA := XMTHETA[1];
          SDTHETA := XSDTHETA[1];
          DOTHETA5(MTHETA,SDTHETA,XT);
          MTHETA := XMTHETA[2];
          SDTHETA := XSDTHETA[2];
        DOTHETA2(MTHETA,SDTHETA,XT);
          PRINTTHETAS(XT);
          LOGPROB(XQA,XT);
END.
```

WMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMW
MWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWMWM

```
PROGRAM CCEQUAT(INPUT/,OUTPUT,LTH5013,LA5013A,LA5013B);

CONST

      NUMQ = 500;
      NUMN = 13;
      NUMM = 200;
      NUML = 20;
      NUMT = 2;

TYPE

     IDXQ = 1..NUMQ;
     IDXN = 1..NUMN;
     IDXM = 1..NUMM;
     IDXL = 1..NUML;
     IDXT = 1..NUMT;

     QTHETAS = ARRAY[IDXQ] OF REAL;
     THETAS = ARRAY[IDXM] OF REAL;
     PARAM = ARRAY[IDXN] OF REAL;
     MAT = ARRAY[IDXT,IDXT] OF REAL;
     VEC = ARRAY[IDXT] OF REAL;
     PARTS = ARRAY[IDXN,IDXM] OF REAL;


VAR

     I : IDXN;      (*INDEXES ITEMS*)
     J : IDXM;      (*INDEXES PERSONS*)
     Q : IDXQ;      (*INDEXES ORIGINAL THETAS*)
     L : IDXL;      (*INDEXES ITERATIONS*)

     ALPHA,BETA : REAL;    (*INITIAL ESTIMATE*)
     A1,B1,C1   : PARAM;   (*ITEM PARAMETERS, GROUP 1*)
     A2,B2,C2   : PARAM;   (*ITEM PARAMETERS, GROUP 2*)


     H, OUTH : MAT;
     T : THETAS;
     TA, TB : QTHETAS;

     DF : VEC;    (*PARTIAL DERIVATIVES OF F AT ALPHA, BETA*)

     X, S : VEC;

     AMAX,AMIN,BMAX,BMIN,SPAN,DELTA,MAX,MIN: REAL;
     A,B,F : REAL;

    LTH5013, LA5013A, LA5013B : TEXT;
```

```
PROCEDURE GRAD(PA, PB : REAL;
                 VAR PDF : VEC;
                 VAR FP : REAL);

VAR

     PI : IDXN;
     PJ : IDXM;
     SUM, ASUM, BSUM, PFA, PFB : REAL;
     T1, TS, SUMA, SUMB, PTA, PTB : THETAS;
     PSF : VEC;
     PAS, PBS, PC : PARAM;
     X1, X2S, P1, PS, PPA, PPB,  PT : PARTS;


BEGIN

   FOR PI := 1 TO NUMN DO
      BEGIN
         PAS[PI] := A2[PI]/PA;
         PBS[PI] := B2[PI]*PA + PB;
         PC[PI] := C2[PI];
      END;
   FOR PI := 1 TO NUMN DO
      FOR PJ := 1 TO NUMM DO
         BEGIN
            X1[PI,PJ] := (-1.7)*A1[PI]*(T[PJ] - B1[PI]);
            X2S[PI,PJ] := (-1.7)*PAS[PI]*(T[PJ] - PBS[PI]);
            P1[PI,PJ] := C1[PI] + ((1-C1[PI])/(1+EXP(X1[PI,PJ])));
            PS[PI,PJ] := C2[PI] + ((1-C2[PI])/(1+EXP(X2S[PI,PJ])));
         END;

   FOR PJ := 1 TO NUMM DO
      BEGIN
         SUM := 0;
         FOR PI := 1 TO NUMN DO
            SUM := SUM + P1[PI,PJ];
         T1[PJ] := SUM;
      END;
   FOR PJ := 1 TO NUMM DO
      BEGIN
         SUM := 0;
         FOR PI := 1 TO NUMN DO
            SUM := SUM + PS[PI,PJ];
         TS[PJ] := SUM;
      END;
```

```
   SUM := 0;
   FOR PJ := 1 TO NUMM DO
      SUM := SUM + SQR(T1[PJ] - TS[PJ]);

   FP := (1/NUMM)*SUM;
   WRITELN;
   WRITELN('THE FUNCTION F (TO BE MINIMIZED) :', FP);

   FOR PI := 1 TO NUMN DO
      FOR PJ := 1 TO NUMM DO
         BEGIN
            PPA[PI,PJ] := (1.7)*(T[PJ]-PBS[PI])*(1-PS[PI,PJ])*(PS[PI,PJ]-C2[PI])
                          /(1-C2[PI]);
            PPB[PI,PJ] := (-1.7)*(PAS[PI])*(1-PS[PI,PJ])*(PS[PI,PJ]-C2[PI])/(1-
                          C2[PI]);
            PT[PI,PJ] := B2[PI]*PPB[PI,PJ] - A2[PI]*PPA[PI,PJ]/SQR(PA);
         END;

   FOR PJ := 1 TO NUMM DO
      BEGIN
         SUMA[PJ] := 0;
         SUMB[PJ] := 0;
         FOR PI := 1 TO NUMN DO
            BEGIN
               SUMA[PJ] := SUMA[PJ] + PT[PI,PJ];
               SUMB[PJ] := SUMB[PJ] + PPB[PI,PJ];
            END;
         PTA[PJ] := SUMA[PJ];
         PTB[PJ] := SUMB[PJ];
      END;

   ASUM := 0;
   BSUM := 0;

   FOR PJ := 1 TO NUMM DO
      BEGIN
         ASUM := ASUM + (T1[PJ]-TS[PJ])*PTA[PJ];
         BSUM := BSUM + (T1[PJ]-TS[PJ])*PTB[PJ];
      END;
   PDF[1] := (-2/NUMM)*ASUM;
   PDF[2] := (-2/NUMM)*BSUM;
   WRITELN;
   WRITELN('THE PARTIAL DERIVATIVES OF F ARE', PDF[1], PDF[2]);
END;
```

```
PROCEDURE NEXTXH(PA,PB : REAL;
                 VAR PS : VEC;
                 VAR INH : MAT;
                 VAR PX : VEC;
                 VAR OUTH : MAT;
                 VAR PDF2 : VEC);
VAR

    Y,PY,PSIG,PDY,PDF : VEC;
    BP, AP , BNUM : MAT;
    SPYS,SPXS,PETA,ETA,FP,PAL,PW,PZ: REAL;
    FY,BD,FP2,BD1,BD2,ADENOM,R1,R2,R3,R4: REAL;

BEGIN

   GRAD(PA,PB,PDF,FP);
   SPXS := PDF[1]*PS[1] + PDF[2]*PS[2];
   PETA := (-2)*FP/SPXS;
   IF PETA < 1
      THEN
         ETA := PETA
      ELSE
         ETA := 1;
   PY[1] := PA + ETA*PS[1];
   PY[2] := PB + ETA*PS[2];

   GRAD(PY[1],PY[2],PDY,FY);

   SPYS := PDY[1]*PS[1] + PDY[2]*PS[2];
   PZ := (3/ETA)*(FP-FY) + SPXS + SPYS;
   PW := SQRT(SQR(PZ) - (SPXS*SPYS));
   PAL := ETA*(1-((SPYS+PW-PZ)/(SPYS-SPXS+(2.0)*PW)));

   PSIG[1] := PAL*PS[1];
   PSIG[2] := PAL*PS[2];

   PX[1] := PA + PSIG[1];
   PX[2] := PB + PSIG[2];

   GRAD(PX[1],PX[2],PDF2,FP2);
```

```
    Y[1] := PDF2[1]-PDF[1];
    Y[2] := PDF2[2]-PDF[2];
    ADENOM := PSIG[1]*Y[1] + PSIG[2]*Y[2];

    AP[1,1] := SQR(PSIG[1])/ADENOM;
    AP[1,2] := PSIG[1]*PSIG[2]/ADENOM;
    AP[2,1] := AP[1,2];
    AP[2,2] := SQR(PSIG[2])/ADENOM;

    R1 := INH[1,1]*Y[1] + INH[1,2]*Y[2];
    R2 := INH[1,1]*Y[1] + INH[2,1]*Y[2];
    R3 := INH[1,2]*Y[1] + INH[2,2]*Y[2];
    R4 := INH[2,1]*Y[1] + INH[2,2]*Y[2];

    BNUM[1,1] := (-1)*R1*R2;
    BNUM[1,2] := (-1)*R1*R3;
    BNUM[2,1] := (-1)*R4*R2;
    BNUM[2,2] := (-1)*R4*R3;

    BD1 := Y[1]*(INH[1,1]*Y[1]+INH[2,1]*Y[2]);
    BD2 := Y[2]*(INH[1,2]*Y[1]+INH[2,2]*Y[2]);

    BD := BD1 + BD2;

    BP[1,1] := BNUM[1,1]/BD;
    BP[1,2] := BNUM[1,2]/BD;
    BP[2,1] := BNUM[2,1]/BD;
    BP[2,2] := BNUM[2,2]/BD;

    OUTH[1,1] := INH[1,1] + AP[1,1] + BP[1,1];
    OUTH[1,2] := INH[1,2] + AP[1,2] + BP[1,2];
    OUTH[2,1] := INH[2,1] + AP[2,1] + BP[2,1];
    OUTH[2,2] := INH[2,2] + AP[2,2] + BP[2,2];

END;
```

```
BEGIN (*LORD*)

   RESET(LTH5013);
   RESET(LA5013A);
   RESET(LA5013B);
   WRITELN;
   WRITELN('ENTER THE FIRST ESTIMATE OF ALPHA AND BETA.');
   WRITELN;
   READLN;
   READ(ALPHA,BETA);

   FOR I := 1 TO NUMN DO
      BEGIN
         READLN(LA5013A,A1[I]);
         READLN(LA5013A,B1[I]);
         READLN(LA5013A,C1[I]);
         READLN(LA5013B,A2[I]);
         READLN(LA5013B,B2[I]);
         READLN(LA5013B,C2[I]);
      END;

   FOR Q := 1 TO NUMQ DO
      READLN(LTH5013,TA[Q]);
   FOR Q := 1 TO NUMQ DO
      READLN(LTH5013,TB[Q]);

   AMAX := 0;
   AMIN := 0;
   BMAX := 0;
   BMIN := 0;

   FOR Q := 1 TO NUMQ DO
      BEGIN
         IF TA[Q] > AMAX
            THEN
               AMAX := TA[Q];
         IF TA[Q] < AMIN
            THEN
               AMIN := TA[Q];
         IF TB[Q] > BMAX
            THEN
               BMAX := TB[Q];
         IF TB[Q] < BMIN
            THEN
               BMIN := TB[Q];
      END;
```

```
    IF AMAX>BMAX
       THEN
          MAX := AMAX
       ELSE
          MAX := BMAX;

    IF AMIN<BMIN
       THEN
          MIN := AMIN
       ELSE
          MIN := BMIN;

    SPAN := MAX - MIN;
    DELTA := SPAN/(NUMM - 1);

    FOR J := 1 TO NUMM DO
       T[J] := MIN + (J - 1)*DELTA;

    WRITELN('MIN/MAX THETA IS',T[1], T[NUMM]);
    WRITELN;

    A := ALPHA;
    B := BETA;

    H[1,1] := 1;
    H[1,2] := 1;
    H[2,1] := 1;
    H[2,2] := 1;

    GRAD(A,B,DF,F);

    FOR L := 1 TO NUML DO
       BEGIN
          S[1] := (-1)*(H[1,1]*DF[1]+H[1,2]*DF[2]);
          S[2] := (-1)*(H[2,1]*DF[1]+H[2,2]*DF[2]);

          NEXTXH(A,B,S,H,X,OUTH,DF);

          WRITELN;
          WRITELN('ITERATION ', L);
          WRITELN;
          WRITELN('*******', X[1],X[2],'*******');
          WRITELN;

          A := X[1];
          B := X[2];
          H[1,1] := OUTH[1,1];
          H[1,2] := OUTH[1,2];
          H[2,1] := OUTH[2,1];
          H[2,2] := OUTH[2,2];

       END;
END.
```

REFERENCES

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

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

Baker, F. B. (1984). Ability Metric Transformations involved in vertical equating under item response theory. Applied Psychological Measurement, 8, 261-271.

Baker, F. B. (1983). Comparison of ability metrics obtained under two latent trait theory procedures. Applied Psychological Measurement, 7, 97-110.

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

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

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

CTB/McGraw-Hill. (1982). Comprehensive tests of basic skills, preliminary technical report, forms U and V. Monterey, CA: Author.

De Gruijter, D. M. N. (1986). The use of item statistics in the calibration of an item bank. Applied Psychological Measurement, 10, 231-237.

Divigi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. Applied Psychological Measurement, 9, 413-415.

Divigi, D. R. (1980). Evaluation of scales for multilevel test batteries. Paper presented at the meeting of the American Educational Research Association, Boston. (Revised.)

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.

Drasgo, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Fletcher, R., & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. The Computer Journal, 6, 163-168.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. Japanese Psychological Research, 22, 144-149.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 48, 467-510.

Hambleton, R. K., & Van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. Applied Psychological Measurement, 6, 373-378.

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

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimension of a set of test items. Applied Psychological Measurement, 10, 287-302.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. Applied Psychological Measurement, 10, 35-43.

Hicks, M. M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. Applied Psychological Measurement, 7, 255-266.

Hopkins, K. D., Glass, G. V., & Hopkins, B. R. (1987). Basic Statistics for the Behavioral Sciences. Englewoods CLiffs, NJ: Prentice-Hall.

Ironson, G. (1982). Chi-square and item response theory techniques. In R. Berk (Ed.), Handbook of-methods for-detecting test bias. Baltimore, MD: John Hopkins University Press.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 147-154.

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., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. Psychometrika, 52, 43-59.

Kolen, M. J. (1985). Standard errors of tucker equating. Applied Psychological Measurement, 9, 209-223.

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

Lord, F. M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

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

Lord, F. M. (1982). Standard error of an equating by item response theory. Applied Psychological Measurement, 9, 413-415.

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

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

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

Phillips, S. E. (1983). Comparison of equipercentile and item response theory equating when the scaling test method is applied to a multilevel achievement battery. Applied Psychological Measurement, 7, 267-281.

Phillips, S. E. (1985). Quantifying equating errors with item response methods. Applied Psychological Measurement, 9, 59-[illegible]1.

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

Spence, I. (1983). Monte Carlo simulation studies. Applied Psychological Measurement, 7, 405-425.

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

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.

Vale, C. D. (1986). Linking item parameters onto a common scale. Applied Psychological Measurement, 10, 333-344.

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

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristics curve parametrs (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service.

Woodruff, D. (1986). Derivations of observed score linear equating methods based on test score models for the common item nonequivalent populations design. Journal of Educational Statistics, 11, 245-257.

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.