1-1-1987

The effect of range restriction on invariance in item response models.

Richard Francis Mooney
University of Massachusetts Amherst

c

THE EFFECT OF RANGE RESTRICTION ON INVARIANCE IN ITEM RESPONSE MODELS

A Dissertation Presented

By

RICHARD FRANCIS MOONEY

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

School of Education

THE EFFECT OF RANGE RESTRICTION ON INVARIANCE IN ITEM RESPONSE MODELS

A Dissertation Presented

By

RICHARD FRANCIS MOONEY

Approved as to style and content by:

Hariharan Swaminathan, Chairperson of Committee

Janice Gifford, Member

Stanley Scarpati, Member

Castellano Turner, Member

George Urch, Acting Dean
School of Education

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Dr. Hariharan Swaminathan, Chairperson of my committee. His guidance, support and patience have made this study possible.

I also wish to thank the members of my committee, Dr. Janice Gifford, Dr. Stanley Scarpati, and Dr. Castellano Turner, for their helpful comments and suggestions. Thanks to Dr. Ronald K. Hambleton for his guidance throughout my doctoral program.

I would also like to add a word of thanks to Robert Gonter, Trina Hosmer, Wayne Johnson, Eva Goldwater, Mary Cushing and Cliff Donath for their technical assistance with computer programming and data management.

I wish to thank Bernadette McDonald for her excellent assistance with manuscript preparation.

I also thank my wife, Dr. Sarah B. Kinder, for her support and encouragement through all stages of this project.

ABSTRACT

The Effect of Range Restriction on Invariance
in Item Response Models
(September, 1987)
Richard Francis Mooney, B.A. Oxford University
M.A. Oxford University, Ed.D. University of Massachusetts
Directed by: Professor Hariharan Swaminathan
Item parameter invariance is a key property of IRT models, and it is a property that sets IRT apart from classical test theory models. Item parameter invariance is important for a number of testing issues, but one of the most direct and straightforward examples of the use of this property arises in the study of item bias. Here, the estimates from different groups are obtained and then compared to determine if individual items behave differently for different groups.

A question that naturally arises in this application is the degree to which parameter invariance holds for different subgroups with different sample sizes and different ability distributions when bias does not exist.

To answer this question, simulated data for three levels of ability and three levels of sample size were generated to yield nine testing situations. Thirty random samples of data from each testing situation were fitted to the three parameter item response model using

sampling with replacement. The difficulty parameter estimates were compared for stability and accuracy of estimation.

The results of the study show that while stability was obtained, accuracy for extreme items was influenced by restriction in the range of ability of the group of examinees. Further, it was shown that the three parameter model appeared to obtain a better fit when a positively skewed distribution of ability was used. Overall, the model generally performed well with items that have difficulty parameters in the middle range of difficulty. Increases in sample size did not generally improve the quality of estimation, although the influence of restriction of ability range persisted and maintained similar patterns even for the largest sample size (n=1,200).

The sampling with replacement technique was seen to be a useful method for examining the sampling error of item parameter estimates. This method may prove useful in the context of determining model data fit or other item response theory applications that depend on the property of parameter invariance.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

Item response theory (IRT) is a measurement theory based on the assumption that examinee test performance for a given item can be explained as a function of underlying examinee traits as well as the particular characteristics of the item. By making assumptions about the form of this relationship and about the dimensionality of the latent space (the number of traits necessary for describing the response of an examinee) inferences can be made about the unobservable traits based on observable test scores.

The relationship between observed scores and unobservable traits is specified through a monotonically increasing mathematical function known as an item characteristic function. In cases where the latent space measures a single underlying trait, the item characteristic function is known as an item characteristic curve (ICC). Currently, only unidimensional models are available for practical application, although a broad range of models both unidimensional and multidimensional, linear and non-linear, are feasible (McDonald, 1982). Typically, the item characteristic curve is taken as the logistic curve, although the less mathematically tractable normal ogive may also be used.

Model-Data Fit

To completely specify the relationship between the probability of a correct response and underlying ability, an item response model that relates the probability to the parameters that characterize an item is needed. If the model does not fit the data then advantages of IRT may not be realized. Three unidimensional item response models are currently available to practitioners working with dichotomously scored items. All three of these models assume that the examinee's response to a given item is completely described as a single or unidimensional ability factor. These unidimensional models are the one-parameter or Rasch model, the two-parameter model, and the three-parameter model. The one-parameter model assumes that items are characterized by one parameter, item difficulty, while the two parameter model assumes that the items are characterized by two parameters, item difficulty and item discrimination. The three-parameter model, the most general of the unidimensional IRT models currently in wide use, assumes that the items are characterized by a guessing parameter as well as item difficulty and item discrimination (Hambleton & Swaminathan, 1985).

Advantages of IRT

The item response function is essentially the regression of item score on ability. Regression functions remain the same in spite of changes in a frequency distribution of the predictor variable. This implies that the parameters that characterize the regression function are invariant. Since the item parameter in IRT describe the regression function, they are therefore invariant across ability distributions.

Advantages of IRT include examinee ability parameters that are independent of the particular set of items administered, and item parameters that are invariant across subgroups of examinees. These features offer potential for solving several important testing problems that were not solvable using classical testing models based on linear characterizations of human testing behavior.

Among the testing problems that may be solved using IRT are: item banking, tailored/adaptive testing, equating test scores and identification of item bias. These applications depend upon the property of invarance of item parameters.

Applications of the Property of Invariance in IRT

Four important areas in which the property of invariance plays a central role are item banking, tailored/adaptive testing, test equating and the study of item bias. These applications are briefly reviewed in the following section.

Item Banking. An item bank is a large pool of pilot tested items that are categorized by objectives or skills. These banks may then be used to build a test to meet particular needs quickly and efficiently.

Item banks constructed using classical item parameters are not optimal in that classical item parameters such as item difficulty and item discrimination are sample dependent. IRT, however, offers a potentially useful theoretical framework for developing items for item banks because of the expected feature of invariant item parameters.

Such invariant item parameters greatly simplify the task of building and using item banks.

Tailored/Adaptive Testing. Tailored testing is another important application of item response theory. The invariance of item and ability parameters permits the "tailoring" of tests to fit particular needs. In the context of norm referenced tests, test builders typically choose items that have a classical item difficulty index of a .50 probability of answering the average item correctly. This also means that examinees of extreme ability obtain more poorly estimated scores as compared to examinees in the middle range of ability.

An ideal solution to this problem is to administer items that correspond to the ability level of an examinee so that the ability of each examinee can be estimated accurately. Using IRT, it is possible to accomplish this goal. The ability of each examinee can be determined from items that are "tailored" to an examinee. Moreover, the invariance property permits the comparison of examinees.

Adaptive testing is a dynamic form of tailored testing. Here, the examinee has an interactive relationship with an item bank, and items are selected for presentation based on the performance of the examinee. Such a strategy offers promise for obtaining high quality estimates of examinee ability, particularly for examinees in the extreme ranges of ability. It has been demonstrated that by using an adaptive testing strategy, test taking time can be considerably reduced.

Test Equating. Test equating is important for comparisons of examinee's performance on non-identical tests. Equating tests at the same ability level is known as horizontal equating, while equating tests over different levels of ability is known as vertical equating. Vertical equating may be used, for example, in comparing children across different school grades. In both cases, the items have to be placed on a single scale. Once again, the invariant item parameters of IRT provide a useful framework for this challenging testing problem.

Item Bias. Item bias exists when groups of examinees of equal ability have an unequal probability of getting correct responses to an item. One way to approach the item bias problem is to compare the item difficulty parameters of a given item across the groups of interest. Lord (1980) has argued that classical item difficulty statistics are not appropriate for the study of item bias because such item statistics are sample dependent. Item response theory, however, offers a better mechanism for testing bias because of the property of invariant item parameters. According to Ironson (1983) "...IRT is less likely (than methods based on classical item statistics) to artificially label an item as biased. Classical measures confound ability differences with differences in discrimination, difficulty and guessing" (p. 55).

Statement of the Problem

The invariance of item parameters is important in the field of testing. Through the expected feature of item parameter invariance, IRT provides a sound theoretical basis for exploring the issue of item bias detection. Although different sub-groups may have different ability distributions they should nevertheless demonstrate equal item parameter estimates when sampling fluctuations are taken into account. When parameter estimates differ, one interpretation would be that the items are behaving differently for the two groups. This, in turn, implies that the item is biased.

One concern, however, is that the expected feature of invariance may be confounded with such estimation issues as sampling error and range restriction. Sampling error describes the differences among parameter estimates with randomly equivalent samples of examinees. Differences between the estimates would be expected to decrease as sample size increases.

Range restriction refers to constriction in the distribution of ability in a particular sample of examinees used to estimate the parameters. For example, a given sample of examinees may be homogenious and have a relatively narrow range of ability. When this happens, the regression function has to be estimated from a set of points that cluster tightly. This results in the regression function being estimated poorly. Small changes in the placement of points may result in dramatically different regression functions and consequently in parameters that are unstable.

The expected property of invariance of item parameters plays a major role in detecting item bias. The comparison of parameter estimates obtained from sub-groups of interest using IRT models has been advocated as a method of detecting item bias. One issue with this approach is that although it is known that range restriction may influence parameter estimation, it is not known precisely what the impact would be in the case of comparing extreme groups for the purpose of item bias examination.

The purpose of this dissertation is to study the effects of the above mentioned factors on the invariance of item parameter estimates where groups are known to have similar characteristics. Questions of interest in this study are:

1) How does range of ability affect the invariance of the estimates of the difficulty parameters in the three parameter IRT model?

2) What is the influence of sample size in the invariance of the estimates of the difficulty parameter?

3) What is the consequence of interaction of range of ability with sample size?

This study assesses the variability of the item difficulty or b parameter estimates of the three parameter IRT model by obtaining parameter estimates for the same items over repeated samples with very similar characteristics. The strategy for this dissertation was to evaluate the extent to which repeated estimates obtained from samples with differing ability distributions and sample sizes would recover the true values for these parameters.

To investigate these questions, three levels of range restriction and three levels of sample size were generated to yield nine testing situations. Thirty random samples from each testing situation were fitted to the three-parameter item response model and compared. If the invariance property holds, parameter estimates should be consistantly homogenious across the full range of items and conditions. The hypothesis was that estimation would not be influenced by changes in the ability distribution because of the invariance property.

Simulated data were used for this study primarily because population parameters could be known. A second advantage of simulated data is as a control for model-data fit and also for bias. Although model-data fit or lack of item bias cannot be established even with simulated data, this approach provides a reasonable intuitive basis for this.

One way to obtain repeated samples is to artificially generate responses for each examinee. This approach was taken by Gifford and Swaminathan (in press). While this is a useful approach for understanding the properties of the estimates, it is not a feasible approach in a practical testing situation. In this approach, samples are drawn, with replacement; for each sample, the item and ability parameters are estimated; and the sampling distribution of these estimates established empirically. The method of resampling from the same set of data has some clear advantages since we do not know theoretically the sampling error of the estimates. These include avoiding the need for collecting more data, while allowing for the possiblity of studying the sampling properties of the estimates.

One contribution of this dissertation is that it provides an empirical understanding of the nature of sampling error in IRT. In particular, the effects of range restriction and sample size on parameter invariance can be investigated.

A further contribution may be in providing a method for determing the standard error of estimate in IRT. Currently, the theoretically derived standard error of estimate is used to understand the sampling fluctuations of the estimates. These standard errors may not be accurate enough for the sample sizes used in practical applications. The resampling method used in this dissertation provides another method of assessing the standard error.

Another contribution is in the assessment of item bias. One method of assessing item bias is to first obtain parameter estimates for groups where bias may be a concern. The parameter estimates, typically the b's, may then be compared using scatterplots. For example, in an examination of possible sex bias, each sex group may be randomly divided into two groups. Parameter estimates may then be obtained for all four groups. If bias does not exist, it would be expected that within group scatterplots would demonstrate about the same degree of scatter as between group scatterplots. This method is advocated by Hambleton and Murray (1983), and will be discussed in Chapter III. The repeated sampling method proposed in this study may provide a clearer picture of bias than would be possible with only two replications for each group.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

IRT is best understood in terms of its historical relationship to classical test theory. Classical test theory predates IRT and is a useful, relatively simple and flexible model that has application for a wide range of testing needs. However, due to a number of limitations of the classical test theory model for solving sophisticated testing issues, and also because of the availability of modern high speed computers, IRT has come to be the test theory model of choice.

This chapter will begin with a review of classical test theory, including a discussion of shortcomings of this model that have led to the use of IRT. Next, IRT will be considered, particularly in relation to the key property of parameter invariance. The method of detecting item bias using IRT estimates obtained from extreme groups will be considered in terms of its potential for investigating item parameter invariance. Finally, a preliminary study of item parameter invariance using repeated samplings will be reviewed.

Review of Classical Test Theory Assumptions

The classical model defines two unobservable scores called true score and error score. This concept is based on the theoretical idea of infinitely replicated testings. For a given examinee, true score is the expected value of the observed scores, while error score is the expected difference between true score and observed score. This model may be written:

$$x = T + e$$

where: x = observed score

T = True score

e = error score

Assumptions for this model are (1) the mean of the error term is zero, (2) the correlation between true score and error score is zero and (3) error terms are uncorrelated over repeated testings on parallel forms. These assumptions describe the conceptual partitioning of the inconsistent performance modeled in the error term from elements that describe consistent performance called true score.

Although several important and useful formulas are derived from the classical test model including the Spearman-Brown formula and others, there are also important limitations to the model. The chief limitation is that classical item parameters measuring item difficulty (p value or proportion correct) and item discrimination (item total correlations) are influenced by examinee characteristics. Lord (1980) says "Proportion of correct answer in a group of examinees is not really a measure of item difficulty. This proportion describes not only the test item, but also the group tested... Item test correlations vary from group to group also. Like other correlations, item-test correlations tend to be high in groups that have a wide range of talent, low in groups that are homogeneous." (P35) Sample dependent item statistics limit the generalizability of test validity to examinee samples that are nearly identical to the sample that is used for item calibration (Hambleton & Swaminathan, 1985).

A related problem is that choice of item is confounded with test reliability. Reliability is enhanced by test variance. One important implication from this is that tests are constructed to maximize observed score variance. The contribution of each item to the test variance cannot be determined precisely. Hence it may not be possible, using classical test theory, to choose items that maximize reliability of the test.

The issue of group dependent item parameters also has implications for the development of parallel forms. Although the notion of the parallel form test is a cornerstone of classical test theory, the parallel form is difficult to realize in practice (Hambleton & Swaminathan, 1985). However, parallel forms are necessary for comparisons of true scores across examinees.

Item Response Theory Assumptions

IRT is based on strong assumptions, while classical test theory is based on weak assumptions. The classical model is flexible because of these weak assumptions and it is very likely to fit nearly all mental measurement test data sets. One problem with classical test theory, however, is that there are inherent limitations with its applications.

IRT models may be less flexible than the classical model as well as more mathematically complex, but when the IRT model fits the data, considerable benefits are realized. While classical test theory models are limited to the first and second moments, item response theory sustains models that support linear and non-linear regression and normal and non-normal frequency distributions (Lord, 1980).

The incorporation of non-linear relationships or equivalently that of higher order moments in item response theory is the key to the added theoretical advantages of IRT over classical test theory. The price to be paid for these advantages include increased stringency of model assumptions, particularly those of local independence and unidimensionality.

Local Independence

Item response theory specifies a probabilistic relationship between examinee test performance and a set of unknown latent traits. A basic assumption in IRT is that the underlying latent space is complete.

When the complete latent space of dimension n is specified, then all the traits T_1, T_2, T_3,...T_n have been taken into account in defining the relationship between examinee response and the individual item characteristics for a given item. This implies that the examinees' responses to items i and j are statistically independent when T_1, $T_2, \ldots T_n$ are given, i.e.,

$$f(y_i, y_j | T_1, T_2, \ldots T_n) = f(y_i | T_1, T_2, \ldots T_n)\ f(y_j | T_1, T_2, \ldots T_n)$$

Local independence is a strong assumption in IRT, and one that is easily violated (Goldstein, 1980). Another way to state the assumption of local independence is that the error terms of the item response models for individual respondents at given levels of T_1, $T_2, \ldots T_n$, should be independent. Violations of local independence would be anticipated in circumstances where a response to one item would influence the examinee's response to another item. This situation may occur in a reading test, for example, when several questions are asked about a single passage.

Unidimensionality

A common assumption in the application of item response theory is that the complete latent space is unidimensional. McDonald (1982) argues that the concept of unidimensionality should flow directly from the concept of local independence.

When unidimensionality does not exist for a given data set, then it is a tautology that a unidimensional model will not provide the best fit. Furthermore, the extent of model robustness is not known, so it cannot be determined to what degree expected features may or may not be obtained given some degree of model data misfit.

The issue of dimensionality is a difficult matter. It opens the possibility of a number of potential explanations of model data fit problems, as well as concerns about the confounding of model data fit problems with other issues such as sampling error, or item bias and so on. Dimensionality is a haunting problem for IRT, precisely because it is elusive and at the same time, central to the expected features that make IRT attractive to measurement specialists.

Mathematical Form of IRT Models

It should be noted that item response models (IRM) are part of a large family of models, including both multidimensional and unidimensional models as well as models that are fully or partly linear or non-linear (McDonald, 1982). Non-linear models are convenient to work with because the eliminate the problem of a probability scale that is not bounded by 0 and 1. Multidimensional models are too complex for practical application at this time.

One parameter model. In the one parameter model, the probability of a correct response may be written:

$$P_{1j}\ (T_i) = \exp D(T_i - b_j)/[1 + \exp D\ (T_i - b_j)]$$

where the correct response for individual i with ability T_i for item j is denoted $P_j\ (T_i)$ and the item difficulty parameter is denoted b_j. The b_j parameter is a location parameter on the ability scale that corresponds to a probability of .5 correctly responding to the item. As items increase in difficulty the curve moves to the right on the ability scale. The scaling factor, D, set at 1.7, is used to maximize correspondence between the normal ogive and the logistic function.

<u>Two parameter model</u>. The two parameter model is appropriate when items vary in difficulty and discrimination. For the two parameter model, the probability of a correct response is given by:

$$P_{2j}\ (T_i) = \exp Da_j\ (T_i - b_j)/[1 + \exp Da_j\ (T_i - b_j)]$$

where a_j is the item discrimination parameter and is the only addition to the previously shown 1-parameter model. This "a" parameter is proportional to the slope at the inflection point (Lord, 1980).

Three parameter model. The probability of a correct response for the three parameter model is given by:

$$P_{3j}\ (T_i) = C_j + (1 - C_j)\{\exp Da_j(T_i - b_j)/ \\ [1 + \exp Da_j(T_i - b_j)]\}$$

where C_j is the guessing parameter. The C parameter corresponds to the lower asymptote. This parameter represents the probability of a randomly selected examinee responding correctly by guessing. This probability is zero for the one- and the two-parameter models.

The guessing parameter is often called the pseudo-guessing parameter or pseudo-chance parameter at the suggestion of Lord (1974) because the estimated chance level is typically below the expected probability for guessing for field data estimates. Lord attributes this to the skill of item writers at providing answer stems that are attractive to examinees who lack sufficient knowledge or technique to answer the question appropriately.

Invariance in Item Response Models

Two key properties of item response models are item and ability parameter invariance. These features are a direct consequence of the assumption that an examinees' ability and the probability of a correct response to an item is related by the item response function. Lord (1980, pp. 34) describes the invariance property as follows:

> "...an item response function can also be viewed as the regression of item score on ability. In many statistical contexts, regression functions remain unchanged. In the present context this should be quite clear: The probability of a correct answer to item i from examinees at a given ability level T_0 depends only on T_0, not on the number of people at T_0, not on the number of people at other ability levels T_1, T_2,...T_n. Since the regression is invariant, its lower asymptote, its point of inflexion, and the slope at this point all stay the same regardless of the distribution of ability in the group tested. Thus a_i, b_i, and c_i are invariant item parameters. According to the model, they remain the same regardless of the group tested."

The Identification Problem

Although item parameters and thetas are invariant from one examinee group to another, they may not appear to be invariant because the scale of the estimates is arbitrary and a linear transformation is required to put the estimates from different groups on the same footing. The arbitrariness or indeterminancy of the scale is formally known as the identification problem (Hambleton & Swaminathan, 1985). To resolve this identification problem it is necessary to fix the scale of estimates that are to be compared across groups.

The three parameter model may be transformed where T becomes T_*, a_i becomes a_{i*}, b_i becomes b_{i*} and c_i becomes c_{i*}, such that:

$$T_* = m\,(T) + n$$

$$b_{i*} = m\,(b_i) + n$$

$$a_{i*} = a_i/m$$

$$c_{i*} = c_i$$

so that an invariant item response function results:

$$P_i(T_* \mid a_{i*}, b_{i*}, c_{i*}) = P_i(T \mid a_i, b_i, c_i)$$

(Hambleton & Swaminathan, 1985). One approach to this problem is to fix the scale of theta to have mean zero and standard deviation one by choosing j and k appropriately.

Factors Influencing Parameter Estimation

Invariance of item parameters and ability estimates is not unlike the concept of invariance of the functional relationship obtained in linear regression (Hambleton & Swaminathan, 1985). A linear regression line is theoretically invariant regardless of the distribution of the independent variable. However, although a true or population regression line exists, proper estimation of the line may be affected by sample size and restriction of range.

The problem of range restriction may be further exacerbated by the non-linearity of the ICC (Hambleton & Swaminathan, 1985). The difficulty is that the non-linear form of the logistic function requires that the curves in the more complex function also be estimated and sufficient data points must be available to achieve proper estimates of these curves.

Although IRT provides a sound theoretical basis for item parameter invariance, an important issue is that the stability of estimates obtained from extreme ability groups is not known. In light of the discussion above, it is clear that range restriction could be an important influence on parameter estimates when extreme groups are used for the detection of item bias. Because range restriction may be an issue in the detection of item bias, the next section will explore the literature on the technique of using estimates based on extreme groups to detect item bias.

Item Bias Detection Methods and Parameter Invariance

Bias arise when groups of examinees (e.g., Males and Females) who are equal in ability, differ in item performance (Hambleton & Swaminathan, 1985). Pine (1977) defined an unbiased item as an item for which different subgroups of equal ability have the same probability of getting the item correct. Given this orientation, IRT provides a natural framework for studying item bias.

Three methods of detecting item bias using IRT models are documented by Hambleton and Swaminathan (1985). Method one is the "area" method, in which differences between the item characteristic curves are compared across subgroups of interest. A second and logically equivalent method is to compare item parameters across subgroups. If invariance is not obtained for a particular item, one potential explanation is that the item is biased. Another way to view this might be to consider such an item multidimensional.

A third approach is to investigate model data fit. If the model fits the data, then the expected feature of item parameter invariance is assured and item bias can be ruled out. These methods should yield equivalent results.

Comparing ICC's and comparing item parameters across subgroups should be identical because the ICC's are defined by the item parameters. However, it has been argued that ICC's may show very little difference while item parameter estimates may seem to be quite different (Linn, Levine, Hastings & Wardrop, 1981). While this implies that ICC's may be more appropriate for the study of item bias, it has also been argued that ICC's may disguise "true" bias (Hambleton & Swaminathan, 1985).

The most sensitive and direct method of checking for item parameter invariance is a method that would compare the item parameters across different groups. According to Lord (1980) "The invariance of item parameters across groups is one of the most important characteristics of item response theory."

Hambleton and Murray (1983) used a technique for comparing item parameters, along with other methods, to explore goodness of fit. The method of assessing model data fit relates to the detection of item bias through a tautology. It is known that model data fit implies obtaining the expected features, and therefore, demonstrating either one should be sufficient to guarantee the other.

The technique, based on an idea by Angoff (1982), was intended to detect bias using classical item statistics, but was adapted by Hambleton and Murray for use with IRT models. Hambleton and Murray adapted Angoff's approach, which is descriptive in nature, because statistical methods of detecting invariance may be inadequate. Because of the large sample sizes required for estimation of parameters when using IRT models, statistical approaches are hampered by their extreme sensitivity to differences that may not be significant in practical terms.

Hambleton and Murray's approach was to split a parent sample of examinees into subgroups according to background variables, such as males and females or blacks and whites, where differences in ability might have been expected. Item difficulty or b estimates are obtained for blacks and whites. If the b estimates were invariant, Hambleton and Murray (1983) argued that scatterplots of the estimates should fall on a straight line, with positive slope. However, because of sampling errors, this may not be realized in practice. To address this problem, Hambleton and Murray (1983) obtained a baseline for comparison.

To obtain a basis for comparison, each examinee subgroup is divided randomly into two groups, parameter estimates obtained, and scatterplots generated for the four groups. Hambleton and Murray (1983) reasoned that scatterplots based on estimates of random samples within each subgroup could be used to demonstrate sampling error. The degree of scatter from the random within groups could then be used as a baseline of comparison for cross subgroup scatterplots. Scatter that exceeds the envelope of scatter established by the random-within plots might be attributed to bias.

Hambleton and Murray (1983) found more scatter for between groups than for random within groups. This implied either the model did not fit the data, (hence invariance was not obtained), or that item bias was pertinent. Another possibility proposed was that parameter invariance may not have been observed because extreme groups were leading to poor estimates due to range restriction.

Another potential influence may have been the effect of sample size on the precision of the estimates. Hambleton & Murray (1983) worked with samples of 165 examinees. These samples may have been too small for obtaining proper estimates.

Preliminary Study of the Invariance Property

Mooney and Swaminathan (1986) sought to establish a better understanding of the problem of sampling error and its effect on the technique used by Hambleton and Murray (1983). Mooney and Swaminathan (1986) obtained thirty estimates for each item difficulty using random samples of 600 subjects drawn with replacement from a single parent sample of 1,200 subjects. Test items were from National Assessment of Educational Progress (NAEP) field data. They obtained distributions of b parameter estimates based on these samples. They reasoned that the distributions based on the random samples would offer a good baseline of comparison for estimates obtained from subgroups from the same population. Comparison groups that differed in educational background were used, where low education included formal education up to and including High School, while high education included all subjects who reported education beyond High School.

Using plus or minus two standard deviations on the sampling distribution of the randomly sampled examinee groups as criteria, Mooney and Swaminathan (1986) found that parameter invariance was found to a higher degree in the low education group than in the high education group. They found that 8 out of the 34 items (24%) were misfitting for the low educational background group, while 20 out of the 34 (59%) were misfitting for the high educational background group.

Because the test was not difficult, Mooney and Swaminathan (1986) reasoned that range restriction may have influenced the estimates more for the high educational background group than for the low educational background group. In other words, for the high education group there were some items that nearly everyone got correct, thereby introducing restriction of range. This phenomena is sometimes termed a "ceiling effect" by psychometricians. For the low education group, on the other hand, the general difficulty of the test demonstrated better balance in relation to the group's ability.

Mooney and Swaminathan (1986) repeated the analysis for samples of size 300 subjects. They split the two educational background groups randomly (designated L1 and L2 for the low education group and H1 and H2 for the high education group), re-estimated the b parameters, and compared them to a baseline of b estimates based on random samples of the same size.

Fit appeared better in this case but was interpreted to have due to the fact that the randomly obtained estimates had about twice the sampling error. (The average standard deviation for the random samples of 600 examines was about .10, while the average standard deviation for the 300 group was about .20.) Looking at Figure 1, each of the two low educational background groups (L1 and L2) had only 4 misfitting items (11%), while the high educational background groups (H1 and H2) had 10 and 11 (29% and 32%), respectively.

In comparing misfitting items for like groups in the 300 sample, Mooney and Swaminathan (1986) found much higher agreement among the low educational background examinee's estimates (82%) than for the high educational background examinee's (50%). (Agreement was calculated by the sum of the diagonal cells of Figure 1 divided by the total number of items.) This finding further supported the idea that differences resulted from range restriction.

One problem with this conclusion, however, was that Mooney and Swaminathan (1986) obtained some out-of-bounds estimates for the high educational background group that required adjustment before the estimates could be compared. Because no established best method exists for determining how to rescale in these circumstances, three methods were compared: no adjustment, missing values, and recoding out-of-bounds estimates to ± 3.00. The recoding method was chosen because it demonstrated the best average fit.

Out-of-bounds estimates for the high educational background group may indeed be a sufficient indicator that range restriction is related to the stability of the estimates although this may be confounded with factors such as item bias and model data fit. It could be, for example, that differences would not have been found, or that they would have been minimized, had the three parameter model been used.

300 H2	300 H1: OUT	300 H1: NOT OUT	
OUT	2	8	10
NOT OUT	9	15	24
	11	23	

Agreement=17/34=50%

300 L2	300 L1: OUT	300 L1: NOT OUT	
OUT	1	3	4
NOT OUT	3	27	30
	4	30	

Agreement=28/34=82%

Figure 1. B-parameter Outliers Within Group Comparisons.

Conclusion

We know that range restriction has an effect on parameter estimation and consequently on the invariance of item parameter estimates. The study by Mooney and Swaminathan (1986) illustrates the need for further investigation into this issue. It is known from the above studies that range restriction of ability may result in extreme out-of-bounds estimates and that this may have serious effect on our ability to study item bias. Although in previous work Mooney and Swaminathan (1986) have confirmed this, they worked with only the two parameter model and the effect of range restriction of ability with the three parameter model needs to be examined. Furthermore, the question of model-data fit could not be assured using field data.

Needed is a study using repeated samplings that would rule out the question of model-data fit as well as item bias while controlling for the two factors of interest: sample size and range restriction of theta.

CHAPTER III

DESIGN OF THE STUDY

Introduction

The purpose of this study is to address the issues outlined in the introduction:

1. How does range of ability affect the invariance of the estimates of the difficulty parameters in the three parameter IRT model?

2. What is the influence of sample size in the invariance of the estimates of the difficulty parameter?

3. What is the consequence of interaction of range of ability with sample size?

To investigate the above questions three sample sizes (n=600, n=900, and n=1200) were completely crossed with three levels of ability range. This yielded a 3 by 3 design with 9 testing situations. Within each of these testing situations, 30 sets of test data were generated using a resampling technique and parameters were estimated using LOGIST4 (Wingersky, M. S., Barton, M. S., & Lord, F. M., 1982). The estimates of the b parameters obtained from LOGIST4 were then compared for comparison across the various testing situations.

Since previous research by Mooney and Swaminathan (1986), has raised issues about the influence of range restriction in a field data study, three levels of range restriction were chosen: 1) a symetric distribution of examinee ability, 2) a moderately positively skewed distribution of examinee ability, and 3) a highly positively skewed distribution of ability. The most extreme level of skewness was determined using empirical methods, and the middle level of skewness was selected as an intermediate position between the most extreme level of skewness and the normal distribution.

Positively skewed distributions were used in this study to facilitate the estimation of the lower asymptote of the three parameter model. The purpose of this study was to explore the relationship between the skewness of the ability distribution and the expected feature of b parameter invariance. Accordingly, either a positively or negatively skewed distribution of ability would be appropriate for this investigation. However, a belief expressed by (Hambleton & Swaminathan, 1985) is that the lower asymptote can be estimated well only when sufficient examinees are available at the lower levels of ability. To avoid confounding poor estimation of the lower asymptote with the quality of estimation of the b parameter in this study, positive skewness of ability was chosen.

Although the influence of range restriction on parameter estimates was the focus of this dissertation, an important factor that may also influence parameter estimation is sample size. Although Lord (1980) recommends samples of approximately 1,000 subjects when using the three parameter model, previous research by Swaminathan and

Gifford (in press) suggests that sample sizes as small as 600 may be reasonable. Accordingly, three sample sizes were chosen for this dissertation, 600, 900 and 1200.

Description of the Data

DATAGEN: For this study, data were generated using the DATAGEN program (Hambleton & Rovinelli, 1973). In order to adequately study item bias detection using extreme groups, the influence of range restriction on parameter estimation of item response models must be studied. This requires knowledge of the true values of the parameters. Simulated data are also an important means of controlling for the influence of model-data misfit.

DATAGEN allows specification of population parameters for the item parameters a_j, b_j, c_j ($j = 1, 2, ..., n$) and for ability parameters T_i, ($i = 1, 2, ..., n$). A uniform or normal distribution may be specified and the true parameter values are then randomly drawn from the distribution.

DATAGEN generates dichotomous examinee responses based on the item response model and the parameter values. (An individual examinee, a, for response, P_{ag}, is then generated based on the given probabilistic item response model for a given item, g.) A random number drawn between the interval (0,1) is then compared to the estimated probability P_{ag}. A score of 1 is given when P_{ag} is greater than the number drawn, otherwise, the examinee obtains a score of 0.

An entire matrix of examinee responses is generated by DATAGEN according to a specified number of examinees and items. This matrix is then available for analysis using LOGIST4.

One concern with analyzing data simulated by DATAGEN is that DATAGEN will generate total examinee responses with perfect and zero scores. Maximum likelihood estimate corresponding to these cannot be obtained. To solve this problem, LOGIST4 removes all cases of perfect and zero scores before the analysis. One concern with this approach in the context of this study, is that these removed cases would influence sample size. To avoid these slight discrepancies, a version of DATAGEN modified by Dr. Janice Gifford was used so that no examinee will have perfect or zero scores.

Review of Specific Steps Taken for Data Generation

In order to obtain the data for this analysis, the following steps were taken:

Step 1: Using the modified version of DATAGEN, a sample of theta values and their associated response vectors were randomly generated to simulate a uniform distribution for 6,000 examinees over 60 items with known item parameters. In generating the values of the item parameters, the a parameters were uniformly distributed over the interval (.60, 1.90), the b parameters over the interval (-1.73, 1.73) and the c parameters over the interval (.15, .25). The theta estimates were also uniformly distributed, and the interval (-1.73, 1.73) was used so that thetas would not be generated that would be beyond the range of the b's.

Step 2: The 1st generation data set was partitioned into 20 equal intervals of theta. Distributions were then obtained by randomly sampling from each of the twenty intervals of ability. Table 1 displays the intervals of theta and the percentages sampled from each of the intervals for each of the three distributions of ability.

Step 3: Three ranges of theta were chosen, level 1, level 2 and level 3. The level 1 distribution centered the majority of the population parameters for ability toward the middle of a symmetric distribution (see Figure 2). The level 2 distribution has a positive skew of ability, with 5% of the population ability parameters in the last five intervals (see Figure 3). The level 3 distribution is more highly positively skewed than the level 2 distribution. The level 3 distribution has 10% of the population parameters for ability in the last 10 intervals. This is displayed in Figure 4.

The arrangement of 3 levels of sample size by three levels of skewness, produces 9 different testing situations. These 9 testing situations are depicted in Figure 5.

In order to construct the distribution at the appropriate ability levels corresponding to the appropriate ability level and sample size in Figure 5, the following steps were taken:

a: From Table 1, the percentage of examinees at a given interval of theta were determined. For example, if interval under consideration was -0.186 to -0.004, 4.0 percent or 36 theta values were selected uniformly in the interval.

b: Thirty samples of item responses were obtained randomly with replacement from the distributions constructed in "a." For convenience, the total data set for the 9 cells were obtained simultaneously.

Table 1

Distributions of Theta

	Range of Interval (n = 6,000)	Level 1 %	Level 2 %	Level 3 %
1	-1.730 to -1.542	1.00	1.90	4.00
2	-1.540 to -1.381	1.00	2.90	8.00
3	-1.386 to -1.205	1.80	3.90	9.00
4	-1.202 to -1.043	2.80	4.90	1.10
5	-1.041 to -0.863	4.00	6.70	1.30
6	-0.861 to -0.709	4.80	8.60	1.30
7	-0.706 to -0.538	6.70	1.21	1.10
8	-0.521 to -0.359	8.00	1.30	9.00
9	-0.355 to -0.182	9.00	1.21	8.00
10	-0.186 to -0.004	1.10	8.60	4.00
11	-0.007 to 0.154	1.10	6.70	1.00
12	0.150 to 0.323	9.00	4.90	1.00
13	0.321 to 0.495	8.00	3.90	1.00
14	0.491 to 0.655	6.70	2.90	1.00
15	0.652 to 0.833	4.80	1.90	1.00
16	0.836 to 1.004	4.00	1.00	1.00
17	1.009 to 1.185	2.80	1.00	1.00
18	1.188 to 1.376	1.80	1.00	1.00
19	1.373 to 1.550	1.00	1.00	1.00
20	1.559 to 1.728	1.00	1.00	1.00

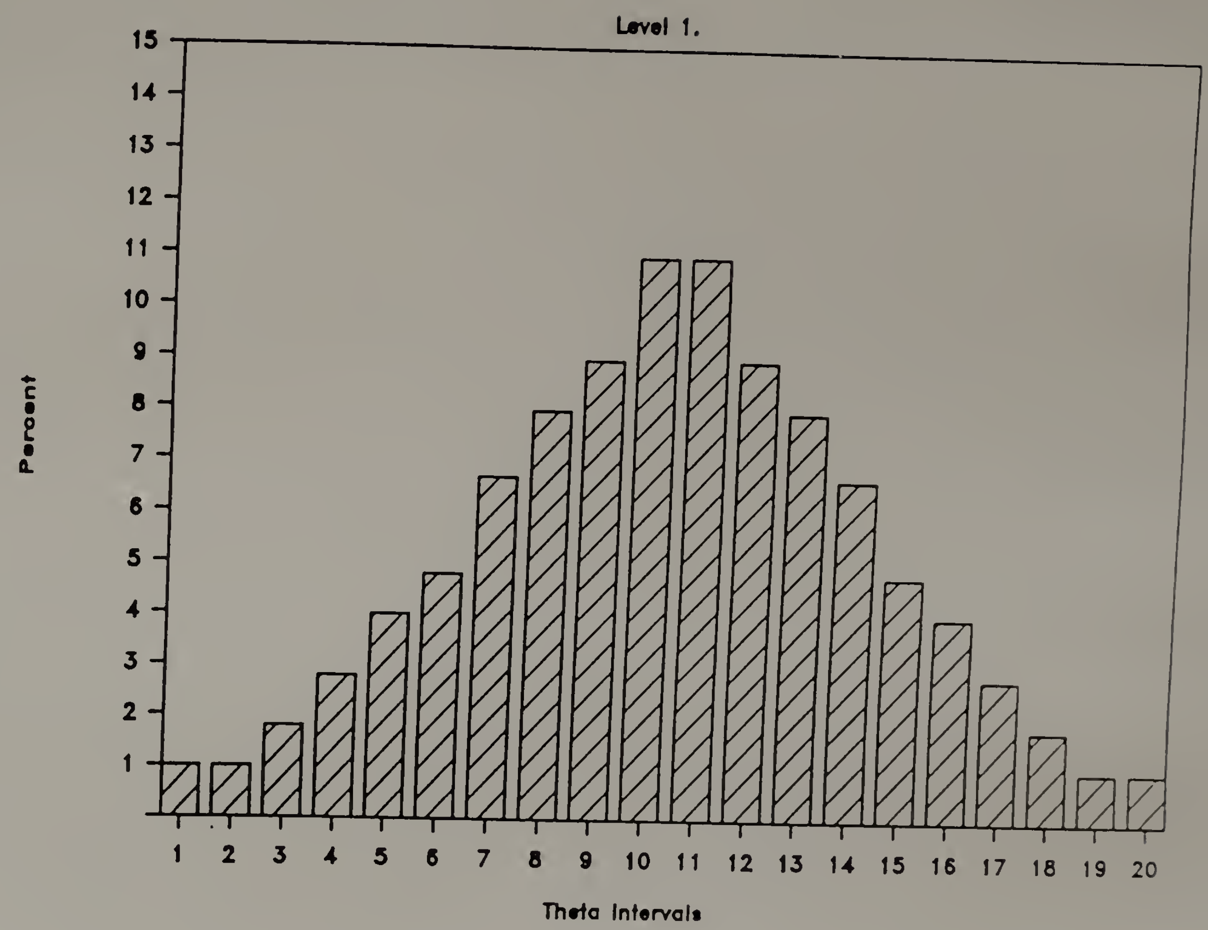

Figure 2, Distribution of Theta.

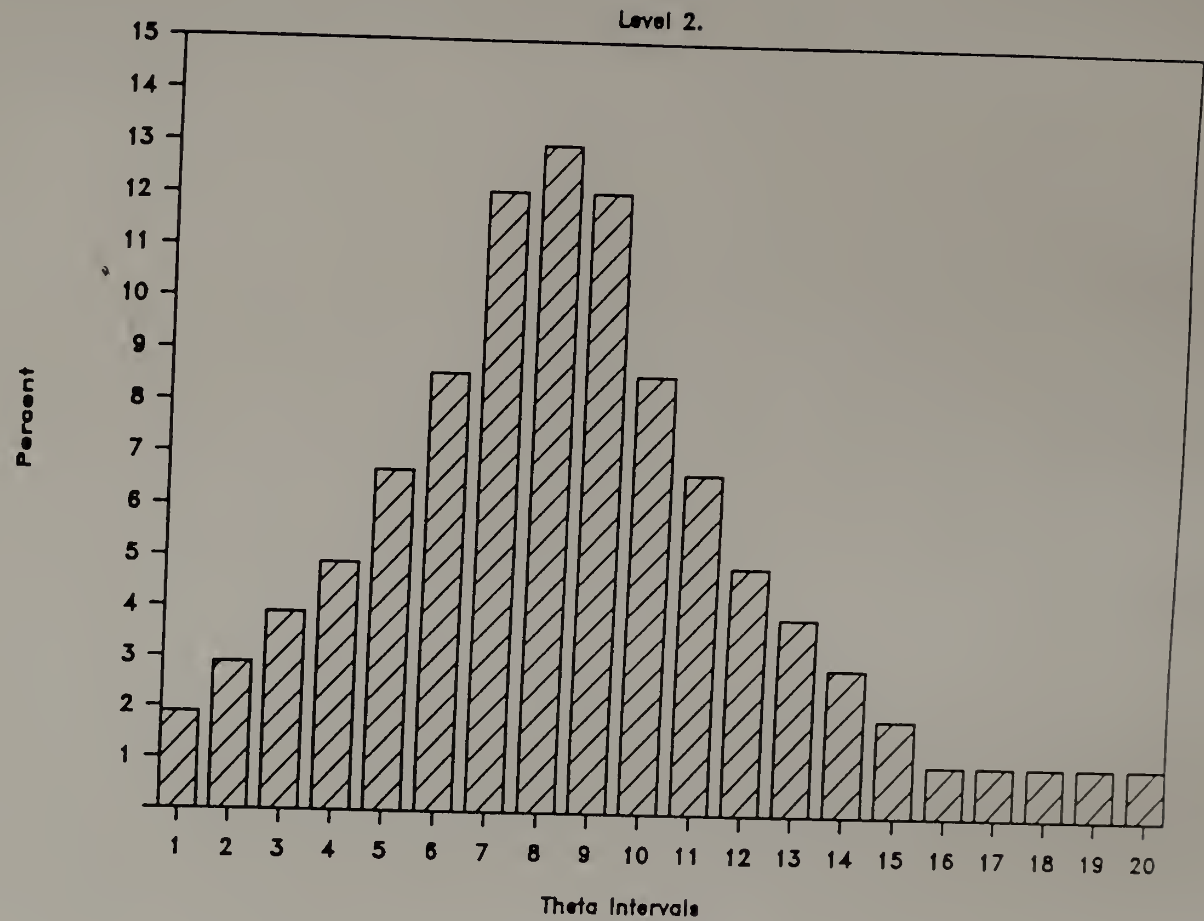

Figure 3. Distribution of Theta.

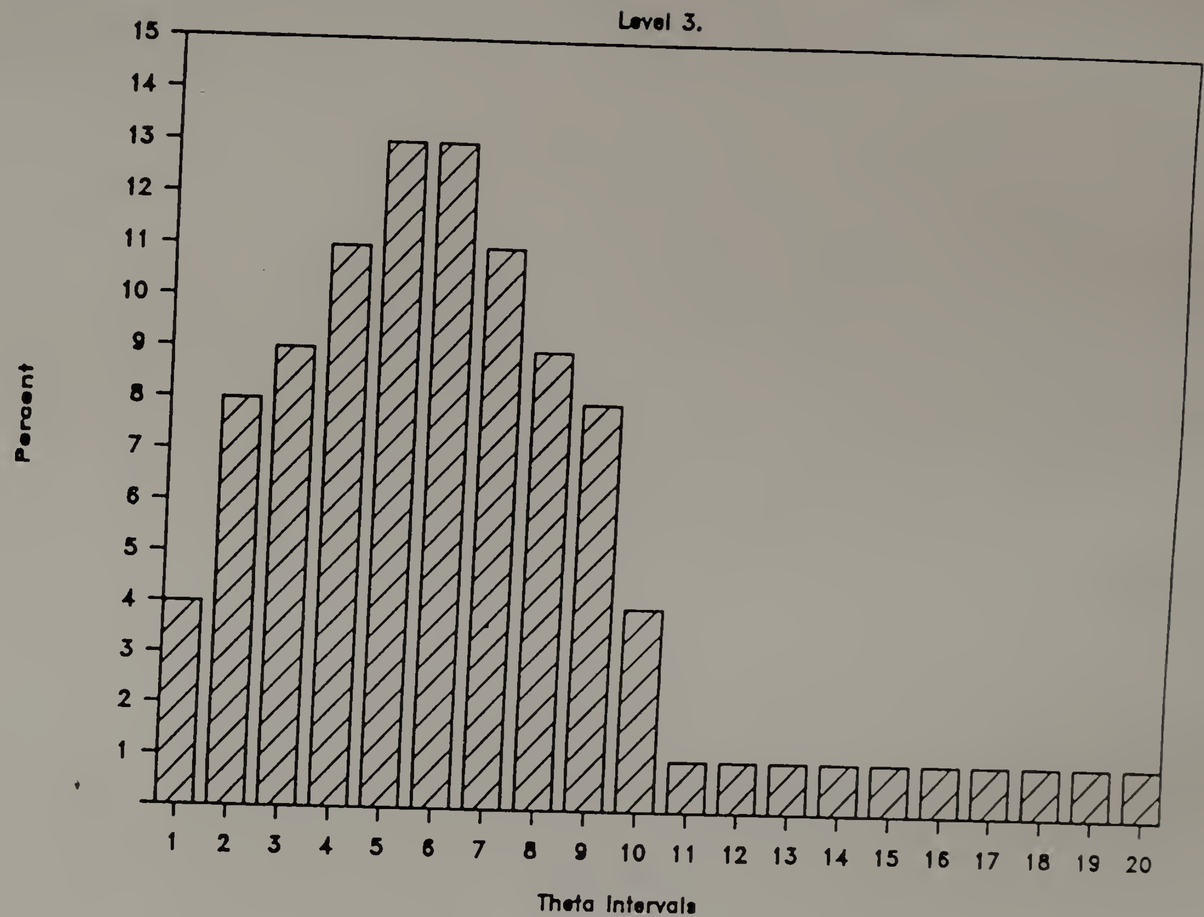

Figure 4. Distribution of Theta.

		Number of Examinees		
		n = 600	n = 900	n = 1200
Distribution of Ability	1			
	2			
	3			

Figure 5. Structure of the Experiment.

Within each cell, parameter estimates were derived from repeated samples randomly obtained with replacement from the parent distributions. For each of the 9 testing situations, 30 sets of test data were generated for 60 items. The b estimates for each of the testing situations were subtracted from the known true values. The sum of the squares of these values were then compared by ranking the items by true score difficulty and displayed as histograms. To improve the interpretability of the results, items were grouped in sets of five. Distributions of bias and variability are also displayed.

Estimation of Parameters

LOGIST4 (Wingersky & Lord, 1976) was used exclusively for parameter estimation in this study. Item parameters estimates were scaled to mean zero and unit variance. This was done to remove the indeterminancy of the item difficulty scale so that item difficulty scale would be comparable across groups. The number of answer choices was five, so that the probability of guessing was 1.0/5. The maximum number of iterations was set at 40, with 6 interations per stage and an overall maximum of 600 seconds for the run. The default settings were chosen for any remaining selections.

Thirty samples were randomly obtained with replacement from each of the three levels of theta distributions. This resampling technique is modeled on the bootstrap method of resampling proposed by Efron (1982).

After obtaining 30 sets of responses cloned from each of the appropriate 2nd generation data sets, LOGIST4 estimates of the item parameters were obtained for each of the cells. In each case, standardized estimates of the item parameters are then subtracted from the standardized true values of the item parameters.

Assessment of Parameter Invariance

In order to address the issue of invariance, two different methods of assessing the invariance property will be used. One method assessed the accuracy of the estimates, and a seperate method assessed the stability of the estimates.

Accuracy: accuracy refers to the degree to which estimates recover the known population value.

In order to assess the accuracy of estimation, the mean squared difference is computed as below:

$$MSD = \sum_{i=1}^{30} (t_i - \text{True})^2 \quad ,$$

Here, t_i is the estimate obtained from an individual replication and True is the true value for a given b. For each item the mean squared difference between the estimates and the true value was calculated.

The accuracy estimates for each level of range restriction are graphically depicted within each level of sample size, while items are ranked according to true value difficulty. Graphs of variance and bias are also provided. By comparing the three levels of range restriction within sample size, the influence of levels of ability on extreme items may be readily interpreted across these three indices.

Accuracy represents the degree to which estimates recover the known population value. Accuracy alone, however, is not enough to answer the question of invariance. The mean squared difference given above can be partitioned into two additive components, variance and bias, i.e.

$$MSD = V(t) + B(T)$$

Gifford and Swaminathan (in press).

Variance: The variance V (t) is given as:

$$V(t) = \sum_{i=1}^{30} (t_i - t.)^2 \quad ,$$

where t. is the mean of the estimates obtained over the 30 replications.

Bias: For each B estimate, bias was calculated as:

$$B(t) = (t. - T)^2 \quad .$$

While MSD is an index of the accuracy of the estimate, it does not provide an explanation of the differences between the estimate and the True value. Partitioning MSD into sampling error and systematic bias provides this explanation. For example, two estimates that obtain the same accuracy may differ with respect to variance and bias. It could be, that estimates are not accurate, but that they are <u>consistent</u> and <u>therefore invariant</u>.

Item Stability: In order to assess invariance, accuracy and variability of the estimates must be assessed under separate conditions. Therefore, stability will be assessed by investigating the nature of the distribution of the estimates by providing an arbitrary benchmark. Following the estimation of item parameters for each cell, the estimates for each item are grouped and rescaled to mean zero, standard deviation one. If the majority of the rescaled estimates (95%) fall within two standard deviations, the estimates will be considered invariant.

CHAPTER IV

RESULTS AND DISCUSSION

Item parameter invariance is a key property of IRT models, and it is a property that sets IRT apart from classical test theory models. Item parameter invariance is important for a number of testing issues, but one of the most direct and straight-forward examples of the use of this property arises in the study of item bias. Here, the estimates from different groups are obtained independently and then compared to determine if individual items are behaving differently for different groups.

A question that naturally arises from this application is the matter of the degree to which parameter invariance holds for different samples. Although parameter invariance is not being questioned, there may exist issues with the quality of parameter estimation that could frustrate the application of the invariance property in practical settings. Hence, the purpose of this dissertation was to answer the following research questions:

1. How does range of ability affect the invariance of the estimates of the difficulty parameters in the three parameter IRT model?

2. What is the influence of sample size in the invariance of the estimates of the difficulty parameter?

3. What is the consequence of interaction of range of ability with sample size?

These questions will be considered in terms of the stability of the parameter estimates, as well as in terms of the accuracy, bias and variance of the estimates as described in Chapter III. To obtain data for this analysis, a sample of simulated responses for 6,000 examinees for 60 items was generated for the three parameter IRT model.

Thirty samples for each of nine testing situations were then constructed from the population of 6,000, varying across three level of sample size and ability distribution. Repeated samples were then obtained from each of the testing situations in order to better understand the behavior of the estimates. The b estimates for each of the 60 items from each situation were then compared in order to establish what, if any, differences exist among the testing situations.

Descriptive Statistics

The population item parameters obtained from DATAGEN are reported in Table 2. These population parameters were then rescaled to mean zero and unit variance. Each item is ranked in order of item difficulty. The purpose of ranking items is to provide a better understanding of item difficulty as it relates to the ability distributor. For example, if the distribution of ability in positively skewed, difficult items may be less estimated with greater variability over replications than would be the case for an item whose difficulty level falls near the mode of the ability distribution.

Table 2

True Item Parameters

item	rank	b	a	c	p
4	1	-1.497	1.864	.167	.917
55	2	-1.345	1.446	.238	.890
44	3	-1.342	.941	.152	.838
40	4	-1.339	1.386	.190	.876
22	5	-1.303	.737	.170	.821
35	6	-1.245	1.392	.189	.865
12	7	-1.123	.722	.168	.787
29	8	-1.118	.912	.176	.801
7	9	-1.107	.983	.217	.815
8	10	-1.096	1.676	.198	.840
56	11	-.910	.992	.158	.770
9	12	-.892	1.166	.238	.796
48	13	-.872	.912	.215	.776
16	14	-.863	1.210	.173	.774
47	15	-.781	1.135	.196	.760
3	16	-.718	1.252	.212	.749
59	17	-.713	.701	.185	.717
36	18	-.626	.785	.226	.721
31	19	-.624	.780	.191	.712
45	20	-.592	1.841	.224	.738
26	21	-.579	1.556	.221	.749
27	22	-.553	1.638	.190	.725
14	23	-.541	1.111	.224	.721
21	24	-.494	1.027	.170	.694
41	25	-.471	1.513	.162	.685
15	26	-.394	.679	.169	.655
50	27	-.393	1.815	.235	.698
28	28	-.370	.641	.207	.670
39	29	-.331	1.778	.227	.680
23	30	-.319	1.189	.161	.649

Table 2 (continued)

item	rank	b	a	c	p
11	31	-.302	1.382	.194	.664
57	32	-.253	1.817	.228	.675
19	33	-.114	.863	.165	.617
43	34	-.059	1.294	.191	.610
18	35	-.053	1.773	.239	.636
34	36	.035	1.306	.186	.578
60	37	.089	.832	.216	.591
2	38	.127	.697	.153	.542
10	39	.168	1.007	.173	.550
20	40	.266	1.207	.217	.541
6	41	.281	.839	.201	.541
37	42	.384	1.157	.237	.539
13	43	.521	1.710	.182	.455
5	44	.538	1.526	.234	.498
32	45	.579	1.061	.239	.500
25	46	.656	1.619	.168	.423
46	47	.679	1.001	.162	.432
49	48	.717	1.794	.190	.431
58	49	.855	1.057	.224	.437
42	50	.907	1.786	.212	.400
1	51	.925	.670	.198	.431
54	52	.973	1.181	.228	.429
38	53	1.327	1.787	.237	.339
30	54	1.381	.805	.163	.329
24	55	1.400	.939	.237	.366
17	56	1.446	1.628	.204	.296
53	57	1.541	1.556	.192	.275
52	58	1.562	1.773	.187	.256
51	59	1.589	.797	.222	.356
33	60	1.728	1.649	.249	.300

Under ideal conditions, all items would have low and consistant variability, with no influence due to item difficulty. The rescaled b parameter population values are reported in Table 3.

Item difficulty parameters show a good range from -1.497 for the least difficult item to 1.728 for the most difficult item. This full range of difficulty is also reflected in the p-values, which range from .917 for the easiest item to .300 for the most difficult item.

To introduce maximum stress to the design, a high degree of range restriction was employed. The patterns of the distributions of the three levels of range restriction are shown in Figures 2, 3 and 4. One concern with employing a high degree of range restriction, however, is the influence that the range restriction may have on the behavior of the estimates, particularly for item discrimination.

Data sets that obtained poor estimation of the discrimination parameters were not included in the study. Figure 6, below, shows the number of runs that had poor estimation of the a parameter. Figure 6 shows the count and percentage of discarded estimation samples for each of the nine testing situations. It can be seen from Figure 6 that 27 (90%) of the runs for the distribution with the most extreme degree of positive skew for sample size 600 included poor estimates for the a parameters. This suggests that this combination of sample size and skewness results in a breakdown of the estimation procedure.

Table 3

Rescaled B Estimates
(n = 6,000)

item	rank	b true (rescaled)	p-value
4	1	-1.579	.917
55	2	-1.410	.890
44	3	-1.407	.838
40	4	-1.403	.876
22	5	-1.363	.821
35	6	-1.299	.865
12	7	-1.163	.787
29	8	-1.157	.801
7	9	-1.145	.815
8	10	-1.133	.840
56	11	-0.926	.770
9	12	-0.906	.796
48	13	-0.883	.776
16	14	-0.873	.774
47	15	-0.782	.760
3	16	-0.712	.749
59	17	-0.706	.717
36	18	-0.610	.721
31	19	-0.607	.712
45	20	-0.572	.738
26	21	-0.557	.749
27	22	-0.528	.725
14	23	-0.515	.721
21	24	-0.463	.694
41	25	-0.437	.685
15	26	-0.351	.655
50	27	-0.350	.698
28	28	-0.325	.670
39	29	-0.281	.680
23	30	-0.268	.649
11	31	-0.249	.664
57	32	-0.194	.675

Table 3 (continued)

item	rank	b true (rescaled)	p-value
19	33	-0.040	.617
43	34	0.022	.610
18	35	0.028	.636
34	36	0.126	.578
60	37	0.186	.591
2	38	0.229	.542
10	39	0.274	.550
20	40	0.384	.541
6	41	0.400	.541
37	42	0.459	.539
13	43	0.667	.455
5	44	0.686	.498
32	45	0.732	.500
25	46	0.818	.423
46	47	0.843	.432
49	48	0.886	.431
58	49	1.039	.437
42	50	1.097	.400
51	1	1.117	.431
54	52	1.171	.429
38	53	1.565	.339
30	54	1.625	.329
24	55	1.646	.366
17	56	1.697	.296
53	57	1.803	.275
52	58	1.826	.256
51	59	1.856	.356
33	60	2.011	.300

Distribution of Ability		Number of Examinees n = 600	n = 900	n = 1200
	1	3 10%	7 23%	1 3%
	2	8 27%	1 3%	7 23%
	3	27 90%	6 20%	6 20%

Figure 6. Unsuccessful LOGIST RUNS.

Data Analysis

The 30 sets of b estimates obtained from each of the nine testing situations were rescaled to mean zero and unit variance. The population parameter values were then subtracted from each of the estimates, and the results are reported in Appendix A. The purpose of looking at this average variance is to better characterize the overall influence of sample size and restriction of range of the variability of the estimates. The average variance of each of the testing situations over 30 samples is reported in Figure 7.

Reading Figure 7 from left to right, for ability level 1 (the normal distribution of ability), variance for the 600 sample was .161, while variance for the 1200 sample is .127 -- a difference of .034. The level 1 distribution of ability produces decreasing variance of estimates as sample size increases.

For ability level 2 and 3, the pattern of decreasing variance with increased sample size does not hold. For ability level 3, for example, the variance of the estimates for the 600 sample is .020 lower than the estimates for the 1200 sample. These small differences do not appear to demonstrate any clear and constant effects due to sample size.

The picture reading down Figure 7 is somewhat different. Here the pattern of difference among the variance of the estimates seems to show a consistant decrease in variance as positive skews increased from the normal distribution to the more positively skewed distribution. For the 600 sample, for example, ability level 1

Distribution of Ability	n = 600	n = 900	n = 1200
1	.161	.151	.127
2	.126	.110	.132
3	.105	.113	.125

Number of Examinees

Figure 7. Average Variance of B Estimates Over Thirty Samples.

variance is .161, which decreases to .126 for ability level 2, and to .105 for ability level 3.

Although there is an indication of decreases in the variability of the estimates with increased skewness of the ability distribution, differences among the mean variance estimates do not appear to demonstrate any dramatic and consistent changes over the two factors of the design. This would seem to imply that the variation among the estimates is not influenced by changes in sample size, ability distribution differences, or by the interaction of the two.

Stability Assessment

An analysis of the stability of parameter estimates is displayed in Figure 8. Stability is defined as the variability of the estimates based on repeated samplings. This analysis provides an empirical investigation of model-data fit. Here each of the rescaled estimates, standardized to mean zero and unit variance, are presented in terms of the percentage of estimates within one and two standard deviations. Each cell includes 60 items by 30 replications. For normally distributed estimates, it would be anticipated that about 68% of the estimates would fall within one standard deviation of the mean and 95% would fall within two standard deviations of the mean. Stable estimates could be anticipated to behave as approximately normal deviates. It is clear from Figure 8 that the estimates are within the expected cut points.

Distribution of Ability	Number of Examinees: n = 600	n = 900	n = 1200
1	1 sd 71.1 2 sd 95.4	1 sd 69.6 2 sd 96.1	1 sd 68.7 2 sd 96.3
2	1 sd 67.1 2 sd 96.4	1 sd 67.4 2 sd 96.1	1 sd 70.0 2 sd 95.7
3	1 sd 69.7 2 sd 95.5	1 sd 69.8 2 sd 95.2	1 sd 67.9 2 sd 96.0

Figure 8. Percentage of Parameter Estimates Within One and Two Standard Deviations.

Taking the symetric distribution (ability level 1) and the smaller sample size (N=600). As an example, Figure 8 shows that 71.1% of the estimates fall within one standard deviation and 95.4% fall within two standard deviations. The greatest expected contrast from the ability level 1, N=600 cell would be ability level 3, N=1200 cell in the lower right hand corner. Here, the ability distribution is at the maximum positive skew and sample size has been doubled. However, for the ability level 3, N=1200 cell, the picture is much the same as was the picture for the level 1, N=600 cell. 67.9% of the ability level 3, N=1200 cell estimates fell within one standard deviation, and 96.0% fell within two standard deviation.

The differences among the nine testing situations appear to be modest. In terms of the three research questions it appears that the range of ability does not influence the invariance of item parameters over sample size or distribution of ability, nor does it appear that these two factors interact.

Figure 8 provides evidence to demonstrate that the model fits the data for all combinations of the two factors. A more detailed inspection of the behavior of the individual items is available in Appendix B. Here, items are ranked by difficulty and compared in terms of the percentage of estimates falling within one, two, three and four standard deviations from the mean. At this more detailed level of inspection, model-data fit appears to hold with a high degree of consistency.

Analysis of Accuracy, Bias and Variance

As described in Chapter III, accuracy can be partitioned into two additive components, variance and bias. Accuracy was interpreted as the degree to which the sample estimates are close to one another, and bias was indicated by the degree to which the means of the estimates differ from the population value. Recall that accuracy, MSD(b), for item difficulty can be partitioned into variance, V(b), and Bias, B(b), i.e., MSD(b) = V(b) + B(b). The mean and standard deviation of MSD(b), V(b), B(b) in each item grouping are compared across the various testing situations. This analysis investigates quality of estimation on the item level and is therefore more highly focussed than the previous anlaysis. Items have been ranked by the population b parameters, and grouped in sets of five to simplify the task of observing change across the 9 testing situations. These items range from 1-12 where 1 indicates the easiest set of items and 12 indicates the most difficult set of items. (Variance bias and accuracy for individual items is presented in Appendix C.)

Table 4 provides the means and standard deviations of V(b) for each of the three ability distributions for sample size of 600. The effect of changing ability distributions on the item difficulty estimates are reflected in the higher V(b) scores. The higher mean scores indicate a large variability of estimates over replications. Ability Level 1, for example, shows lower variability consistency at the extreme ranges of item difficulty.

Table 4

Means and Standard Deviation of V(b) for Item Groups
(n=600)

Item Group	Ability Level 1		Ability Level 2		Ability Level 3	
	mean	s.d.	mean	s.d.	mean	s.d.
1	3.0784	4.3264	1.1473	.3775	.4903	.2011
2	.6857	.3823	.4099	.2106	.2202	.0877
3	.5643	.3276	.3338	.1962	.1813	.0544
4	.3412	.2514	.3788	.3444	.1873	.1051
5	.1626	.0360	.0936	.0508	.1366	.0901
6	.2579	.2884	.1770	.1604	.2645	.3262
7	.1741	.2034	.1391	.0703	.1803	.0868
8	.2111	.0660	.3944	.2879	.4132	.1788
9	.4319	.2994	.6096	.5289	.6451	.2718
10	.4047	.2694	.5520	.2633	1.0443	.2973
11	2.9828	1.0368	5.1462	7.4245	5.2976	3.8564
12	6.0240	5.6314	5.7674	5.8057	8.4046	7.2634
TOTAL	1.2766	2.5745	1.2624	3.1213	1.4555	3.3172

In Table 5, which presents the means and standard deviations of the variance of the b's for the sample of 900, the trend is similar, although estimates show improvement with larger sample size. Figure 10 provides a graph of the means for the sample size 900. Similarly, Table 6, which presents the means and standard deviations of the variance of the b's for the sample of 1200 simulated responses, also expresses the trend shown for the two smaller sample sizes. Figure 11 is a graph of the results for the 1200 sample.

The overall pattern of Figures 9, 10, and 11, ignoring levels, shows that the middle difficulty items have the lowest variability. Variability gradually increases symmetrically as item difficulty increases or decreases.

Looking at Figure 9 little distinction can be made among the three levels of ability distribution for the middle difficulty items (item groups 5 to 7). This suggests that differences in ability distribution have little influence on the variation of estimates of these middle range items.

Differences among levels of ability distribution are more apparent with items that have either low or high difficulty values. Item groups 2, 3 and 4, for example, appear to mirror item groups 8, 9 and 10. Although the general pattern of increasing variability is about the same for these two groups, a subtle difference due to ability distribution may be detected.

For low difficulty items, the following pattern exists: level 3 has less variability than level 2, and level 2 less than level 1. For high difficulty items, the opposite pattern occurs. For item groups

Table 5

Means and Standard Deviations of V(b) for Item Groups
(n=900)

Item Group	Ability Level 1 mean	Ability Level 1 s.d.	Ability Level 2 mean	Ability Level 2 s.d.	Ability Level 3 mean	Ability Level 3 s.d.
1	.8547	.4144	.4125	.1394	.3696	.1068
2	.5555	.1693	.3535	.2748	.2051	.0904
3	.2729	.0810	.1610	.0673	.1479	.0322
4	.2029	.0896	.1357	.0784	.1314	.0803
5	.1374	.0578	.0965	.0577	.0893	.0275
6	.1602	.1504	.1464	.1471	.2516	.2528
7	.0866	.0158	.1060	.0659	.1344	.0805
8	.2900	.1072	.2397	.1512	.3021	.1305
9	.2709	.1679	.3443	.1878	.6968	.1879
10	.3332	.0934	.6222	.3434	.8482	.4968
11	1.4571	.9035	1.7837	1.0775	1.7271	1.1474
12	5.1282	8.1630	1.5997	1.2294	6.9539	9.6047
TOTAL	.8125	2.5406	.5001	.7173	.9881	3.1407

Table 6

Means and Standard Deviations of V(b) for Item Groups
(n=1200)

Item Group	Ability Level 1		Ability Level 2		Ability Level 3	
	mean	s.d.	mean	s.d.	mean	s.d.
1	.5515	.1761	.4592	.1105	.2981	.0510
2	.5834	.5782	.2293	.0711	.1756	.0784
3	.1649	.0445	.1315	.0402	.0930	.0874
4	.2188	.1838	.1096	.0556	.0914	.0390
5	.0750	.0326	.0544	.0277	.0643	.0191
6	.1523	.1448	.1088	.0893	.1498	.1109
7	.1335	.1249	.0938	.0541	.1234	.0335
8	.1679	.0703	.1988	.1251	.3862	.1293
9	.1902	.1128	.2365	.0928	.4562	.1441
10	.3159	.1320	.3314	.1698	.5995	.2389
11	.4175	.1610	1.5481	.8163	1.5232	.5466
12	3.1379	3.4065	4.7931	5.0666	6.6331	7.1019
TOTAL	.5051	1.2188	.6912	1.8708	.8828	2.5800

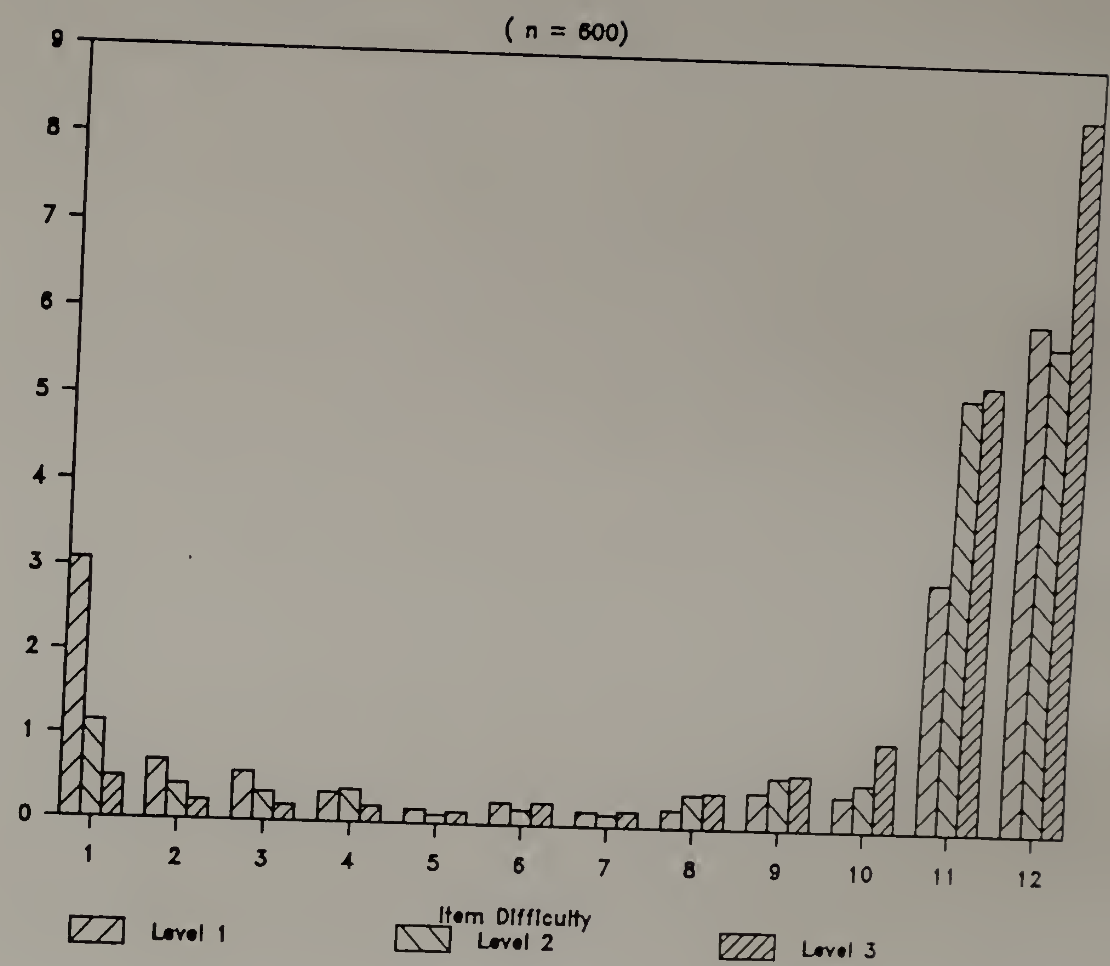

Figure 9. Relationship between V(b) and Item Difficulty Level of Item Groups.

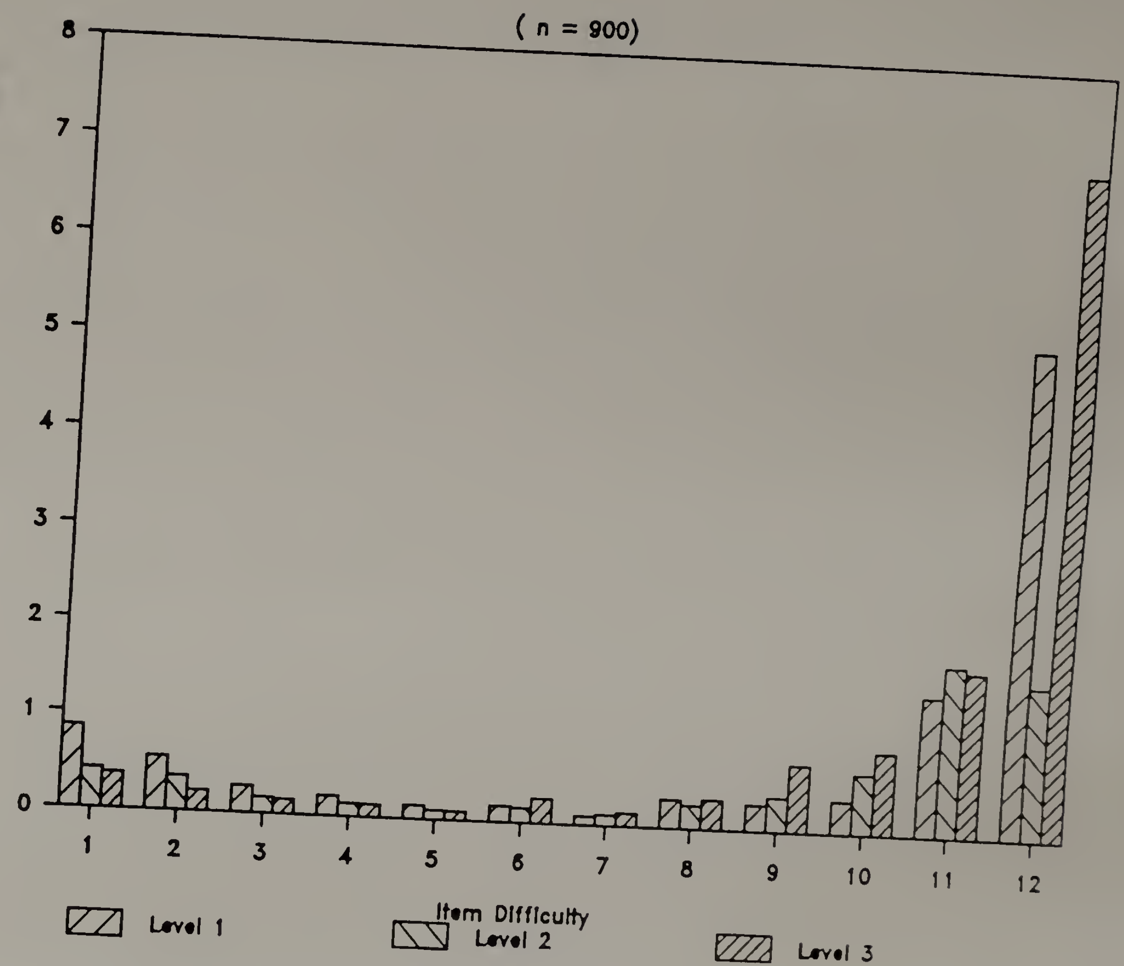

Figure 10. Relationship between V(b) and Item Difficulty Level of Item Groups.

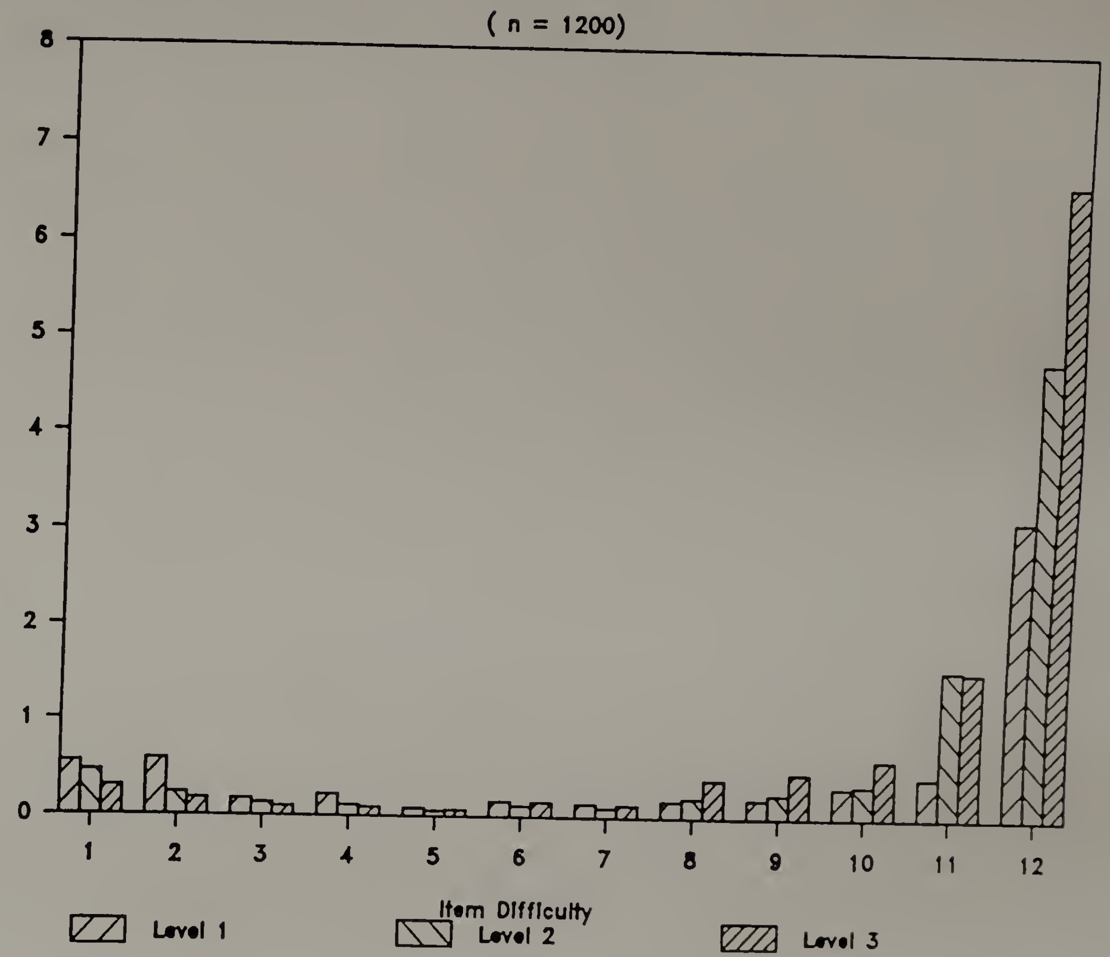

Figure 11. Relationship between V(b) and Item Difficulty Level of Item Groups.

8, 9 and 10, level 3 has the most variability, level 2 less than level 3 and level 1 less than level 2. This pattern among the levels also persists for the remaining items for the extreme left (item group 1) and to the extreme right (item groups 11 and 12).

This shift in variability of the estimates is interpreted to mean that variability among the estimate increases as a function of decreasing distributional density. That is, as the number of examinees decreases at the high range of ability, estimates for the extremely difficult items become more variable. For level 1, the normal distribution of ability, variance is low through the middle ranges, and gradually increases uniformally in both directions as the trails of the distribution thin out in both directions.

For ability level 2, the distribution is positively skewed. This yields low variance for item difficulty estimates that are in the middle range. Variance among the easier items is somewhat reduced as compared to the level 1 variability for the same items. For more difficult items, however, variability for level 2 is higher than variability for level 1.

For ability level 3 this pattern continues. Low difficulty items show decreased variability as the distributions move from level 1 to level 3. Higher difficulty items, to the right of the middle difficulty items, show a commensurate increase in variability as the distributions move from level 2 to level 3.

The general pattern described above is thought to be due to the relative density of the ability distributions. Where density is low (i.e., sample size is small), there is more variance over estimates obtained from samples. Where density is high (i.e., sample size is large), variance is reduced.

These results appear to support earlier work by Mooney and Swaminathan (1986) which explored the quality of b parameter estimation for restricted ability ranges. It is evident from this study that accuracy was not as good for restricted ability ranges. As the ability distribution becomes positively skewed (level 2 and Level 3), more difficult items are less well estimated. The means of Table 3 are also presented in graph form in Figure 9.

The patterns described for variance also holds true for accuracy and bias. Accuracy is presented in Tables 7, 8 and 9 and appear in graph form in Figures 12, 13 and 14. Although the pattern of movement across the axis of the three distributional levels is somewhat less clear than was the pattern for item variance, it is nevertheless still apparent.

The pattern for item bias is the least clear, particularly for the smallest sample size. This information is given in Tables 10, 11 and 12, and is repeated in graph form in Figures 15, 16 and 17. Similarly the pattern of accuracy MSD(b) and item bias B(b) indicates that they are both influenced by the distribution of ability in a manner that echoes the pattern established in the analysis of accuracy. This result is important because it demonstrates that estimates for extreme items are not only more variable, but biased as well.

Table 7

Means and Standard Deviations of MSD(b) for Item Groups
(n = 600)

item group	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
1	5.7343	8.6887	1.2227	.4120	.8293	.4753
2	.9987	.3715	.9992	.7526	.7283	.6068
3	.8736	.6087	.8504	.2322	.4093	.3457
4	.4738	.2708	.4884	.4578	.3405	.1641
5	.5046	.2067	.2167	.1883	.3811	.2110
6	.8173	1.2817	.2214	.1500	.5773	.7636
7	.2301	.1800	.1885	.0470	.3930	.2220
8	.3146	.0892	.4708	.2489	1.3246	.4981
9	.7985	.3055	.8525	.5775	1.5091	1.4752
10	.9379	.2086	1.1311	.5802	1.2481	.4394
11	3.6411	1.3306	8.3742	8.7456	7.2441	4.0847
12	9.0330	7.1301	6.6257	4.7866	14.0488	10.5489
TOT:	2.0298	3.9842	1.8035	3.7025	2.4195	4.9824

Table 8

Means and Standard Deviations of MSD(b) for Item Groups
(n = 900)

item group	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
1	2.2343	2.0641	.4461	.1300	.7307	.5373
2	.7226	.2538	.6255	.4747	.4698	.1579
3	.6424	.4067	.2653	.1226	.3139	.1079
4	.5887	.3428	.1757	.1183	.3234	.2412
5	.3502	.3729	.1728	.1097	.5194	.6860
6	.3073	.2293	.3437	.2976	.9650	1.6432
7	.1721	.0499	.1267	.0681	.5272	.3335
8	.4901	.1531	.6542	.7200	.8177	.4475
9	.3814	.1851	.6289	.3823	1.2802	.6183
10	.4215	.0655	1.2603	.8666	1.4476	.6609
11	1.9756	.7338	2.6378	1.4935	3.2079	1.6015
12	7.6694	9.7059	5.7245	5.3239	16.3373	25.2613
TOT:	1.3296	3.2971	1.0885	2.1525	2.2450	7.9400

Table 9

Means and Standard Deviations of MSD(b) for Item Groups
(n = 1200)

item group	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
1	1.0573	.3808	.9531	.6728	.5514	.2540
2	1.4054	1.6062	.6243	.4617	.4959	.2079
3	.3233	.1930	.2379	.0909	.3167	.1368
4	.5921	.5662	.1960	.0871	.1509	.0616
5	.1570	.0726	.1838	.1333	.2496	.1939
6	.2669	.2221	.3424	.4730	.4516	.5440
7	.2058	.1392	.1623	.0763	.1778	.0628
8	.4818	.4636	.4457	.4879	.5942	.2346
9	.4204	.2051	.3131	.1407	.6599	.1580
10	.4513	.1717	.8273	.3677	1.7580	.5978
11	1.6137	1.2756	2.4459	.7699	3.5079	1.6427
12	6.5821	10.7605	8.4969	10.4122	11.5667	15.6564
TOTAL	1.1298	3.3395	1.2691	3.5614	1.7067	5.1665

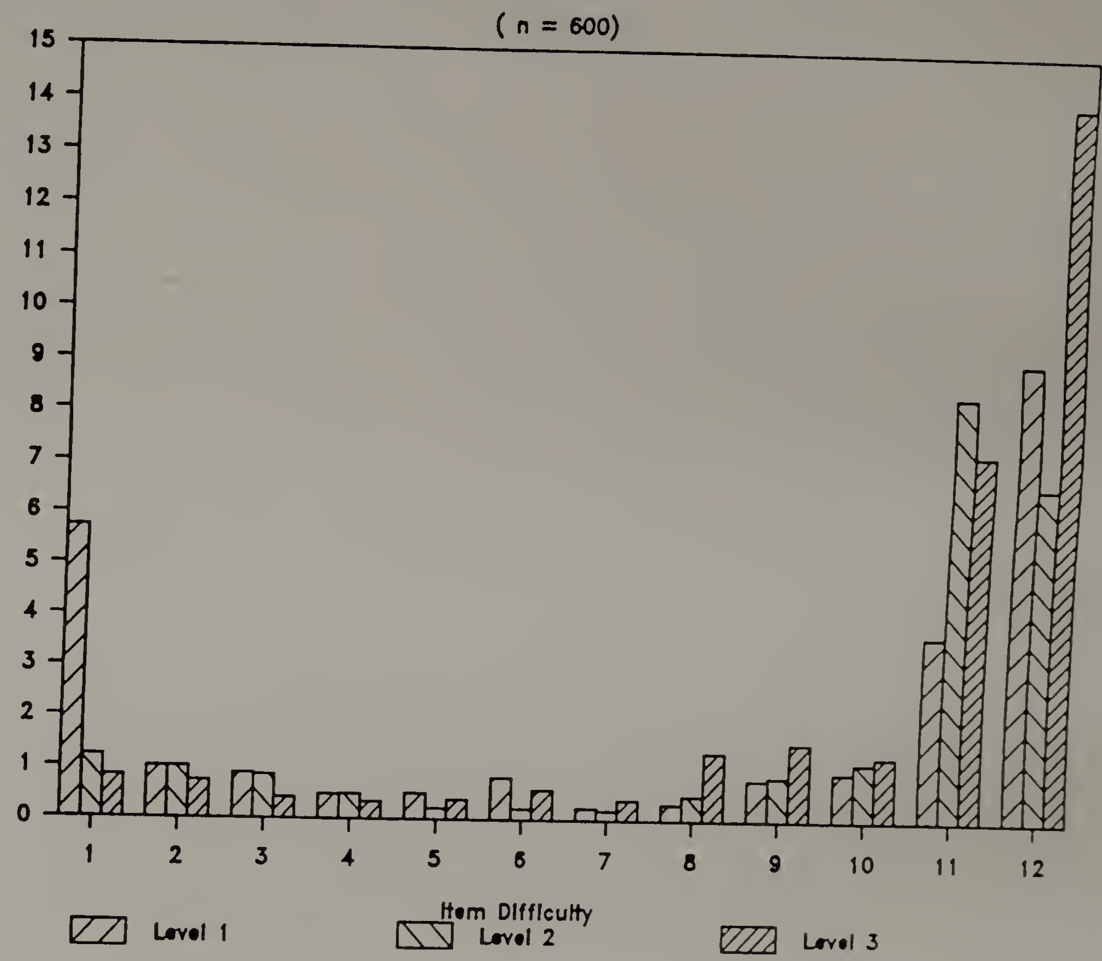

Figure 12. Relationship between MSD(b) and Item Difficulty Level.

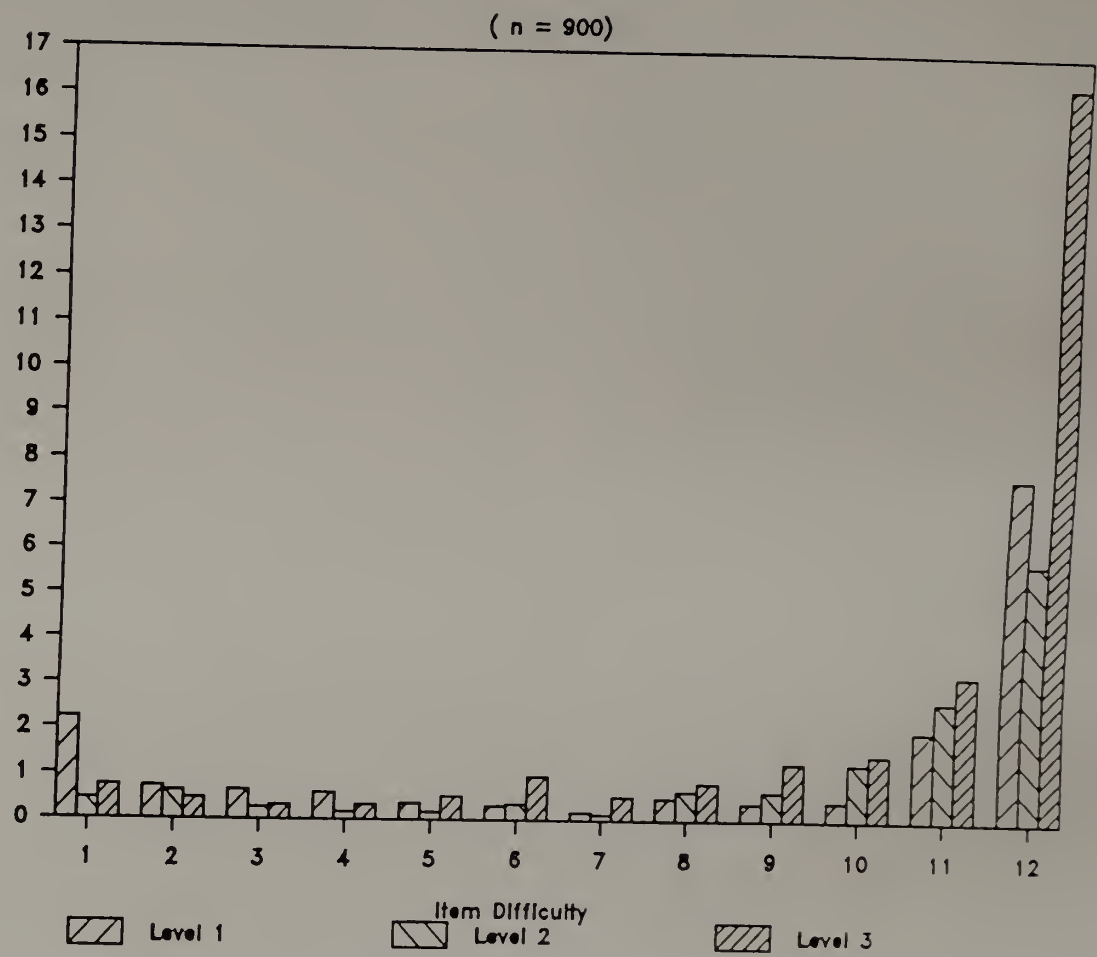

Figure 13. Relationship between MSD(b) and Item Difficulty Level.

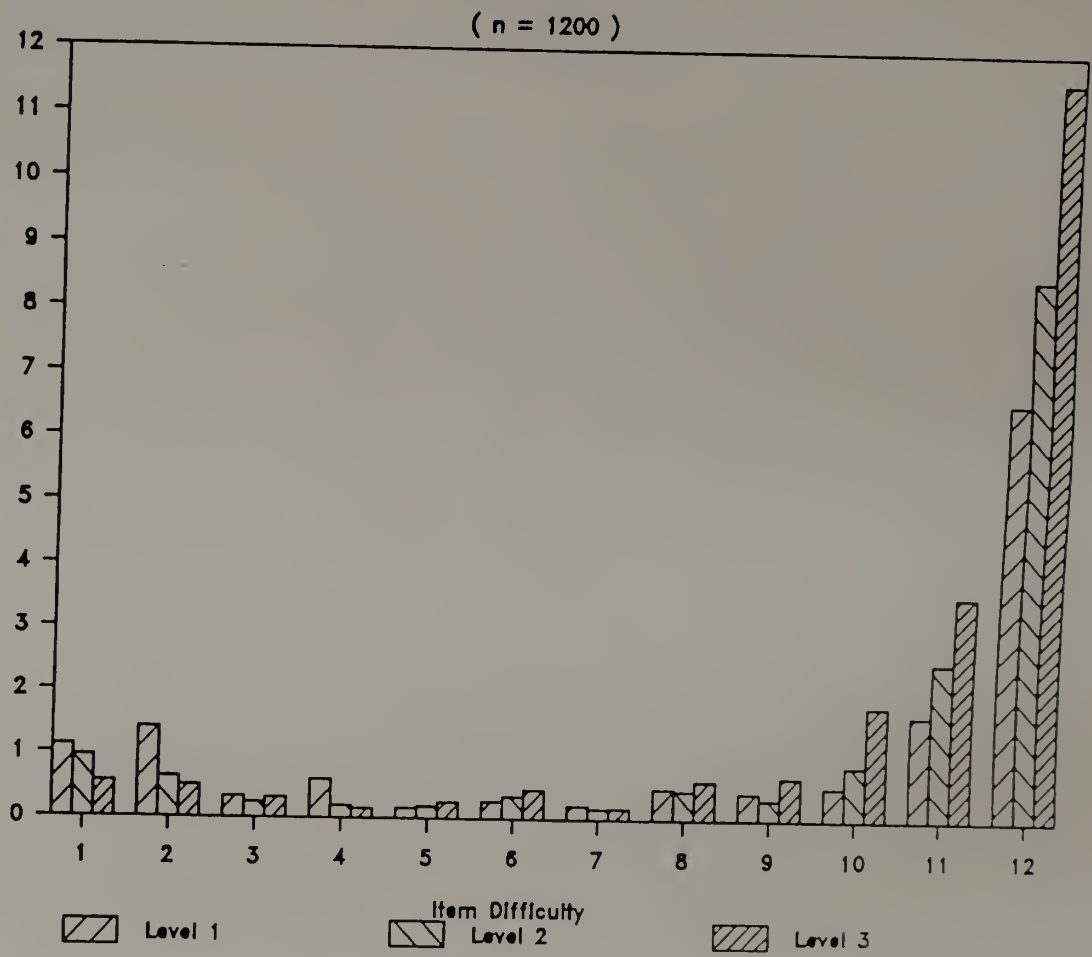

Figure 14. Relationship between MSD(b) and Item Difficulty Level.

Table 10

Means and Standard Deviations of B(b) for Item Groups
(n=600)

Item Group	Level 1		Level 2		Level 3	
	mean	s.d.	mean	s.d.	mean	s.d.
1	2.6559	4.5145	.0754	.0625	.3389	.4675
2	.3130	.2419	.5892	.8819	.5081	.6040
3	.3093	.3223	.5166	.3719	.2279	.3622
4	.1326	.0891	.1095	.1318	.1532	.1222
5	.3420	.2050	.1232	.1834	.2445	.2364
6	.5595	1.2344	.0444	.0619	.3127	.4710
7	.0559	,0459	.0494	.0560	.2127	.2441
8	.1035	.1035	.0764	.0595	.9114	.5544
9	.3666	.4228	.2428	.2900	.8639	1.5173
10	.5332	.3671	.5791	.6064	.2035	.2222
11	.6543	.6091	3.4280	2.7562	1.9465	2.0976
12	4.2090	4.5760	2.0583	3.8868	5.6442	4.1757
TOTAL	.8532	2.1102	.6577	1.6250	.9640	2.0003

Table 11

Means and Standard Deviations of B(b) for Item Groups
(n=900)

Item Group	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
1	1.3796	1.6772	.0336	.0394	.3611	.5913
2	.1671	.3238	.2720	.2361	.3611	.5413
3	.3696	.3789	.1042	.1330	.1660	.1176
4	.3858	.2900	.0400	.0649	.1920	.1769
5	.2128	.3632	.0763	.1038	.4301	.6724
6	.1470	.1708	.1973	.2897	.7313	1.4392
7	.0855	.0493	.0208	.0168	.3928	.3392
8	.2001	.0732	.4146	.6100	.5156	.4243
9	.1105	.1028	.6381	.7034	.6014	.5658
10	.0883	.1028	.6381	.7034	.6014	.5658
11	.5185	.7164	.8541	1.0060	1.4808	1.7990
12	2.5392	3.2420	4.1247	5.7249	9.3834	16.0596
TOTAL	.5170	1.2124	.5884	1.8940	1.2585	4.9181

Table 12

Means and Standard Deviations of B(b) for Item Groups
(n=1200)

Item Group	Level 1		Level 2		Level 3	
	mean	s.d.	mean	s.d.	mean	s.d.
1	.5624	.3625	.4939	.6901	.2534	.2933
2	.8221	1.0654	.3950	.4159	.3203	.2018
3	.1584	.2092	.1064	.0751	.2237	.1741
4	.3734	.4574	.0864	.0711	.0595	.0470
5	.0820	.0589	.1294	.1293	.1854	.1862
6	.1146	.1886	.2336	.4075	.3018	.5066
7	.0723	.0995	.0685	.0860	.0544	.0455
8	.3139	.4013	.2469	.3756	.2080	.1730
9	.2302	.1610	.0767	.0926	.2038	.0941
10	.1986	.1698	.4959	.3903	1.1584	.8008
11	1.1962	1.2356	.8978	1.0342	1.9847	1.5448
12	3.4442	7.4494	3.7045	5.4521	4.9337	8.7747
TOTAL	.6246	2.1969	.5779	1.7680	.8239	2.7047

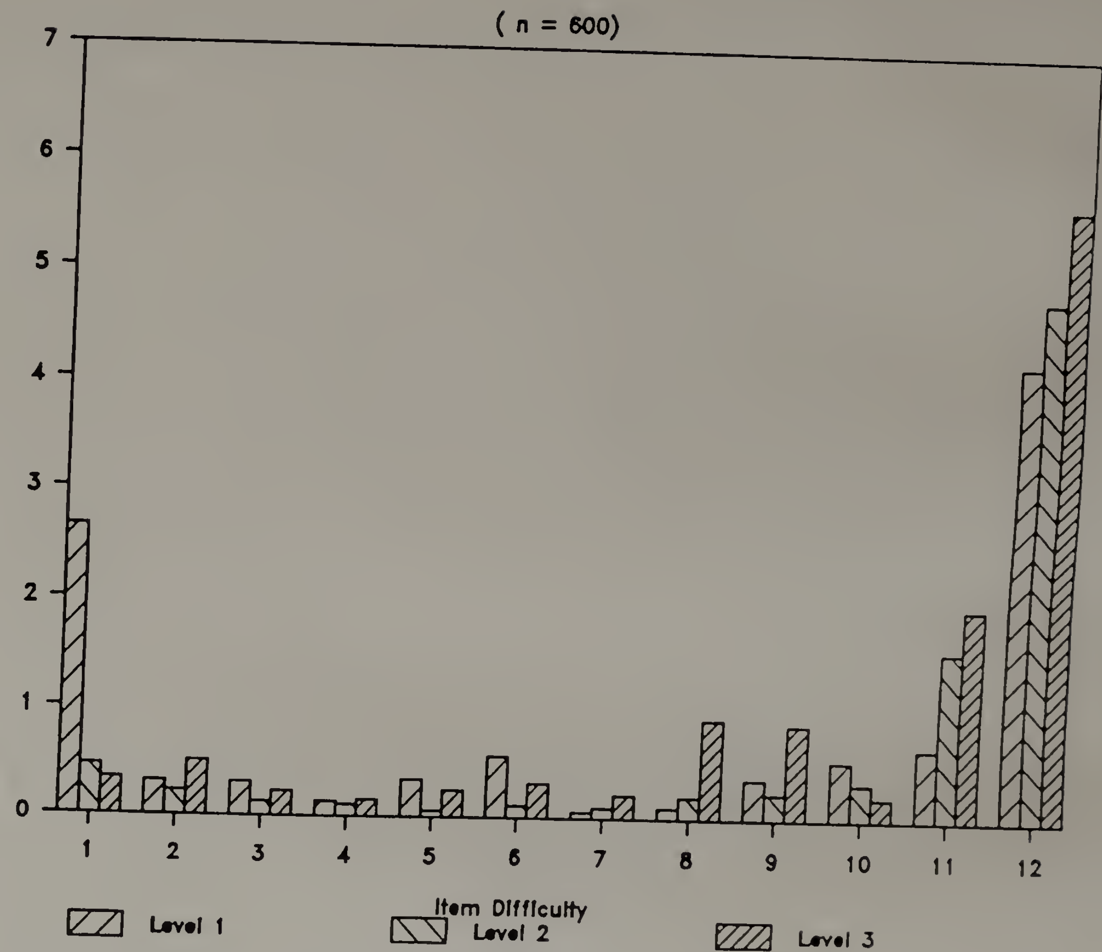

Figure 15. Relationship between B(b) and Item Difficulty Level.

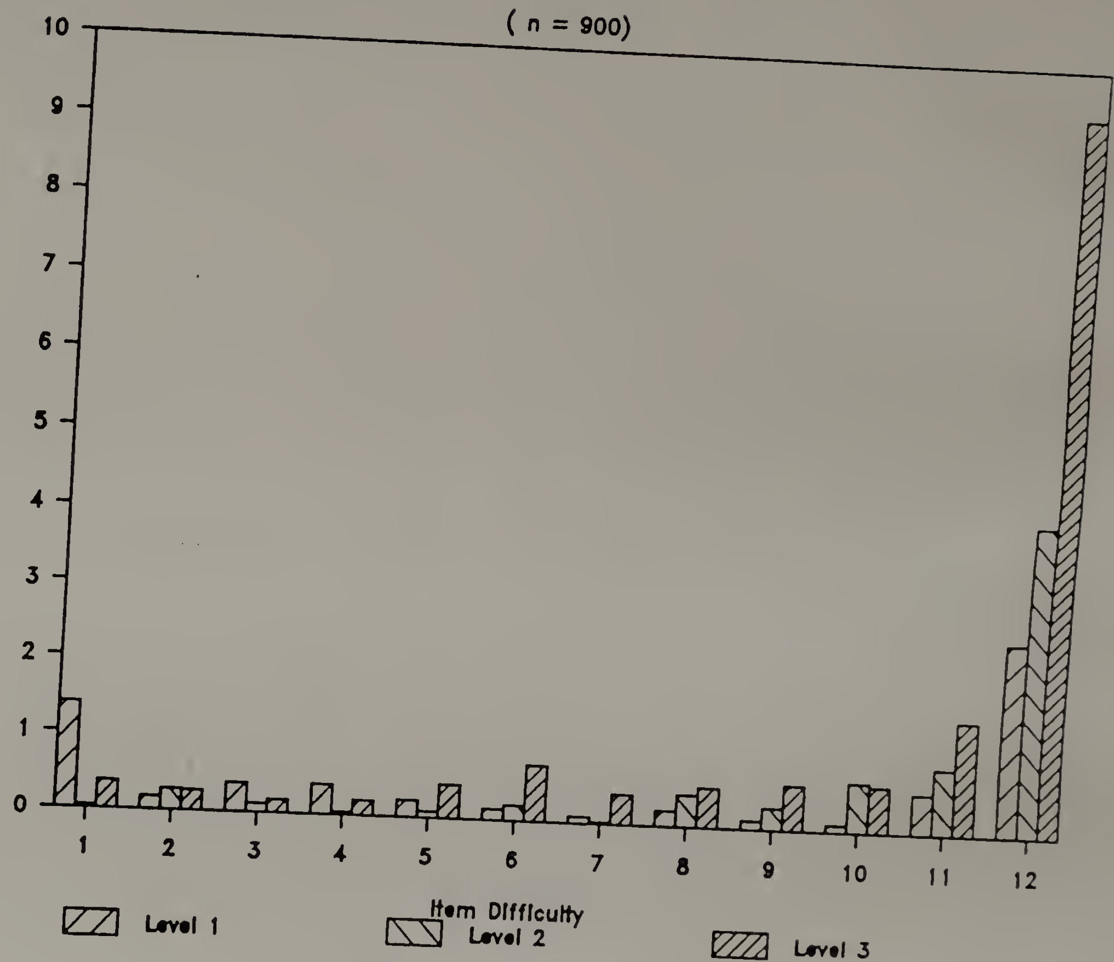

Figure 16. Relationship between B(b) and Item Difficulty Level.

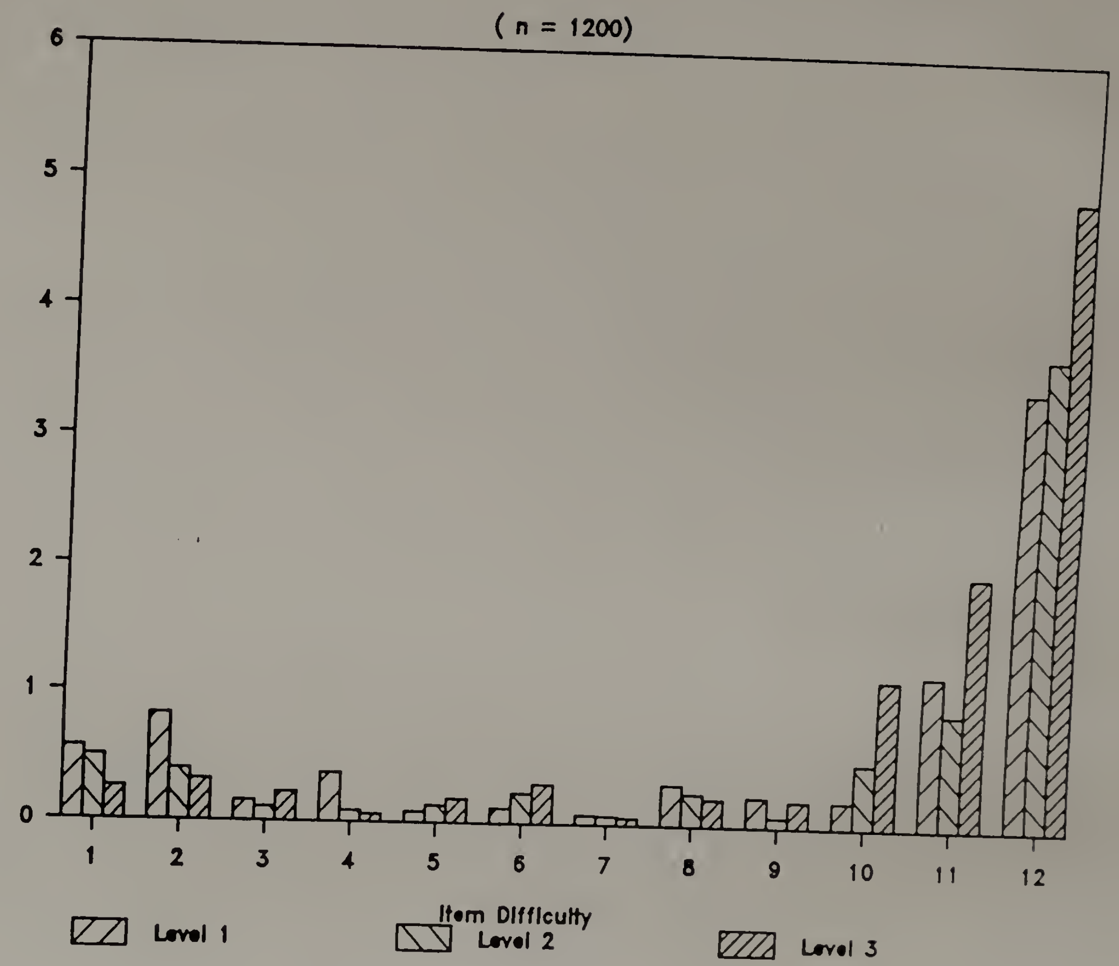

Figure 17. Relationship between B(b) and Item Difficulty Level.

Overall Fit from the Prospective of Item Accuracy

Figure 18 looks at the means and standard deviations of the accuracy index, MSD(b) for the 9 cells. Mean and standard deviation of accuracy averaged on the 12 difficulty levels of MSD(b) are taken here to indicate a global measure of fit. As in the case of the previous analysis of accuracy, a lower accuracy score means that estimates are close to one another over replications.

One interesting finding is that overall fit appears to be best for level 2, rather than for level 1. One possible explanation for this is that the somewhat skewed distribution of ability provides better fit for the three parameter model because there are more subjects of lower ability in the skewed distribution that may provide better c parameter estimates. This same phenomenon can be observed in the graphs of item accuracy. Across all three sample sizes, it can be seen that ability level 2 estimates are nearly always consistently best for the middle ranges of item difficulty.

The above observation raises the question of why ability level 3 does not show a commensurate increase in accuracy over level 2. Level 3 may provide somewhat better accuracy over the easier items as compared to level 2, but may provide a disproportionate decrease in accuracy for the more difficult items. The improved accuracy obtained with the easier items is probably not sufficient to outweigh the decrease in accuracy obtained with the more difficult items.

This effect may be explained by the differences among the ability distributions. Figures 2, 3 and 4 show the differences among the ability distributions.

Distribution of Ability	Number of Examinees: n = 600	n = 900	n = 1200
1	U = 2.0298 SD = 3.9842	U = 1.3296 SD = 3.2971	U = 1.1298 SD = 3.3395
2	U = 1.8035 SD = 3.7025	U = 1.0885 SD = 2.1525	U = 1.2691 SD = 3.5614
3	U = 2.4195 SD = 4.9824	U = 2.2450 SD = 7.9400	U = 1.7067 SD = 5.1665

Figure 18. Means and Standard Deviations of Accuracy, MSD (b).

As skewness increases positively the items become more and more difficult for the group. For the ability level 1 distribution of ability, 50% of the distribution falls to the right of interval 10. For the level 2 distribution, 37% of the distribution falls to the right of interval 10. For the level 3 distribution only 10% of the distribution falls to the right of interval 10. The consequence of this is that variability of the estimates for the difficult items increases disproportionately from level 1 to level 3.

One important implication is that b parameter estimates are influenced by the skewness of the ability distribution. Usually poor estimation of the c parameter may be expected to have an influence on the b parameter estimates. However, in this case, positively skewed distributions of ability were chosen so as to obtain good estimates of the c parameter. The poor estimates of the b parameter must therefore be the result of the influence of skewness of the ability distribution itself.

In addition, it is also notable that occasionally a middle range item or two becomes unstable (see Appendix C). This problem may be attributable to artifacts of estimation using LOGIST4.

Conclusions

The results of this study demonstrate that the accuracy of estimation of extremely easy or extremely difficult items is influenced by restrictions in the range of ability. Invariance, based on the test of stability, appears to hold. One concern with this method of assessment, however, is that it may not be sufficiently sensitive to detect lack of invariance. In addition, it was shown that the three parameter IRT model appeared to obtain relatively better overall accuracy when a somewhat positively skewed distribution was used. This result was attributed to better quality estimation of the guessing parameter.

In general, the three parameter IRT model generally performed well through the middle item difficulty ranges. However, within each of the 9 testing situations, one or two of the middle range items demonstrated some degree of inaccuracy. This was attributed to be the result of artifacts of estimation using LOGIST4.

CHAPTER V

CONCLUSIONS AND IMPLICATIONS

Review

The chief advantage of IRT over classical test theory is that the item parameters are invariant. The purpose of this dissertation is to explore the quality of the estimation of these item difficulty parameters and its effect on the detection of the property of invariance.

One of the most direct ways of assessing the invariance property is to compare the item difficulty estimates for different groups. For example, in item bias studies the technique often used is to evaluate the scatterplots of the item difficulty estimates. When some item estimates fall beyond the degree of scatter displayed by the majority of the estimates then those items are flagged as not invariant across groups and studied for possible bias. It may be that sampling error varies widely from one item to another, such that an item that appears to be not invariant may simply be an item with greater sampling error. Therefore this method that does not take sampling error into account may not be adequate.

In addition to the sampling error issue model-data fit poises another problem. To date, no sure method of assessing model-data fit exists. It is not known to what degree expected features such as invariant item parameters may be obtained in circumstances where model-data fit is not perfect. Finally, range restriction of ability and fluctuations in sample size may be expected to influence the quality of parameter estimation and hence the property of invariance. Because range restriction, sample size, and model-data fit concerns exist in every IRT application and may well be confounded with one another, it is difficult to assess the influence of each of these factors individually.

In order to investigate sampling error and its effect on invariance in greater detail, thirty samples were taken for each of the nine testing situations that vary over range restriction and sample size. Item sets could then be compared over each of the nine testing situations for stability and for accuracy of estimates. If range restriction were not an issue, it would be expected that variance among parameter estimates would not change over ability distributions.

Simulated data were used for this dissertation primarily because population parameters could be known. A second advantage of simulated data is as a control for model-data fit and also for bias. Although model-data fit or lack of item bias cannot be established even with simulated data, this approach provides a reasonable intuitive basis for this.

The strategy for this dissertation was to evaluate the extent to which repeated estimates obtained from samples with differing ability distributions and sample sizes would recover the true values for these parameters. The hypothesis was that estimation would not be influenced by changes in the ability distribution because of the invariance property.

The research questions for this study were:

1. How does range of ability affect the invariance of the estimates of the difficulty parameters in the three parameter IRT model?

2. What is the influence of sample size on the invariance of the estimates of the difficulty parameters?

3. What is the consequence of interaction of range of ability with sample size?

To evaluate these questions three different levels of ability and three different sample sizes were completely crossed for a total of nine testing situations. For each testing situation, response patterns were sampled to fit the required specifications for range restriction and sample size. The data for each of these nine testing situations were then replicated thirty times using sampling with replacement and estimates of the item difficulty parameters were obtained. The degree to which parameters obtained stability and the accuracy of estimation were studied.

Conclusions

The main conclusion of this dissertation is that the ability to establish invariance depends upon the quality of estimation of parameters. Through sampling with replacement it was shown that sampling error was a function of the ability distribution. Estimates for extremely difficult or extremely easy items that were obtained with relatively few subjects of extreme ability levels showed greater variability than estimates obtained where there were more subjects at the appropriate ability level for a given item. This conclusion was confirmed by studying the accuracy of the estimation. Estimates of item difficulty parameters for easy and for difficult items showed more sampling fluctuation and were clearly affected by the distributions of ability.

It was also shown that overall model data fit for a given test was improved when sufficient low ability subjects were available. This was attributed to better model-data fit for the three-parameter IRT model, where a guessing parameter is estimated.

In applications of IRT much has been made of the importance of large sample size. This study has shown that large sample size alone is not sufficient to ensure proper estimation of parameters. There must be enough subjects at each ability level in order to be sure of proper estimation of parameters. In turn, when the item parameters are estimated properly important features such as invariance can be assertained.

Implications

An important conclusion from this study is that extreme range restriction influences accuracy of estimation. This would be an issue when IRM's are applied to cases where ability distributions are apt to be skewed, as in the case of Criterion Referenced Testing (CRT), and would be further exacerbated when CRT examinee samples are not large.

Problems may be anticipated in all IRT applications where the expected feature of parameter invariance is applied without taking into account the accuracy of estimation for the extreme items. Problems may arise in item banking, for example, because items at the extreme ranges of difficulty may not be well estimated. In the case of building a test to determine the best candidates for a scholarship, for example, a high proportion of difficult items would be chosen for such a test from the item bank. However, this study has demonstrated that the parameter estimates for such items may not possess the high degree of accuracy that might be available from items selected from more moderate ranges of difficulty.

In the case of traditional item bias studies where only two groups are compared, estimates may look different and therefore flagged as biased, when, in fact, the estimates may be within the range of the sampling error. One possible solution to this problem is to be sure, when estimating the item parameters for items, that candidates in the appropriate ability range for the level of item difficulty are well represented.

Another issue noted is that some items in the middle range of ability appear to go out-of-bounds. This could be an artifact of estimation using LOGIST4. One obvious concern here is the possiblity that such an item or items may be interpreted as biased.

Finally, there is the question of model choice. Results from this study indicate better fit for the three parameter model when a positively skewed distribution is used. This is interpreted to reflect the applicability of the three parameter model to cases where sufficient low ability examinees are available. In cases where sufficient low ability examinees are not available, the appropriateness of the three parameter model may be in question. This finding supports the idea that range restriction may have impact on parameter estimation for IRT models.

This study demonstrates that the resampling method is useful for providing empirical evidence of the consistancy of parameter estimates. An important drawback of this method, however, is that it is expensive and time consuming. Therefore, this method would seem applicable only in those cases where these issues are critical. Item parameter invariance does work for most items, however items that behave very badly could be investigated using this method.

Another potential application for this method of resampling is in model-data fit. Using two standard deviations from the mean of the b parameter estimates as a benchmark, LOGIST4 estimates were well behaved using data generated from DATAGEN. This finding may have utility for the examination of field data, where no known method of establishing model data fit exists.

The techniques demonstrated in this dissertation could be used to establish model data fit and also for item bias detection. To establish model data fit, repeated random samplings of the total sample of examinees could be fit to the chosen item response model. B parameter estimates should be transformed to mean zero and unit variance. B parameter estimates could then be grouped by item and transferred to mean zero and unit variance once again. When the model fits the data, transformed estimates should fall within the expected range of normal deviates (i.e., 68% of estimates within one standard deviation and 95% within two standard deviations).

Item bias may be investigated by comparing the accuracy of repeated random samplings of b parameter estimates from the group for whom bias may be considered a possible concern, to estimates obtained from random samplings from a similar ability group sample.

The accuracy of the standardized estimates from both groups may be compared for each item. If the accuracy of each group's estimates for a given item are about the same then bias is probably not a serious concern. If the accuracy is not about the same, then perhaps the item should be carefully investigated for possible bias. However, if other items appearing to show bias are from the extremes of the difficulty scale they should be looked at carefully. These estimates may be relatively unstable because of range restriction alone, and not necessarily because of bias.

One concern about the approach described above is that the ability distributions of the two groups should be compared using raw scores to see that they are reasonably comparable. If these distributions are grossly unalike, this will probably also influence parameter estimation.

In summary, a useful method of examing the stability of item parameters has been demonstrated, and this method may prove useful in the context of item bias investigations. It was also shown that different levels of ability distributions influenced the estimation of extremely difficult or extremely easy items. Further work is necessary, however, to characterize these issues in more detail. One possibility might be to investigate the accuracy of the difficulty estimates of extreme items using a uniform distribution. The accuracy of estimates from the uniform distribution could then be compared to accuracy estimates derived from ability distributions with different levels of tail thicknesses.

In terms of conventional item bias studies using IRT models, items showing bias when the items are either extremely easy or extremely difficult ought to be investigated with care. It could be that such items are influenced by small sample size and this may account for the apparent invariance in the difficulty estimates.

Further work also needs to be done with the two parameter model, especially in cases where few low ability examinees exist in the sample. It may be, for example, that the two parameter model would provide more accurate fit in cases where the ability distribution is approximately normal, whereas the three parameter model may provide better fit in cases where the ability distribution is positively skewed.

APPENDIX A

Table 13

Mean Scores and Standard Deviations of B Value Differences
(n = 600)

item	rank	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
4	1	-.593	.606	.004	.164	-.041	.139
55	2	.283	.097	-.055	.238	.106	.170
44	3	.067	.218	.038	.210	.195	.231
40	4	.012	.140	.049	.161	-.029	.134
22	5	.082	.294	.075	.211	.067	.167
35	6	.076	.133	-.044	.144	.038	.164
12	7	.094	.205	-.052	.151	.147	.138
29	8	.132	.166	.263	.087	.224	.149
7	9	-.020	.159	.155	.085	.084	.110
8	10	.140	.078	.016	.112	.067	.123
56	11	-.068	.147	.157	.082	.058	.097
9	12	-.160	.195	-.098	.107	-.049	.113
48	13	.132	.121	.110	.141	-.020	.099
16	14	-.028	.117	.187	.058	.171	.100
47	15	.055	.097	.069	.127	.053	.071
3	16	.052	.095	-.011	.064	.092	.107
59	17	.064	.164	-.081	.171	.096	.077
36	18	.093	.088	-.005	.094	.078	.087
31	19	.076	.102	-.099	.146	-.016	.085
45	20	.029	.070	.041	.039	.038	.053
26	21	-.038	.075	-.032	.040	-.109	.075
27	22	.110	.062	.025	.046	.011	.053
14	23	.095	.085	-.055	.078	.023	.084
21	24	.141	.079	.031	.060	.137	.075
41	25	.120	.072	.122	.052	.098	.050
15	26	.304	.100	.008	.125	.194	.117
50	27	-.002	.062	-.025	.056	-.063	.043
28	28	-.030	.155	.041	.077	-.014	.087
39	29	.004	.062	-.007	.050	-.020	.039
23	30	-.008	.052	.071	.056	-.100	.053
11	31	.053	.045	.056	.048	-.096	.055
57	32	.025	.055	-.064	.060	.070	.059

Table 13 (continued)

item	rank	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
19	33	.015	.136	.007	.092	-.142	.057
43	34	.062	.059	.008	.078	-.007	.055
18	35	-.042	.054	-.030	.060	-.034	.046
34	36	-.012	.089	-.062	.061	.134	.070
60	37	-.074	.085	-.011	.174	-.227	.115
2	38	.046	.103	.072	.105	-.201	.091
10	39	.092	.069	-.043	.113	-.074	.098
20	40	-.033	.077	-.041	.101	-.191	.118
6	41	.026	.143	-.157	.102	.059	.078
37	42	-.077	.092	-.068	.114	-.048	.065
13	43	-.182	.089	-.007	.125	.097	.085
5	44	-.142	.084	-.057	.112	-.099	.100
32	45	-.034	.175	.089	.231	-.345	.138
25	46	-.129	.084	-.028	.093	-.022	.102
46	47	-.158	.103	-.189	.136	-.032	.123
49	48	-.185	.083	.112	.153	-.068	.082
58	49	-.050	.171	-.046	.178	-.097	.115
42	50	-.104	.126	-.213	.113	.136	.108
1	51	-.112	.227	-.167	.287	-.001	.175
54	52	.164	.299	.042	.211	-.270	.159
38	53	.022	.371	.405	.796	-.146	.248
30	54	-.125	.316	.442	.291	.225	.184
24	55	-.233	.368	-.427	.204	-.424	.316
17	56	-.131	.237	.162	.482	.563	.277
53	57	-.272	.312	-.009	.285	-.484	.244
52	58	.280	.611	.010	.190	-.051	.226
51	59	.452	.679	-.547	.209	.551	.134
33	60	-.358	.227	-.131	.628	-.289	.824

Table 14
Mean Scores and Standard Deviations of B Value Differences
(n = 900)

		Level 1		Level 2		Level 3	
item	rank	mean	s.d.	mean	s.d.	mean	s.d.
4	1	.010	.139	.003	.110	-.024	.132
55	2	-.125	.170	.018	.128	.107	.090
44	3	-.363	.231	-.013	.133	.215	.101
40	4	.096	.134	.052	.082	.039	.114
22	5	.270	.167	-.049	.135	.017	.123
35	6	-.043	.164	-.147	.167	.048	.079
12	7	.029	.138	.057	.113	.098	.088
29	8	-.012	.149	.051	.101	.130	.078
7	9	.158	.110	.107	.078	.043	.109
8	10	-.012	.123	.079	.063	.116	.059
56	11	.069	.097	.106	.061	.094	.064
9	12	-.078	.113	-.048	.093	.007	.069
48	13	.123	.099	-.026	.086	-.054	.082
16	14	.182	.100	.043	.061	.095	.065
47	15	.049	.071	.034	.065	.083	.076
3	16	-.134	.107	-.007	.060	.045	.049
59	17	.162	.077	-.035	.094	.097	.093
36	18	.102	.087	.071	.074	-.066	.063
31	19	.097	.085	-.005	.061	.124	.075
45	20	.021	.053	.017	.041	.032	.043
26	21	-.168	.075	-.093	.051	-.062	.055
27	22	.007	.053	.039	.032	.048	.045
14	23	-.021	.084	.013	.063	-.086	.047
21	24	.009	.075	.026	.052	.062	.064
41	25	.082	.050	.042	.080	.233	.063
15	26	-.068	.117	-.002	.056	.332	.154
50	27	-.013	.043	-.088	.042	.001	.055
28	28	-.062	.087	-.027	.118	-.052	.096
39	29	-.036	.039	.038	.045	-.017	.052
23	30	.121	.053	.151	.066	.094	.071
11	31	.064	.055	.017	.055	.070	.097
57	32	-.016	.059	-.024	.040	.010	.047

Table 14 (continued)

item	rank	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
19	33	-.059	.057	.027	.087	-.135	.056
43	34	.064	.055	.040	.053	-.170	.064
18	35	-.047	.046	-.015	.057	-.116	.066
34	36	.078	.070	.012	.067	.029	.093
60	37	.105	.115	-.027	.101	-.086	.113
2	38	.075	.091	.215	.128	.170	.130
10	39	.070	.098	-.020	.079	-.121	.090
20	40	.075	.118	-.147	.062	-.185	.075
6	41	-.083	.078	-.066	.095	-.151	.133
37	42	-.032	.065	-.063	.090	-.136	.141
13	43	.039	.085	.071	.085	-.054	.147
5	44	-.082	.100	.071	.152	-.228	.164
32	45	.049	.138	-.170	.109	.032	.184
25	46	.040	.102	.068	.159	-.222	.153
46	47	-.019	.123	.232	.186	-.168	.152
49	48	-.092	.082	-.092	.070	-.104	.128
58	49	-.061	.115	-.196	.120	.102	.244
42	50	-.020	.108	.032	.168	-.044	.155
1	51	.048	.175	.087	.239	-.217	.166
54	52	-.236	.159	-.259	.111	-.168	.157
38	53	.055	.248	-.022	.231	-.131	.313
30	54	-.159	.184	.251	.340	-.391	.202
24	55	.000	.316	-.065	.263	.043	.326
17	56	.008	.277	.059	.165	-.159	.253
53	57	-.115	.244	.247	.351	.072	.444
52	58	-.111	.226	.279	.259	.073	.212
51	59	-.488	.134	-.689	.134	-.530	.266
33	60	.400	.824	-.264	.199	1.117	.907

Table 15
Mean Scores and Standard Deviations of B Value Differences
(n = 1200)

		Level 1		Level 2		Level 3	
item	rank	mean	s.d.	mean	s.d.	mean	s.d.
4	1	.125	.112	-.085	.147	.069	.099
55	2	.153	.104	.044	.118	-.022	.103
44	3	.064	.134	.129	.107	.154	.091
40	4	.180	.153	-.237	.122	.028	.114
22	5	.122	.154	-.019	.132	.112	.099
35	6	.113	.101	.086	.080	-.111	.102
12	7	-.299	.233	.128	.108	.083	.086
29	8	-.036	.143	.190	.095	.144	.066
7	9	.149	.098	.055	.078	.051	.074
8	10	.105	.076	.056	.080	.105	.051
56	11	.118	.076	.081	.075	.126	.042
9	12	.019	.077	.021	.057	.019	.071
48	13	-.017	.090	.040	.077	.068	.070
16	14	.109	.063	.065	.070	.083	.049
47	15	-.004	.069	.070	.053	.098	.045
3	16	-.061	.084	.077	.059	.065	.045
59	17	-.019	.095	.052	.079	.051	.063
36	18	-.178	.132	-.022	.072	.037	.068
31	19	.162	.050	.067	.051	.040	.061
45	20	.015	.040	-.027	.038	-.008	.038
26	21	-.066	.045	-.105	.042	-.066	.036
27	22	.040	.034	-.020	.034	.061	.043
14	23	.029	.061	.051	.059	-.025	.052
21	24	.073	.061	.038	.035	.072	.050
41	25	.040	.048	.078	.041	.130	.053
15	26	.004	.069	.179	.088	.084	.107
50	27	-.122	.055	-.040	.031	.000	.048
28	28	-.051	.118	.020	.079	-.200	.076
39	29	-.031	.046	-.031	.048	-.017	.047
23	30	.025	.049	.063	.037	.056	.064
11	31	-.040	.059	.017	.042	-.008	.059
57	32	-.035	.048	-.084	.037	-.037	.053

Table 15 (continued)

item	rank	Level 1 mean	Level 1 s.d.	Level 2 mean	Level 2 s.d.	Level 3 mean	Level 3 s.d.
19	33	-.031	.111	.012	.075	-.046	.075
43	34	.006	.049	-.051	.069	-.065	.064
18	35	-.091	.051	-.036	.051	-.036	.072
34	36	.037	.051	.023	.085	.031	.118
60	37	-.156	.093	-.072	.072	-.113	.100
2	38	.161	.088	.174	.118	.105	.144
10	39	.011	.075	-.057	.047	.024	.103
20	40	-.025	.067	-.042	.076	-.096	.106
6	41	-.038	.075	-.033	.096	.056	.134
37	42	-.118	.056	-.085	.103	-.101	.117
13	43	-.099	.074	-.063	.067	-.085	.090
5	44	-.052	.072	.003	.071	.064	.134
32	45	-.102	.115	-.020	.106	-.097	.144
25	46	.005	.108	.013	.086	-.256	.129
46	47	.053	.126	-.071	.143	-.010	.185
49	48	-.087	.059	-.168	.076	-.224	.114
58	49	-.101	.118	-.163	.099	-.234	.133
42	50	-.060	.088	-.151	.116	-.151	.147
1	51	.004	.134	-.285	.151	-.120	.240
54	52	-.247	.088	-.086	.190	-.174	.211
38	53	.003	.094	.087	.307	-.282	.160
30	54	-.310	.136	-.225	.214	-.383	.243
24	55	-.205	.137	-.056	.261	-.245	.276
17	56	-.052	.123	-.117	.192	.045	.300
53	57	.098	.305	.271	.439	-.021	.175
52	58	-.050	.187	-.191	.174	-.026	.261
51	59	.012	.298	-.218	.334	.378	.592
33	60	.748	.556	.668	.674	.822	.777

A P P E N D I X   B

Table 16

Percentages Within 1, 2, 3 and 4 Standard Deviation Units of B Values

Ability Level 1
(n = 600)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	66.7	92.4	100.0	-
55	2	60.0	100.0	-	-
44	3	63.3	96.6	100.0	-
40	4	73.3	96.9	100.0	-
22	5	83.3	96.6	96.6	100.0
35	6	66.7	96.7	100.0	-
12	7	70.0	93.3	100.0	-
29	8	73.3	96.6	100.0	-
7	9	66.7	96.7	100.0	-
8	10	70.0	96.7	100.0	-
56	11	70.0	96.7	100.0	-
9	12	70.0	93.3	100.0	-
48	13	66.7	96.7	100.0	-
16	14	60.0	96.7	100.0	-
47	15	60.0	100.0	-	-
3	16	70.0	93.3	100.0	-
59	17	70.0	93.3	100.0	-
36	18	76.7	96.7	100.0	-
31	19	66.7	93.4	100.0	-
45	20	73.3	96.6	100.0	-
26	21	66.7	96.7	100.0	-
27	22	83.3	96.6	96.6	100.0
14	23	60.0	93.3	100.0	-
21	24	66.7	96.7	100.0	-
41	25	70.0	86.7	100.0	-
15	26	70.0	93.3	100.0	-
50	27	70.0	93.3	100.0	-
28	28	80.0	93.3	100.0	-
39	29	66.7	96.7	96.7	100.0
23	30	56.7	93.4	100.0	-
11	31	76.7	96.7	100.0	-
57	32	73.3	96.6	96.6	100.0

Table 16 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
19	33	73.3	93.3	100.0	-
43	34	73.3	96.6	100.0	-
18	35	76.7	96.7	100.0	-
34	36	63.3	100.0	-	-
60	37	73.3	96.6	100.0	-
2	38	70.0	96.7	100.0	-
10	39	66.7	96.7	100.0	-
20	40	80.0	90.0	100.0	-
6	41	66.7	92.4	100.0	-
37	42	80.0	93.3	100.0	-
13	43	63.3	96.6	100.0	-
5	44	66.7	100.0	-	-
32	45	66.7	96.7	100.0	-
25	46	66.7	96.7	100.0	-
46	47	63.3	100.0	-	-
49	48	76.7	93.4	100.0	-
58	49	76.7	96.7	96.7	100.0
42	50	70.0	96.7	100.0	-
1	51	80.0	96.7	96.7	100.0
54	52	73.3	93.3	100.0	-
38	53	90.0	93.3	96.6	100.0
30	54	80.0	93.3	100.0	-
24	55	86.7	96.7	96.7	100.0
17	56	66.7	93.4	100.0	-
53	57	80.0	96.7	96.7	-
52	58	70.0	93.3	100.0	-
51	59	70.0	93.3	100.0	-
33	60	80.0	93.3	96.6	100.0

Table 17
Percentages Within 1, 2, 3 and 4 Standard Deviation Units of B Values

Ability level 2
(n = 600)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	60.0	96.7	100.0	-
55	2	70.0	96.7	100.0	-
44	3	63.3	96.6	100.0	-
40	4	63.3	100.0	-	-
22	5	60.0	96.7	100.0	-
35	6	63.3	96.6	100.0	-
12	7	70.0	93.3	100.0	-
29	8	70.0	96.7	100.0	-
7	9	63.3	100.0	-	-
8	10	66.7	96.7	100.0	-
56	11	73.3	93.3	100.0	-
9	12	66.7	93.4	100.0	-
48	13	66.7	96.7	100.0	-
16	14	73.3	93.3	100.0	-
47	15	66.7	93.4	100.0	-
3	16	73.3	93.3	100.0	-
59	17	76.7	93.4	100.0	-
36	18	63.3	96.6	100.0	-
31	19	76.7	93.4	96.7	100.0
45	20	63.3	96.6	100.0	-
26	21	56.7	96.7	100.0	-
27	22	70.0	93.3	100.0	-
14	23	70.0	96.7	100.0	-
21	24	63.3	100.0	-	-
41	25	66.7	96.7	100.0	-
15	26	66.7	96.7	100.0	-
50	27	56.7	96.7	100.0	-
28	28	66.7	96.7	100.0	-
39	29	63.3	100.0	-	-
23	30	66.7	96.7	100.0	-
11	31	60.0	96.7	100.0	-
57	32	66.7	100.0	-	-
19	33	66.7	93.4	100.0	-

Table 17 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	70.0	96.7	100.0	-
18	35	70.0	93.3	100.0	-
34	36	66.7	100.0	-	-
60	37	66.7	100.0	-	-
2	38	70.0	96.7	100.0	-
10	39	66.7	93.4	100.0	-
20	40	66.7	96.7	100.0	-
6	41	63.3	96.6	100.0	-
37	42	70.0	93.3	100.0	-
13	43	63.3	100.0	-	-
5	44	56.7	100.0	-	-
32	45	56.7	100.0	-	-
25	46	56.7	96.7	100.0	-
46	47	70.0	93.3	100.0	-
49	48	73.3	96.6	96.7	100.0
58	49	66.7	100.0	-	-
42	50	66.7	96.7	100.0	-
1	51	76.7	96.7	96.7	100.0
54	52	63.3	96.3	100.0	-
38	53	80.0	96.7	100.0	-
30	54	66.7	96.7	100.0	-
24	55	63.3	100.0	-	-
17	56	73.3	93.3	100.0	-
53	57	66.7	96.7	100.0	-
52	58	66.7	96.7	100.0	-
51	59	66.7	96.7	100.0	-
33	60	93.3	93.3	96.6	100.0

Table 18
Percentages Within 1, 2, 3 and 4 Standard Deviation Units of B Values

Skewness Level 3
(n=600)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	60.0	96.7	100.0	-
55	2	70.0	96.7	100.0	-
44	3	63.3	96.6	100.0	-
40	4	63.3	100.0	-	-
22	5	60.0	96.7	100.0	-
35	6	63.3	96.6	100.0	-
12	7	70.0	93.3	100.0	-
29	8	70.0	96.7	100.0	-
7	9	63.3	100.0	-	-
8	10	66.7	96.7	100.0	-
56	11	73.3	93.3	100.0	-
9	12	66.7	93.4	100.0	-
48	13	66.7	96.7	100.0	-
16	14	73.3	93.3	100.0	-
47	15	66.7	93.4	100.0	-
3	16	73.3	93.3	100.0	-
59	17	76.7	93.4	100.0	-
36	18	63.3	96.6	100.0	-
31	19	76.7	93.4	96.7	100.0
45	20	63.3	96.6	100.0	-
26	21	56.7	96.7	100.0	-
27	22	70.0	93.3	100.0	-
14	23	70.0	96.7	100.0	-
21	24	63.3	100.0	-	-
41	25	66.7	96.7	100.0	-
15	26	66.7	96.7	100.0	-
50	27	56.7	96.7	100.0	-
28	28	66.7	96.7	100.0	-
39	29	63.3	100.0	-	-
23	30	66.7	96.7	100.0	-
11	31	60.0	96.7	100.0	-
57	32	66.7	100.0	-	-
19	33	66.7	93.4	100.0	-

Table 18 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	70.0	96.7	100.0	-
18	35	70.0	93.3	100.0	-
34	36	66.7	100.0	-	-
60	37	66.7	100.0	-	-
2	38	70.0	96.7	100.0	-
10	39	66.7	93.4	100.0	-
20	40	66.7	96.7	100.0	-
6	41	63.3	96.6	100.0	-
37	42	70.0	93.3	100.0	-
13	43	63.3	100.0	-	-
5	44	56.7	100.0	-	-
32	45	56.7	100.0	-	-
25	46	56.7	96.7	100.0	-
46	47	70.0	93.3	100.0	-
49	48	73.3	93.3	100.0	-
58	49	66.7	100.0	-	-
42	50	66.7	96.7	100.0	-
1	51	76.7	96.7	96.7	100.0
54	52	63.3	93.3	100.0	-
38	53	80.0	96.7	100.0	-
30	54	66.7	96.7	100.0	-
24	55	63.3	100.0	-	-
17	56	73.3	93.3	100.0	-
53	57	66.7	96.7	100.0	-
52	58	66.7	96.7	100.0	-
51	59	66.7	96.7	100.0	-
33	60	93.3	96.6	100.0	-

Table 19
Percentages Within 1, 2, 3 and 4 Standard Deviation Units of B Values

Ability Level 1
(n = 900)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	66.7	96.7	100.0	-
55	2	76.7	93.4	93.4	100.0
44	3	66.7	93.4	100.0	-
40	4	63.3	96.6	100.0	-
22	5	66.7	96.7	100.0	-
35	6	63.3	100.0	-	-
12	7	56.7	96.7	100.0	-
29	8	56.7	100.0	-	-
7	9	66.7	100.0	-	-
8	10	76.7	96.7	100.0	-
56	11	63.3	96.6	100.0	-
9	12	76.7	93.4	100.0	-
48	13	73.3	96.6	100.0	-
16	14	66.7	96.7	100.0	-
47	15	70.0	96.7	100.0	-
3	16	56.7	96.7	100.0	-
59	17	63.3	100.0	-	-
36	18	73.3	93.3	100.0	-
31	19	70.0	96.7	100.0	-
45	20	50.0	100.0	-	-
26	21	70.0	96.7	100.0	-
27	22	80.0	96.7	100.0	-
14	23	73.3	93.3	100.0	-
21	24	70.0	93.3	100.0	-
41	25	70.0	96.7	100.0	-
15	26	73.3	93.3	100.0	-
50	27	70.0	100.0	-	-
28	28	66.7	96.7	100.0	-
39	29	73.3	96.6	100.0	-
23	30	70.0	96.7	100.0	-
11	31	76.7	86.7	100.0	-
57	32	73.3	93.6	100.0	-
19	33	70.0	96.7	100.0	-

Table 19 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	70.0	96.7	100.0	-
18	35	76.7	96.7	100.0	-
34	36	73.3	96.6	100.0	-
60	37	66.7	96.7	100.0	-
2	38	73.3	93.3	100.0	-
10	39	70.0	100.0	-	-
20	40	63.3	96.6	100.0	-
6	41	56.7	96.7	100.0	-
37	42	60.0	93.3	100.0	-
13	43	66.7	93.4	100.0	-
5	44	73.3	93.3	100.0	-
32	45	56.7	100.0	-	-
25	46	73.3	96.6	100.0	-
46	47	66.7	96.7	100.0	-
49	48	80.0	96.7	100.0	-
58	49	60.0	100.0	-	-
42	50	70.0	96.7	100.0	-
1	51	70.0	96.7	100.0	-
54	52	66.7	96.7	100.0	-
38	53	90.0	93.0	96.0	100.0
30	54	76.7	93.4	100.0	
24	55	70.0	96.7	100.0	-
17	56	83.3	96.6	100.0	-
53	57	76.7	93.4	100.0	-
52	58	80.0	93.3	93.3	100.0
51	59	73.3	93.3	100.0	-
33	60	73.3	96.6	100.0	-

Table 20
Percentages Within 1, 2, 3 and 4 Standard Deviation Units
of B Values

Ability Level 2
(n = 900)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	63.3	96.6	100.0	-
55	2	70.0	96.7	100.0	-
44	3	76.7	93.4	100.0	-
40	4	63.3	100.0	-	-
22	5	56.7	93.4	100.0	-
35	6	80.0	93.3	93.3	100.0
12	7	76.7	93.4	100.0	-
29	8	73.3	93.3	100.0	-
7	9	66.7	96.7	100.0	-
8	10	66.7	96.7	100.0	-
56	11	56.7	96.7	100.0	-
9	12	60.0	96.7	100.0	-
48	13	63.3	96.6	100.0	-
16	14	66.7	93.4	100.0	-
47	15	76.7	96.7	100.0	-
3	16	56.7	100.0	-	-
59	17	73.3	96.6	100.0	-
36	18	76.7	96.7	100.0	-
31	19	63.3	93.3	100.0	-
45	20	66.7	100.0	-	-
26	21	63.3	100.0	-	-
27	22	66.7	96.7	100.0	-
14	23	56.7	96.7	100.0	-
21	24	66.7	93.4	100.0	-
41	25	80.0	93.3	100.0	-
15	26	70.0	96.7	100.0	-
50	27	66.7	96.7	100.0	-
28	28	70.0	96.7	100.0	-
39	29	70.0	93.3	100.0	-
23	30	70.0	96.7	100.0	-
11	31	63.3	100.0	-	-
57	32	63.3	93.3	100.0	-
19	33	63.3	96.69	100.0	-

Table 20 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	70.0	96.7	100.0	-
18	35	66.7	96.7	100.0	-
34	36	70.0	93.3	100.0	-
60	37	63.3	93.3	100.0	-
2	38	63.3	93.3	100.0	-
10	39	70.0	96.7	100.0	-
20	40	63.3	96.6	100.0	-
6	41	60.0	100.0	-	-
37	42	63.3	96.6	100.0	-
13	43	73.3	96.6	100.0	-
5	44	63.3	96.6	100.0	-
32	45	56.7	100.0	-	-
25	46	63.3	93.3	100.0	-
46	47	56.7	100.0	-	-
49	48	63.3	96.6	100.0	-
58	49	56.7	96.7	100.0	-
42	50	73.3	93.3	100.0	-
1	51	66.7	100.0	-	-
54	52	70.0	96.7	100.0	-
38	53	80.0	96.7	96.7	100.0
30	54	80.0	96.7	96.7	100.0
24	55	73.3	93.3	100.0	-
17	56	66.7	96.7	100.0	-
53	57	70.0	93.3	93.3	100.0
52	58	76.7	96.7	100.0	-
51	59	73.3	96.6	100.0	-
33	60	70.0	96.7	100.0	-

Table 21
Percentages Within 1, 2, 3 and 4 Standard Deviation Units of B Values

Ability Level 3
(n = 900)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	73.3	96.6	100.0	-
55	2	80.0	90.0	100.0	-
44	3	63.3	96.6	100.0	-
40	4	70.0	93.3	100.0	-
22	5	66.7	96.7	100.0	-
35	6	66.7	96.7	100.0	-
12	7	70.0	96.7	100.0	-
29	8	70.0	96.7	100.0	-
7	9	66.7	96.7	100.0	-
8	10	73.3	96.6	100.0	-
56	11	76.7	96.7	100.0	-
9	12	76.7	93.4	100.0	-
48	13	70.0	96.7	100.0	-
16	14	66.7	96.7	100.0	-
47	15	73.3	96.6	100.0	-
3	16	66.7	93.4	100.0	-
59	17	63.3	93.3	100.0	-
36	18	70.0	96.7	100.0	-
31	19	70.0	93.3	100.0	-
45	20	73.3	93.3	100.0	-
26	21	76.7	93.4	100.0	-
27	22	63.3	96.6	100.0	-
14	23	70.0	96.7	100.0	-
21	24	63.3	96.6	100.0	-
41	25	76.7	93.4	100.0	-
15	26	73.3	93.3	100.0	-
50	27	70.0	96.7	100.0	-
28	28	73.3	93.3	100.0	-
39	29	76.7	96.7	100.0	-
23	30	66.7	96.7	100.0	-
11	31	63.3	96.6	100.0	-
57	32	63.3	96.6	100.0	-
19	33	66.7	100.0	-	-

Table 21 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	76.7	90.0	100.0	-
18	35	70.0	93.3	100.0	-
34	36	73.3	93.3	100.0	-
60	37	73.3	93.3	100.0	-
2	38	70.0	96.7	100.0	-
10	39	73.3	93.3	100.0	-
20	40	60.0	96.7	100.0	-
6	41	66.7	96.7	100.0	-
37	42	73.3	90.0	100.0	-
13	43	70.0	96.7	100.0	-
5	44	73.3	93.3	100.0	-
32	45	73.3	93.3	100.0	-
25	46	70.0	96.7	100.0	-
46	47	70.0	93.3	100.0	-
49	48	63.3	96.6	100.0	-
58	49	76.7	96.7	96.7	100.0
42	50	60.0	93.3	100.0	-
1	51	66.7	93.4	100.0	-
54	52	66.7	96.7	100.0	-
38	53	83.3	96.6	96.6	100.0
30	54	70.0	93.3	100.0	-
24	55	56.7	100.0	-	-
17	56	46.7	100.0	-	-
53	57	86.7	93.4	96.7	100.0
52	58	73.3	93.3	100.0	-
51	59	70.0	93.3	100.0	-
33	60	66.7	96.7	100.0	-

Table 22
Percentages Within 1, 2, 3 and 4 Standard Deviation Units of B Values

Ability Level 1
(n = 1200)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	80.0	96.7	96.7	100.0
55	2	73.3	93.3	100.0	-
44	3	70.0	93.3	100.0	-
40	4	70.0	96.7	100.0	-
22	5	73.3	93.3	100.0	-
35	6	66.7	93.4	100.0	-
12	7	73.3	93.3	100.0	-
29	8	63.3	96.6	100.0	-
7	9	63.3	100.0	-	-
8	10	73.3	93.3	100.0	-
56	11	63.3	100.0	-	-
9	12	66.7	96.7	100.0	-
48	13	66.7	100.0	-	-
16	14	70.0	96.7	100.0	-
47	15	63.3	100.0	-	-
3	16	73.3	96.6	100.0	-
59	17	66.7	100.0	-	-
36	18	70.0	93.3	100.0	-
31	19	66.7	96.7	100.0	-
45	20	70.0	96.7	100.0	-
26	21	66.7	96.7	100.0	-
27	22	63.3	96.6	100.0	-
14	23	63.3	100.0	-	-
21	24	73.3	96.6	100.0	-
41	25	73.3	96.6	100.0	-
15	26	56.7	96.7	100.0	-
50	27	70.0	93.3	100.0	-
28	28	73.3	96.6	96.6	100.0
39	29	66.7	96.7	100.0	-
23	30	66.7	96.7	100.0	-
11	31	66.7	100.0	-	-
57	32	73.3	93.3	100.0	-
19	33	70.0	90.0	100.0	-

Table 22 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	63.3	100.0	-	-
18	35	70.0	96.7	100.0	-
34	36	66.7	96.7	100.0	-
60	37	80.0	96.7	96.7	100.0
2	38	76.7	96.7	96.7	100.0
10	39	66.7	96.7	100.0	-
20	40	70.0	93.3	100.0	-
6	41	70.0	96.7	100.0	-
37	42	60.0	100.0	-	-
13	43	73.3	100.0	-	-
5	44	70.0	96.7	100.0	-
32	45	80.0	96.7	100.0	-
25	46	73.3	93.3	100.0	-
46	47	70.0	96.7	100.0	-
49	48	56.7	100.0	-	-
58	49	70.0	93.3	100.0	-
42	50	70.0	96.7	100.0	-
1	51	50.0	100.0	-	-
54	52	63.3	96.6	100.0	-
38	53	70.0	93.3	100.0	-
30	54	66.7	96.7	100.0	-
24	55	73.3	96.6	96.6	100.0
17	56	73.3	93.3	100.0	-
53	57	56.7	96.7	100.0	-
52	58	63.3	96.6	100.0	-
51	59	80.0	96.7	96.7	100.0
33	60	73.3	96.6	100.0	-

Table 23
Percentages Within 1, 2, 3 and 4 Standard Deviation Units of B Values

Ability Level 2
(n = 1200)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	66.7	96.7	100.0	-
55	2	66.7	96.7	100.0	-
44	3	73.3	96.6	100.0	-
40	4	66.7	96.7	100.0	-
22	5	70.0	93.3	100.0	-
35	6	63.3	96.6	100.0	-
12	7	66.7	96.7	100.0	-
29	8	66.7	96.7	100.0	-
7	9	70.0	96.7	100.0	-
8	10	63.3	96.6	100.0	-
56	11	73.3	93.3	100.0	-
9	12	70.0	93.3	100.0	-
48	13	76.7	93.4	100.0	-
16	14	73.3	96.6	100.0	-
47	15	76.7	93.4	100.0	-
3	16	83.3	96.6	96.6	100.0
59	17	56.7	96.7	100.0	-
36	18	70.0	96.7	100.0	-
31	19	63.3	100.0	-	-
45	20	63.3	96.6	100.0	-
26	21	66.7	96.7	100.0	-
27	22	60.0	100.0	-	-
14	23	70.0	93.3	100.0	-
21	24	66.7	96.7	100.0	-
41	25	80.0	93.3	96.6	100.0
15	26	73.3	96.6	100.0	-
50	27	53.3	100.0	-	-
28	28	76.7	93.4	96.7	100.0
39	29	73.3	96.6	96.6	100.0
23	30	63.3	96.6	100.0	-
11	31	70.0	96.7	100.0	-
57	32	73.3	96.6	100.0	-
19	33	80.0	93.3	96.3	100.0

Table 23 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	76.7	90.0	100.0	-
18	35	63.3	96.6	100.0	-
34	36	73.3	90.0	100.0	-
60	37	73.3	96.6	100.0	-
2	38	70.0	96.7	100.0	-
10	39	73.3	93.3	100.0	-
20	40	50.0	100.0	-	-
6	41	70.0	96.7	100.0	-
37	42	60.0	96.7	100.0	-
13	43	63.3	100.0	-	-
5	44	56.7	96.7	100.0	-
32	45	63.3	93.3	100.0	-
25	46	66.7	93.4	100.0	-
46	47	80.0	90.0	100.0	-
49	48	66.7	100.0	-	-
58	49	70.0	93.3	100.0	-
42	50	63.3	100.0	-	-
1	51	66.7	96.7	100.0	-
54	52	73.3	93.3	100.0	-
38	53	86.7	93.4	96.7	100.0
30	54	76.7	96.7	96.7	100.0
24	55	70.0	96.7	100.0	-
17	56	83.3	93.3	96.6	100.0
53	57	93.3	96.6	96.6	100.0
52	58	80.0	93.3	96.6	100.0
51	59	73.3	93.3	100.0	-
33	60	70.0	96.7	100.0	-

Table 24
Percentages Within 1, 2, 3 and 4 Standard Deviation Units
of B Values

Ability Level 3
(n = 1200)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
4	1	70.0	93.3	100.0	-
55	2	76.7	90.0	100.0	-
44	3	66.7	93.4	100.0	-
40	4	66.7	96.7	100.0	-
22	5	66.7	96.7	100.0	-
35	6	70.0	96.7	100.0	-
12	7	66.7	96.7	100.0	-
29	8	66.7	93.4	100.0	-
7	9	70.0	96.7	100.0	-
8	10	73.3	93.3	100.0	-
56	11	63.3	100.0	-	-
9	12	70.0	96.7	100.0	-
48	13	60.0	100.0	-	-
16	14	56.7	100.0	-	-
47	15	63.3	96.6	100.0	-
3	16	60.0	100.0	-	-
59	17	63.3	36.7	100.0	-
36	18	73.3	96.6	100.0	-
31	19	70.0	96.7	100.0	-
45	20	70.0	93.3	100.0	-
26	21	63.3	100.0	-	-
27	22	70.0	93.3	100.0	-
14	23	66.7	93.4	100.0	-
21	24	70.0	90.0	100.0	-
41	25	73.3	93.3	100.0	-
15	26	56.7	100.0	-	-
50	27	7.3	96.6	100.0	-
28	28	63.3	100.0	-	-
39	29	70.0	90.0	100.0	-
23	30	70.0	96.7	100.0	-
11	31	66.7	96.7	100.0	-
57	32	70.0	93.3	100.0	-
19	33	70.0	96.7	100.0	-

Table 24 (continued)

item	rank	1 s.d. %	2 s.d. %	3 s.d. %	4 s.d. %
43	34	73.3	93.3	100.0	-
18	35	70.0	96.7	100.0	-
34	36	70.0	96.7	100.0	-
60	37	73.3	96.6	100.0	-
2	38	66.7	100.0	-	-
10	39	66.7	96.7	100.0	-
20	40	66.7	96.7	100.0	-
6	41	63.3	96.6	100.0	-
37	42	73.3	96.6	100.0	-
13	43	70.0	93.3	100.0	-
5	44	63.3	100.0	-	-
32	45	63.3	96.6	100.0	-
25	46	70.0	96.7	100.0	-
46	47	53.3	96.6	100.0	-
49	48	63.3	96.6	100.0	-
58	49	76.7	90.0	100.0	-
42	50	66.7	96.7	100.0	-
1	51	66.7	96.7	100.0	-
54	52	73.3	93.3	100.0	-
38	53	63.3	96.6	100.0	-
30	54	80.0	93.3	100.0	-
24	55	66.7	96.7	100.0	-
17	56	76.7	96.7	96.7	100.0
53	57	70.0	96.7	100.0	-
52	58	63.3	93.3	100.0	-
51	59	70.0	96.7	100.0	-
33	60	70.0	96.7	100.0	-

APPENDIX C

Table 25
Ability Level 1 B Estimates
(n = 600)

item	rank	group	variance	bias	accuracy
4	1	1	10.66171	10.53406	21.19577
55	2	1	.27389	2.40154	2.67542
44	3	1	1.38364	.13561	1.51925
40	4	1	.56867	.00466	.57333
22	5	1	2.50416	.20369	2.70785
35	6	2	.50977	.17480	.68457
12	7	2	1.21606	.26320	1.47926
29	8	2	.79602	.52378	1.31980
7	9	2	.72920	.01192	.74112
8	10	2	.17735	.59136	.76871
56	11	3	.62546	.13926	.76473
9	12	3	1.10518	.76768	1.87286
48	13	3	.42163	.52589	.94752
16	14	3	.39361	.02420	.41781
47	15	3	.27554	.08965	.36519
3	16	4	.26005	.08206	.34211
59	17	4	.77878	.12250	.90128
36	18	4	.22508	.25910	.48417
31	19	4	.29947	.17358	.47306
45	20	4	.14260	.02587	.16847
26	21	5	.16295	.04226	.20521
27	22	5	.11069	.36520	.47589
14	23	5	.20992	.27265	.48258
21	24	5	.18024	.59587	.77611
41	25	5	.14931	.43392	.58323
15	26	6	.29123	2.76762	3.05885
50	27	6	.11282	.00008	.11289
28	28	6	.69532	.02748	.72280
39	29	6	.11126	.00053	.11179
23	30	6	.07863	.00172	.08035
11	31	7	.05989	.08491	.14480
57	32	7	.08712	.01820	.10532
19	33	7	.53706	.00642	.54349

Table 25 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.10128	.11669	.21797
18	35	7	.08532	.05343	.13875
34	36	8	.22785	.00406	.23191
60	37	8	.21061	.16295	.37356
2	38	8	.31057	.06247	.37304
10	39	8	.13611	.25466	.39077
20	40	8	.17046	.03333	.20380
6	41	9	.58947	.01997	.60944
37	42	9	.24473	.17849	.42322
13	43	9	.23219	.99190	1.22409
5	44	9	.20545	.60805	.81349
32	45	9	.88757	.03468	.92225
25	46	10	.20250	.49897	.70147
46	47	10	.31006	.74482	1.05488
49	48	10	.20096	1.02268	1.22364
58	49	10	.84677	.07530	.92207
42	50	10	.46336	.32406	.78742
1	51	11	1.49138	.37408	1.86547
54	52	11	2.59902	.80721	3.40623
38	53	11	3.99248	.01443	4.00691
30	54	11	2.89870	.47226	3.37095
24	55	11	3.93254	1.62355	5.55608
17	56	12	1.62505	.51667	2.14171
53	57	12	2.82058	2.21789	5.03847
52	58	12	10.81724	2.35480	13.17204
51	59	12	13.35797	12.12008	19.47805
33	60	12	1.49914	3.83562	5.33475

Table 26
Ability Level 2 B Estimates
(n = 600)

item	rank	group	variance	bias	accuracy
4	1	1	.77670	.00053	.77723
55	2	1	1.64005	.09219	1.73223
44	3	1	1.27832	.04439	1.32271
40	4	1	.75566	.07086	.82652
22	5	1	1.28582	.16905	1.45487
35	6	2	.59914	.05755	.65670
12	7	2	.66027	.07967	.73994
29	8	2	.21888	2.07665	2.29553
7	9	2	.20945	.72447	.93393
8	10	2	.36198	.00774	.36972
56	11	3	.19283	.74230	.93512
9	12	3	.33004	.28714	.61718
48	13	3	.57891	.36124	.94015
16	14	3	.09778	1.05057	1.14834
47	15	3	.46925	.14173	.61098
3	16	4	.11974	.00335	.12309
59	17	4	.85188	.19732	1.04920
36	18	4	.25887	.00069	.25956
31	19	4	.61895	.29681	.91575
45	20	4	.04473	.04953	.09426
26	21	5	.04738	.03046	.07784
27	22	5	.06133	.01930	.08063
14	23	5	.17557	.08933	.26489
21	24	5	.10600	.02920	.13521
41	25	5	.07749	.44750	.52499
15	26	6	.45532	.00172	.45704
50	27	6	.09205	.01910	.11116
28	28	6	.17378	.04994	.22372
39	29	6	.07160	.00164	.07324
23	30	6	.09244	.14939	.24183
11	31	7	.06745	.09263	.16008
57	32	7	.10337	.12442	.22779
19	33	7	.24366	.00141	.24508

Table 26 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.17540	.00200	.17740
18	35	7	.10561	.02664	.13226
34	36	8	.10728	.11544	.22273
60	37	8	.87763	.00379	.88141
2	38	8	.32247	.15581	.47827
10	39	8	.36913	.05581	.42495
20	40	8	.29531	.05109	.34640
6	41	9	.30443	.73916	1.04359
37	42	9	.37979	.14077	.52055
13	43	9	.45172	.00137	.45309
5	44	9	.36120	.09724	.45844
32	45	9	1.55097	.23568	1.78665
25	46	10	.25204	.02274	.27478
46	47	10	.53854	1.07050	1.60904
49	48	10	.68055	.37565	1.05620
58	49	10	.92144	.06450	.98594
42	50	10	.36720	1.36235	1.72955
1	51	11	2.39531	.84135	3.23667
54	52	11	1.29053	.05250	1.34303
38	53	11	18.38563	4.91508	23.30071
30	54	11	2.45553	5.86357	8.31910
24	55	11	1.20424	5.46731	5.67155
17	56	12	12.72539	.78279	7.50818
53	57	12	2.36159	.00258	2.36417
52	58	12	1.04445	.00273	1.04717
51	59	12	1.26273	8.98502	10.24775
33	60	12	11.44289	.51824	11.96114

Table 27
Ability Level 3 β Estimates
(n = 600)

item	rank	group	variance	bias	accuracy
4	1	1	.40705	.04986	.45690
55	2	1	.31744	.33793	.65537
44	3	1	.44727	1.14583	1.59310
40	4	1	.44213	.02558	.46771
22	5	1	.83779	.13548	.97327
35	6	2	.29864	.04309	.34174
12	7	2	.17143	.64621	.81765
29	8	2	.23467	1.50618	1.74084
7	9	2	.29196	.21017	.50213
8	10	2	.10416	.13507	.23923
56	11	3	.17260	.09965	.27225
9	12	3	.11337	.07183	.18520
48	13	3	.24564	.01196	.25761
16	14	3	.14927	.87313	1.02240
47	15	3	.22573	.08311	.30884
3	16	4	.07675	.25447	.33122
59	17	4	.27441	.27821	.55262
36	18	4	.24482	.18127	.42609
31	19	4	.27143	.00758	.27902
45	20	4	.06914	.04431	.11346
26	21	5	.07318	.35360	.42678
27	22	5	.08609	.00350	.08959
14	23	5	.29457	.01629	.31086
21	24	5	.11351	.56006	.67357
41	25	5	.11585	.28910	.40495
15	26	6	.79616	1.12830	1.92446
50	27	6	.05306	.11945	.17251
28	28	6	.36364	.00569	.36932
39	29	6	.05483	.01228	.06711
23	30	6	.05493	.29800	.35293
11	31	7	.03944	.27399	.31343
57	32	7	.17997	.14756	.32753
19	33	7	.18177	.60549	.78726

Table 27 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.26663	.00151	.26815
18	35	7	.23351	.03509	.26860
34	36	8	.21548	.53841	.75390
60	37	8	.29990	1.54950	1.84941
2	38	8	.48019	1.20681	1.68700
10	39	8	.67970	.16266	.84235
20	40	8	.39078	1.09978	1.49056
6	41	9	.99233	.10396	1.09629
37	42	9	.46274	.06999	.53273
13	43	9	.36825	.28053	.64877
5	44	9	.87246	.29304	1.16550
32	45	9	.52985	3.57213	4.10198
25	46	10	.68563	.01417	.69980
46	47	10	1.36671	.03030	1.39801
49	48	10	.78421	.13858	.92280
58	49	10	1.12440	.28324	1.40764
42	50	10	1.26069	.55135	1.81204
1	51	11	4.14777	.00007	4.14783
54	52	11	.83352	2.18106	3.01458
38	53	11	5.73723	.64094	6.37817
30	54	11	11.39596	1.51875	12.91471
24	55	11	4.37376	5.39158	9.76534
17	56	12	18.28393	9.50569	27.78962
53	57	12	1.52761	7.02187	8.54948
52	58	12	1.37240	.07783	1.45023
51	59	12	12.48871	9.11685	21.60556
33	60	12	8.35050	2.49870	10.84920

Table 28
Ability Level 1 B Estimates
(n = 900)

item	rank	group	variance	bias	accuracy
4	1	1	.55689	.00306	.55995
55	2	1	.83790	.47201	1.30991
44	3	1	1.55105	3.96324	5.51429
40	4	1	.52204	.27763	.79968
22	5	1	.80571	2.18214	2.98785
35	6	2	.78430	.05607	.84037
12	7	2	.55395	.02575	.57970
29	8	2	.64562	.00406	.64968
7	9	2	.35307	.74513	1.09821
8	10	2	.44051	.00456	.44508
56	11	3	.27495	.14394	.41888
9	12	3	.36931	.18127	.55059
48	13	3	.28706	.45362	.74068
16	14	3	.28862	.99627	1.28489
47	15	3	.14434	.07272	.21705
3	16	4	.33140	.53681	.86821
59	17	4	.17351	.78408	.95759
36	18	4	.21839	.31314	.53153
31	19	4	.20857	.28169	.49026
45	20	4	.08267	.01319	.09586
26	21	5	.16097	.84437	1.00534
27	22	5	.08238	.00146	.08384
14	23	5	.20621	.01319	.21940
21	24	5	.16527	.00248	.16775
41	25	5	.07202	.20271	.27473
15	26	6	.39728	.13940	.53668
50	27	6	.05466	.00469	.05935
28	28	6	.22200	.11694	.33894
39	29	6	.04464	.03795	.08258
23	30	6	.08262	.43609	.51871
11	31	7	.08759	.12455	.21214
57	32	7	.10151	.00768	.10919
19	33	7	.09490	.10597	.20087

Table 28 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.08882	.12237	.21119
18	35	7	.06018	.06674	.12692
34	36	8	.14191	.18299	.32489
60	37	8	.38347	.32907	.71254
2	38	8	.24222	.17025	.41247
10	39	8	.27893	.14812	.42705
20	40	8	.40339	.17025	.57365
6	41	9	.17820	.20634	.38454
37	42	9	.12365	.03085	.15449
13	43	9	.21065	.04493	.25558
5	44	9	.29121	.19992	.49113
32	45	9	.55086	.07057	.62142
25	46	10	.30368	.04905	.35273
46	47	10	.44160	.01121	.45281
49	48	10	.19390	.25669	.45059
58	49	10	.38590	.11249	.49838
42	50	10	.34075	.01208	.35283
1	51	11	.88763	.06950	.95714
54	52	11	.73574	1.67749	2.41324
38	53	11	1.78342	.09031	1.87373
30	54	11	.97777	.75525	1.73303
24	55	11	2.90085	.00000	2.90085
17	56	12	2.22356	.00215	2.22571
53	57	12	1.73052	.39354	2.12405
52	58	12	1.47991	.36896	1.84887
51	59	12	.51859	7.15115	7.66974
33	60	12	19.68835	4.78040	24.47875

Table 29
Ability Level 2 B Estimates
(n = 900)

item	rank	group	variance	bias	accuracy
4	1	1	.35404	.00020	.35424
55	2	1	.47452	.00994	.48446
44	3	1	.51258	.00502	.51760
40	4	1	.19467	.08050	.27517
22	5	1	.52652	.07242	.59894
35	6	2	.81238	.64945	1.46183
12	7	2	.36714	.09588	.46302
29	8	2	.29684	.07926	.37610
7	9	2	.17460	.34626	.52086
8	10	2	.11653	.18897	.30551
56	11	3	.10922	.33984	.44906
9	12	3	.25201	.07037	.32238
48	13	3	.21378	.02086	.23464
16	14	3	.10621	.05453	.16074
47	15	3	.12392	.03564	.15955
3	16	4	.10519	.00141	.10661
59	17	4	.25755	.03633	.29388
36	18	4	.15963	.15308	.31271
31	19	4	.10662	.00067	.10729
45	20	4	.04974	.00850	.05824
26	21	5	.07569	.25891	.33460
27	22	5	.02905	.04532	.07437
14	23	5	.11564	.00486	.12051
21	24	5	.07798	.02033	.09831
41	25	5	.18394	.05208	.23603
15	26	6	.09079	.00014	.09092
50	27	6	.05227	.23214	.28441
28	28	6	.40409	.02117	.42526
39	29	6	.05826	.04447	.10273
23	30	6	.12640	.68857	.81497
11	31	7	.08745	.00891	.09636
57	32	7	.04715	.01786	.06501
19	33	7	.21933	.02182	.24115

Table 29 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.08211	.04872	.13083
18	35	7	.09372	.00654	.10026
34	36	8	.13105	.00454	.13559
60	37	8	.29667	.02149	.31816
2	38	8	.47735	1.38976	1.86711
10	39	8	.18307	.01224	.19532
20	40	8	.11025	.64475	.75499
6	41	9	.26022	.13002	.39024
37	42	9	.23729	.11869	.35598
13	43	9	.21034	.15151	.36186
5	44	9	.66762	.15237	.81999
32	45	9	.34618	.87040	1.21659
25	46	10	.73030	.14036	.87065
46	47	10	1.00306	1.61658	2.61963
49	48	10	.14026	.25539	.39565
58	49	10	.41539	1.14700	1.56239
42	50	10	.82192	.03117	.85309
1	51	11	1.66314	.22568	1.88882
54	52	11	.35543	2.00881	2.36423
38	53	11	1.54081	.01408	1.55489
30	54	11	3.35974	1.89556	5.25530
24	55	11	1.99957	.12636	2.12593
17	56	12	.79186	.10538	.89724
53	57	12	3.57837	1.83620	5.41457
52	58	12	1.95044	2.34249	4.29293
51	59	12	.52441	14.24439	14.76879
33	60	12	1.15364	2.09511	3.24875

Table 30
Ability Level 3 B Estimates
(n = 900)

item	rank	group	variance	bias	accuracy
4	1	1	.50155	.01704	.51859
55	2	1	.23525	.34497	.58021
44	3	1	.29652	1.38890	1.68542
40	4	1	.37605	.04602	.42207
22	5	1	.43871	.00840	.44711
35	6	2	.17988	.06941	.24928
12	7	2	.22451	.28832	.51282
29	8	2	.17458	.50363	.67820
7	9	2	.34587	.05677	.40263
8	10	2	.10061	.40531	.50591
56	11	3	.12021	.26489	.38510
9	12	3	.13648	.00131	.13779
48	13	3	.19415	.08835	.28250
16	14	3	.12115	.27037	.39152
47	15	3	.16753	.20501	.37255
3	16	4	.06901	.06120	.13021
59	17	4	.25264	.27995	.53259
36	18	4	.11668	.12949	.24617
31	19	4	.16420	.45781	.62201
45	20	4	.05460	.03130	.08590
26	21	5	.08927	.11532	.20459
27	22	5	.05897	.06855	.12752
14	23	5	.06499	.22309	.28808
21	24	5	.11983	.11507	.23490
41	25	5	.11348	1.62867	1.74215
15	26	6	.68373	3.29876	3.89248
50	27	6	.08766	.00001	.08767
28	28	6	.26508	.08164	.34672
39	29	6	.07706	.00898	.08604
23	30	6	.14472	.26734	.41206
11	31	7	.27155	.14714	.41869
57	32	7	.06404	.00308	.06712
19	33	7	.09161	.54648	.63809

Table 30 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.11818	.86632	.98450
18	35	7	.12638	.40113	.52751
34	36	8	.25112	.02460	.27572
60	37	8	.37007	.21948	.58955
2	38	8	.49365	.86768	1.36133
10	39	8	.23279	.43899	.67178
20	40	8	.16281	1.02712	1.18993
6	41	9	.51664	.68312	1.19976
37	42	9	.57478	.55788	1.13265
13	43	9	.63055	.08619	.71674
5	44	9	.77740	1.55815	2.33555
32	45	9	.98478	.03130	1.01608
25	46	10	.67760	1.47541	2.15301
46	47	10	.67229	.84202	1.50431
49	48	10	.47237	.32282	.79519
58	49	10	1.72168	.30927	2.03095
42	50	10	.69689	.05773	.75462
1	51	11	.79800	1.41093	2.20894
54	52	11	.71882	.84538	1.56419
38	53	11	2.84198	.51562	3.35760
30	54	11	1.18846	4.57549	5.76394
24	55	11	3.08838	.05659	3.14497
17	56	12	1.85486	.75557	2.61043
53	57	12	5.70499	.15380	5.85879
52	58	12	1.30022	.15914	1.45936
51	59	12	2.05853	8.42912	10.48765
33	60	12	23.85095	37.41950	61.27045

Table 31
Ability Level 1 B Estimates
(n = 1200)

item	rank	group	variance	bias	accuracy
4	1	1	.36343	.46725	.83069
55	2	1	.31296	.70564	1.01860
44	3	1	.52436	.12224	.64660
40	4	1	.67871	.97164	1.65035
22	5	1	.69012	.45019	1.14031
35	6	2	.29820	.38194	.68014
12	7	2	1.57859	2.68562	4.26421
29	8	2	.59382	.03982	.63364
7	9	2	.27655	.66961	.94616
8	10	2	.16961	.33349	.50309
56	11	3	.16621	.41772	.58393
9	12	3	.17177	.01030	.18207
48	13	3	.23291	.00867	.24158
16	14	3	.11358	.35469	.46827
47	15	3	.13997	.00043	.14041
3	16	4	.20497	.11187	.31684
59	17	4	.26370	.01129	.27499
36	18	4	.50527	.95016	1.45543
31	19	4	.07297	.78635	.85932
45	20	4	.04698	.00718	.05416
26	21	5	.05861	.12949	.18810
27	22	5	.03361	.04864	.08226
14	23	5	.10873	.02471	.13344
21	24	5	.10751	.15958	.26709
41	25	5	.06668	.04744	.11412
15	26	6	.13903	.00055	.13959
50	27	6	.08619	.44823	.53442
28	28	6	.40553	.07661	.48213
39	29	6	.06084	.02920	.09004
23	30	6	.06969	.01850	.08820
11	31	7	.10029	.04689	.14718
57	32	7	.06800	.03710	.10510
19	33	7	.35581	.02846	.38427

Table 31 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.06879	.00117	.06995
18	35	7	.07486	.24770	.32256
34	36	8	.07527	.04174	.11701
60	37	8	.24946	.72541	.97487
2	38	8	.22343	.78021	1.00364
10	39	8	.16106	.00394	.16501
20	40	8	.13029	.01835	.14864
6	41	9	.16401	.04317	.20718
37	42	9	.09231	.41866	.51097
13	43	9	.15883	.29304	.45187
5	44	9	.15070	.08143	.23214
32	45	9	.38533	.31457	.69990
25	46	10	.33963	.00072	.34036
46	47	10	.46110	.08438	.54548
49	48	10	.09960	.22585	.32545
58	49	10	.40671	.30401	.71072
42	50	10	.22517	.10944	.33462
1	51	11	.52423	.00053	.52476
54	52	11	.22512	1.83620	2.06132
38	53	11	.25833	.00026	.25859
30	54	11	.53671	2.87928	3.41599
24	55	11	.54296	1.26485	1.80782
17	56	12	.43866	.08206	.52072
53	57	12	2.69304	.28989	2.98292
52	58	12	1.00997	.07610	1.08608
51	59	12	2.57186	.00437	2.57623
33	60	12	8.97604	16.76867	25.74471

Table 32
Ability Level 2 B Estimates
(n = 1200)

item	rank	group	variance	bias	accuracy
4	1	1	.62460	.21879	.84339
55	2	1	.40073	.05914	.45987
44	3	1	.33424	.50104	.83528
40	4	1	.43420	1.67986	2.11406
22	5	1	.50213	.01068	.51280
35	6	2	.18413	.22429	.40843
12	7	2	.33998	.48845	.82843
29	8	2	.26200	1.07996	1.34196
7	9	2	.17645	.08933	.26578
8	10	2	.18385	.09285	.27670
56	11	3	.16325	.19797	.36122
9	12	3	.09588	.01353	.10941
48	13	3	.17240	.04736	.21977
16	14	3	.14324	.12571	.26895
47	15	3	.08262	.14756	.23018
3	16	4	.10164	.17818	.27982
59	17	4	.18108	.08258	.26367
36	18	4	.14831	.01461	.16292
31	19	4	.07460	.13521	.20980
45	20	4	.04261	.02133	.06395
26	21	5	.05097	.33054	.38151
27	22	5	.03389	.01212	.04602
14	23	5	.10212	.07834	.18046
21	24	5	.03594	.04401	.07995
41	25	5	.04910	.18221	.23131
15	26	6	.22556	.95873	1.18429
50	27	6	.02829	.04752	.07581
28	28	6	.18226	.01244	.19470
39	29	6	.06758	.02871	.09629
23	30	6	.04041	.12071	.16112
11	31	7	.05229	.00919	.06147
57	32	7	.03927	.21336	.25263
19	33	7	.16324	.00444	.16768

Table 32 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.13758	.07681	.21439
18	35	7	.07681	.03852	.11533
34	36	8	.20996	.01610	.22606
60	37	8	.15063	.15408	.30471
2	38	8	.40135	.91246	1.31381
10	39	8	.06413	.09884	.16297
20	40	8	.16772	.05300	.22072
6	41	9	.26861	.03188	.30049
37	42	9	.30942	.21931	.52873
13	43	9	.12967	.12021	.24988
5	44	9	.14598	.00030	.14629
32	45	9	.32862	.01156	.34019
25	46	10	.21332	.00474	.21806
46	47	10	.59454	.14925	.74379
49	48	10	.16908	.84739	1.01648
58	49	10	.28703	.79446	1.08149
42	50	10	.39312	.68373	1.07685
1	51	11	.66445	2.42991	3.09436
54	52	11	1.04297	.22447	1.26743
38	53	11	2.73072	.22724	2.95796
30	54	11	1.32722	1.51425	2.84147
24	55	11	1.97507	.09296	2.06803
17	56	12	1.06373	.40973	1.47346
53	57	12	5.59203	2.19944	7.79146
52	58	12	.88275	1.10017	1.98292
51	59	12	3.24117	1.42354	4.66471
33	60	12	13.18563	13.38939	26.57202

Table 33
Ability Level 3 B Estimates
(n = 1200)

item	rank	group	variance	bias	accuracy
4	1	1	.28238	.14269	.42507
55	2	1	.30939	.01461	.32400
44	3	1	.23763	.70748	.94511
40	4	1	.37667	.02330	.39997
22	5	1	.28427	.37879	.66306
35	6	2	.30307	.36675	.66982
12	7	2	.21572	.20833	.42406
29	8	2	.12488	.61978	.74466
7	9	2	.15780	.07864	.23644
8	10	2	.07652	.32823	.40475
56	11	3	.05102	.47754	.52856
9	12	3	.14454	.01091	.15545
48	13	3	.14069	.13818	.27887
16	14	3	.06881	.20634	.27515
47	15	3	.05971	.28577	.34549
3	16	4	.05927	.12805	.18732
59	17	4	.11483	.07854	.19337
36	18	4	.13418	.04159	.17576
31	19	4	.10679	.04736	.15415
45	20	4	.04214	.00184	.04399
26	21	5	.03673	.13227	.16899
27	22	5	.05279	.11322	.16601
14	23	5	.07916	.01915	.09831
21	24	5	.07122	.15696	.22818
41	25	5	.08141	.50518	.58659
15	26	6	.33358	.21218	.54576
50	27	6	.06698	.00000	.06698
28	28	6	.16533	1.19520	1.36054
39	29	6	.06375	.00908	.07283
23	30	6	.11941	.09263	.21204
11	31	7	.10228	.00198	.10426
57	32	7	.08183	.04048	.12231
19	33	7	.16404	.06487	.22890

Table 33 (continued)

item	rank	group	variance	bias	accuracy
43	34	7	.12018	.12507	.24524
18	35	7	.14881	.03953	.18834
34	36	8	.40525	.02908	.43433
60	37	8	.28804	.38420	.67224
2	38	8	.60338	.33264	.93602
10	39	8	.30951	.01723	.32675
20	40	8	.32461	.27686	.60148
6	41	9	.51980	.09274	.61254
37	42	9	.39860	.30401	.70261
13	43	9	.23445	.21607	.45052
5	44	9	.52385	.12275	.64660
32	45	9	.60420	.28324	.88744
25	46	10	.48409	1.97018	2.45426
46	47	10	.99556	.00331	.99886
49	48	10	.37611	1.49946	1.87557
58	49	10	.51370	1.63707	2.15077
42	50	10	.62807	.68222	1.31029
1	51	11	1.67280	.43320	2.10600
54	52	11	1.28772	.91072	2.19843
38	53	11	.73909	2.38798	3.12707
30	54	11	1.70937	4.39072	6.10009
24	55	11	2.20704	1.80075	4.00779
17	56	12	2.61727	.05985	2.67712
53	57	12	.88624	.01306	.89931
52	58	12	1.97265	.02107	1.99372
51	59	12	10.17946	4.29560	14.47506
33	60	12	17.50968	20.27874	37.78842

REFERENCES

Angoff, W. H. (1982). Use of difficutly and discrimination indicies for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting item bias. Baltimore, MD: The Johns Hopkins University Press.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, 6, 379-396.

Gifford, J. & Swaminathan, H. Accuracy, priors and bias in Baysian estimation of parameters of item response models. Applied Psychological Measurement. (in press).

Goldstein, H. (1980). Dimensionality, bias, independence, and measurement problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

Hambleton, R. K. & Murray, L. N. (1983). Some goodness of fit investigations for item response models 1,2,3. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K. & Rovinelli, R. (1973). A FORTRAN IV program for generating examinees response data from logistic trait models. Behavioral Science, 18, 74.

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

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Linn, R. L., Levine, M. V., Hastings, C. N. & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

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

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

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379-396.

Mooney, R. F. & Swaminathan, H. (1986). An exploration of b parameter estimate invariance in random and non-equivalent samples. Laboratory of Psychometric and Evaluation Research Report. Amherst, MA: School of Education, University of Massachusetts.

Pine, S. M. (1977). Applications of item response theory to the problem of test bias. In D. J. Weiss (Ed.), Applications of computerized adaptive testing. (Research Report 77-1). Minneapolis: University of Minnesota, Psychometric Methods Program, Department of Psychology.

Wingersky, M. S., Barton, M. A. & Lord, F. M. (1982). LOGIST user's guide. Princeton, N.J.: Educational Testing Service.