VARIABILITY IN THE ESTIMATION OF ITEM OPTION CHARACTERISTIC CURVES FOR THE MULTIPLE-CATEGORY SCORING MODEL

By

DIANNE C. BUHR

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1989

ACKNOWLEDGMENTS

The author wishes to thank all of those who contributed to bringing this study to completion. Her special thanks go to Dr. Linda Crocker for her direction and support as major advisor.

The author is grateful also to Dr. James Algina, Dr. M. David Miller, and Dr. William Hedges for their advice as members of her committee.

Her sincere appreciation is extended to Dr. Jeaninne Webb and Dr. Sue Legg for their understanding and support in this endeavor.

Special thanks go to her two sons, Aaron and Joshua, who helped with computer programming and their support. Last, but certainly not least, the author wishes to express her deep appreciation for the contribution of her husband, Ken, in his encouragement and patience throughout this long process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

VARIABILITY IN THE ESTIMATION OF ITEM OPTION CHARACTERISTIC CURVES FOR THE MULTIPLE-CATEGORY SCORING MODEL

By

Dianne C. Buhr

August 1989

Chairman: Linda Crocker
Major Department: Foundations of Education

The present study was conducted to investigate the variability of the item option characteristic curves (OCCs) produced by the multiple-category scoring model under different conditions of estimation. Factors which were expected to affect estimation were test length, sample size, option, ability distribution of the sample of examinees, and difficulty distribution of the set of items.

Data were drawn from two recent administrations of the mathematics subtest of the College Level Academic Skills Test (CLAST), a minimum competency examination required for college sophomores in Florida. To investigate the effects of test length and sample size, ten random samples of 500 examinees each and ten random samples of 1000 examinees each were selected. The multiple-category scoring model of MULTILOG was used to estimate item parameters with subsets of four, six, eight, and ten items, with four items common to all sets. Option characteristic curves were calculated

for the ten samples under each condition, and the variability of the OCCs across samples was computed by a variation of the Brown-Forsythe statistic. A series of repeated measures analyses of variance were conducted to determine whether the variability of the OCCs was affected by test length, sample size, or option. This analysis was replicated for a separate set of four common items.

To investigate the effects of ability and difficulty distribution, item parameters were estimated with ten negatively skewed and ten normal samples of 1050 examinees each, for five-item subsets of "easy" and "difficult" items, with three common items. Again, analyses of variance were conducted with the Brown-Forsythe measure of the variability among OCCs as the dependent variable.

Significant effects due to sample size were found for seven of the eight items, with more variability among OCCs for samples of 500 examinees than for samples of 1000 examinees. Significant effects due to test length were found for three of the eight items, with four-item subsets showing the greatest variability among OCCs. Ability distribution had a significant effect for one of three items only. There was no significant effect of difficulty distribution. Differences among options were related to the number of examinees selecting the option.

CHAPTER I

INTRODUCTION

Examinations used in large-scale educational testing programs are almost universally comprised of multiple-choice items. Although these items have multiple response options, typically they are scored on a right-wrong basis, assigning a score of 1 to the correct response and a score of 0 to any incorrect response. This is known as binary scoring. When multiple-choice tests are scored in this way, any information about the examinee's ability which may be available in the incorrect responses is ignored. With this method of scoring no consideration is given to the fact that examinees do not respond randomly to incorrect options of an item but make use of what Gullikson (1950) has referred to as partial knowledge in responding.

The development of scoring models for multiple choice tests that take into account partial knowledge reflected by examinees' option choices has a long history in the measurement literature. The goal of this research usually has been to identify scoring methods that yield increased reliability and validity of the test scores. Numerous schemes for weighting the various response options of an

item have been proposed and examined in empirical studies. These include using expert judgment to set weights for the various responses (Hambleton, Roberts, & Traub, 1970; Jacobs & Vandeventer, 1970; Nedelsky, 1954; Patnaik & Traub, 1973), using raw scores to generate weights for combining responses (Davis & Fifer, 1959; Guttman, 1941; Hendrickson, 1971; Reilly & Jackson, 1973; Sabers & White, 1969), requiring the examinee to eliminate responses considered to be correct or responses judged to be incorrect (Coombs, Milholland, & Womer, 1956; Dressel & Schmid, 1953; Willey, 1960), and requiring the examinee to indicate level of confidence in the correctness of the response alternatives, known as confidence weighting (Dressel & Schmid, 1953; Hambleton, Roberts & Traub, 1970; Rippey, 1968; Shuford, Albert & Massingill, 1966).

In general, however, results of studies using the above weighting schemes have not been sufficiently promising in terms of increasing either reliability or validity to encourage further use (Crocker & Algina, 1986; Echternacht, 1972; Hambleton & Cook, 1977). Hambleton and Cook suggested that one problem with such studies has been the use of a group statistic to assess the effectiveness of the scoring system rather than a criterion which reflects the change in precision of ability estimation at different ability levels. Another problem may have been that the ability estimation itself was not based on a strong

mathematical model for relating response choice to examinee ability level.

A method that could provide more precise ability estimation through the estimation of parameters for each of the response options for multiple-choice items is Thissen and Steinberg's (1984) multiple-category scoring model. The multiple-category scoring model is a latent trait or item response theory (IRT) model. A latent trait model specifies the relationship between examinee test performance, which is observable, and the trait underlying the test, which is unobservable. This relationship is represented by an item characteristic curve (sometimes referred to as a trace line or an item characteristic function) which is a mathematical function relating the probability of success on the item to the ability measured by the test. The item characteristic curve (ICC) is represented by a graph (usually in the form of an S-shaped curve), with ability level, denoted by θ, measured on the horizontal axis and the probability of a correct response, $P_g(\theta)$, measured on the vertical axis (see Figure 1). In binary IRT models, because only the probability of making a correct response is considered, only one ICC is needed for one item.

In the multiple-category model, each possible response is represented by an "item option characteristic curve." Figure 2 shows a graph of the option characteristic curves

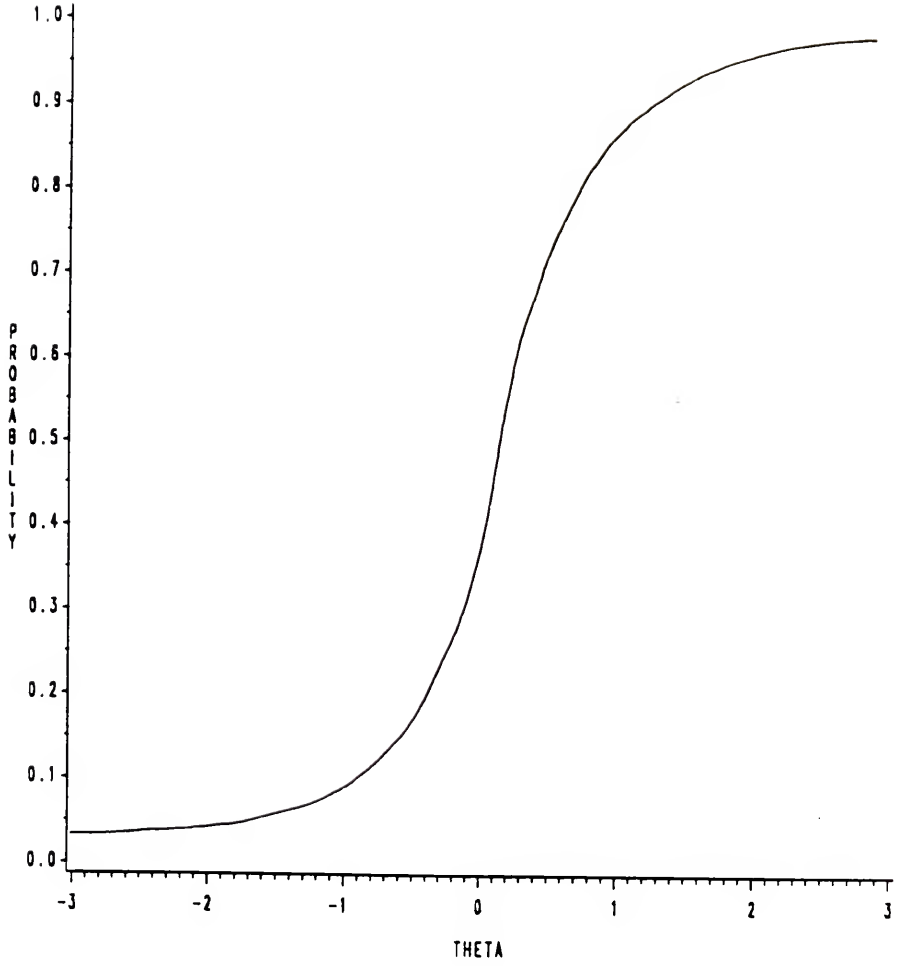

Figure 1. Item Characteristic Curve (ICC) based on the normal ogive model with binary scoring

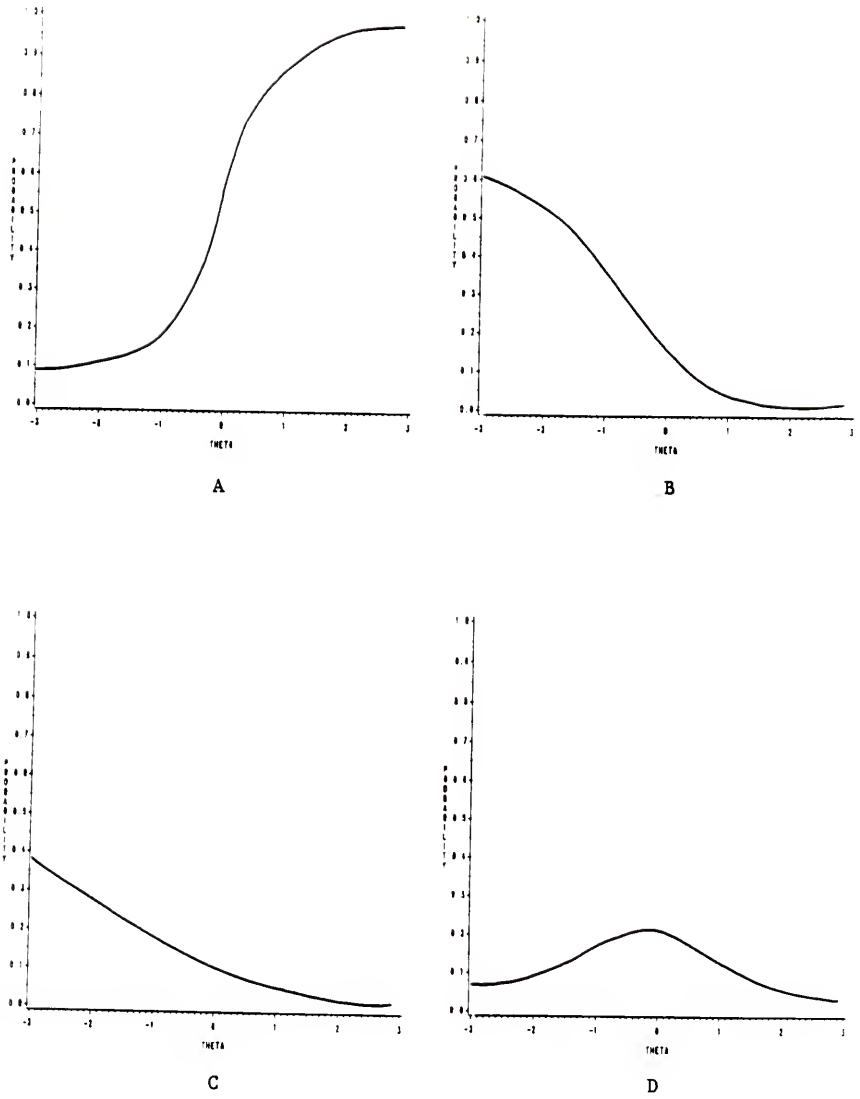

Figure 2. Item option characteristic curves (OCCs) for the multiple-category scoring model

(OCCs) for a hypothetical item with four response options (A, B, C, and D). The curve for the correct response generally is monotonically increasing, as illustrated by response A in Figure 2. The OCCs for the incorrect responses may be monotonically decreasing, as are the curves for responses B and C in Figure 2, but are not restricted to monotonicity by the model. The OCC for option D, for example, indicates that the likelihood of choosing that response is greatest at a theta level of about zero. An additional response category (DK) is included (Samejima, 1979) to represent the latent response category labelled "don't know."

Application of the multiple-category scoring model has been extremely limited. Thissen (1976) applied the model to a subset of 20 items from Raven's Progressive Matrices. He found that the use of multiple-category scoring rather than binary scoring increased information for the lower half of the ability range but made no difference at the upper half. For IRT models information is similar to a measure of reliability for the classical test model. The more precisely the model estimates an examinee's ability, the more information the model provides (Hambleton & Traub, 1971). An important finding in Thissen's study was that the individual "category characteristic curves" produced by the multiple-category model could be meaningfully interpreted.

Using subsets of 4, 12, and 35 items from a military accession test, Thissen and Steinberg (1984) demonstrated that the multiple-category scoring model provided a better fit to the data than a binary IRT scoring model. In this study again, the authors made substantive interpretations of the responses to these items based on visual inspection of the option characteristic curves. The justification for interpreting curves rather than item parameters was that many apparently different sets of parameters result in very similar option characteristic curves.

Results of these studies have indicated the potential of the multiple-category scoring model for providing more information about examinee ability through the analysis of wrong responses. In addition, Thissen and Steinberg have suggested that the multiple-category IRT scoring model permits rigorous item analysis for small item sets.

Considerable research has been devoted to the investigation of the variability of item and ability parameters for the binary IRT models (Bock & Aitken, 1981; Buhr & Algina, 1986; Haberman, 1975; Hulin, Lissak, & Drasgow, 1982; Lord, 1975; Qualls & Ansley, 1985; Ree, 1979; Ree & Jensen, 1980; Swaminathan, 1985; Swaminathan & Gifford, 1979, 1983, 1985, 1986; Wingersky & Lord, 1984; Yen, 1987). Among the factors that have been found to affect parameter estimation are the number of items, the number of

examinees, the method of estimation (Hulin, Drasgow, & Parsons, 1983), the ability distribution of the examinees, and the homogeneity or heterogeneity of the item set within which the items were calibrated (Cook, Eignor, & Petersen, 1982).

For the multiple-category scoring model, however, the only research on the variability of the option characteristic curves is a recent study by Crocker (1987) which explored options of the MULTILOG procedure (Thissen, 1984) for multiple-category scoring. Crocker examined the variability of item option trace lines over two random samples of 1000 each for a professional education test and two samples of 2000 each for a mathematics test. Visual comparisons of OCCs for items from the same item sets showed relatively minor differences. Only one subset of four items was considered in each comparison, and other test properties that might have affected the variability of the item parameter estimates were not systematically varied or controlled.

Purpose

The purpose of the present study was to investigate the variability of the option characteristic curves produced by the multiple-category scoring procedure of MULTILOG under a variety of controlled conditions. The specific questions to be addressed were

1. Does the number of examinees in the sample affect the variability of the option characteristic curves?
2. Does the number of items in the item set affect the variability of the option characteristic curves?
3. Does the difficulty distribution of the item set affect the variability of the option characteristic curves?
4. Does the ability distribution of the sample of examinees affect the variability of the option characteristic curves?
5. Is the variability of the option characteristic curves affected by the option?

Assumptions and Limitations

With a latent trait model, such as Thissen and Steinberg's multiple-category scoring model, certain assumptions must be made. One of these assumptions is that of unidimensionality, which means that a single latent trait underlies performance on the test. Lord and Novick (1968) suggested that a unidimensional model may be appropriate for "tests that appear as though they ought to be homogeneous" (p. 381). However, a mathematics test made up half of arithmetic reasoning items and half of plane geometry items would be expected to show at least two

latent dimensions. Therefore, for this study, subsets of items which tested one content area of mathematics only, algebra for example, were used.

The particular item sets available constituted a limitation of the study, because the ranges of item difficulty and discrimination were limited for these small item subsets. The use of simulated data would have allowed for more control over these variables. However, an empirical study such as this provides information about the performance of the multiple-category scoring model with actual test data, which is important for application of the model.

Significance of the Study

The multiple-category scoring model seems promising for application in several areas of measurement research, including the examination of differential item functioning (DIF) for different ethnic and/or sex groups through the analysis of differences in option characteristic curves. Currently, one of the most prominent approaches to the detection of items that function differently for different groups is through comparisons of their ICCs for correct options only. This approach, however, does little to increase our understanding of the underlying cause in the item which may result in differential performance.

Several researchers (Donlon, 1984; Frary & Giles, 1980; Jensen, 1976; Scheuneman, 1982; Veale & Foreman, 1983; Wendler & Carleton, 1987) have suggested that the study of bias in multiple-choice items must include an examination of examinee responses to incorrect options, a process which is termed distractor analysis. Previous research on distractor analysis has been limited by the lack of a procedure which takes into account underlying differential abilities of examinees. The multiple-category scoring model overcomes this limitation.

The potential value in this method of analysis is that if differential performance can be traced to a particular foil (or set of foils written to the same item specifications), the bias can be eliminated by rewriting test foils and/or avoided by changing item specifications. For this potential to be realized, however, it is important to know that the option characteristic curves produced by the multiple-category scoring model are stable under different conditions of estimation. Otherwise, apparent differences in OCCs estimated for different ethnic or gender groups could simply be due to sampling variability in the estimation of the curves.

The specific questions posed for this study were chosen as those that would have the most direct implications for future research and application of multiple-category IRT for the detection of item bias.

CHAPTER II
REVIEW OF LITERATURE

Traditional Test Scoring Models

Differential Option Weighting Methods

The multiple-category scoring model is based on the idea of partial knowledge (Gullikson, 1950), which refers to knowledge that enables examinees to discriminate "better" from "worse" wrong responses when they cannot identify the correct response. Knowing an examinee's choice of incorrect options increases our ability to estimate his level of knowledge.

Various methods have been used in attempts to incorporate information available in wrong responses to multiple-choice items. Nedelsky (1954) made the first step in the direction of differential weighting of incorrect responses. He instructed experts to identify distractors which were so grossly incorrect as to be attractive only to an "F" student. Students received three scores: (a) R, the number of items answered correctly; (b) F, the number of F options chosen; and (c) a composite score computed from the formula R-F/f, where f was the average number of F options per item. Nedelsky found that for students scoring

in the D or F range with conventional scoring, the F score was the most reliable score, and that for all students, the composite score was considerably more reliable than the conventional score.

Others who have used judgment to provide weights for concatenating all responses into the test score include Jacobs and Vandeventer (1970), Hambleton, Roberts, and Traub (1970), and Patnaik and Traub (1973). Jacobs and Vandeventer (1970) used facet analysis as the basis of a procedure for systematically ordering the distractors on Raven's Coloured Progressive Matrices Test (Raven, 1956) as to degree of correctness. This a priori method of keying responses was shown to have a moderate degree of test-retest reliability as well as concurrent and predictive validity against the criterion of the number of correct answers on the same and a different form of the test.

Hambleton et al. (1970) compared a procedure based on the average rank assigned to options by judges who ranked all options for correctness with a confidence testing procedure that required an examinee to indicate the probability of correctness of each option by marking on a percentage-scale graph. Predictive validity was increased with both procedures, and reliability was increased with the first procedure, although none of the increases was statistically significant.

Patnaik and Traub (1973) used degree-of-correctness judgments for the set of response alternatives to develop weights for a group test of general intelligence. Weighted scores resulted in a significant increase in reliability over conventional and formula scoring. However, validity decreased.

Others have suggested empirical weighting schemes that use the raw scores to generate weights for combining responses (Davis & Fifer, 1959; Hendrickson, 1971; Reilly & Jackson, 1973; Sabers & White, 1969), employing variations of Guttman's (1941) empirical weighting technique in which a weight is assigned to each option based on the mean total score of the examinees marking the option in question. Davis and Fifer (1959) constructed two parallel tests of 45 arithmetic items each. From the responses of a sample of examinees, response-option weights were derived empirically via the correlation between marking the option and the total score on all 90 items. A new sample of examinees took both forms of the test, and both weighted and unweighted scores were obtained. Parallel-forms reliability for the weighted scores was .76, while that for the unweighted scores was just .68, a difference equivalent to lengthening the test from 45 to 67 items. However, no increase in validity, as measured by teacher's rating and the score on a free-response form of the same test, resulted.

Sabers and White (1969) stated that unless increases in reliability due to weighted scoring are accompanied by increases in criterion-related validity, it may be assumed that what is being added by using weights is irrelevant variance. In their study, where the percentages of the upper and lower 27% of a sample marking an item option were used to set Guttman weights for options, there was virtually no increase in reliability or validity with the weighted scores.

Hendrickson (1971) asserted that the lack of evidence of increased validity in the previous two studies was a result of the techniques utilized. She also applied Guttman's technique for maximizing internal consistency to obtain option weights for the Scholastic Aptitude Test (SAT). Weights were derived separately for males and females for four subtests of the SAT via an iterative procedure which began by assigning to a response category a weight equal to the mean total score on the remaining items of the subtest obtained by the examinees who marked that response category. Hendrickson found an average effective increase in test length, estimated from the Spearman-Brown prophecy formula, of 49% for the weighted tests. An interesting result of Hendrickson's study was that the factor structure of the test changed with weighting, indicating that the test was measuring different aspects of the underlying ability with weighted scores.

Reilly and Jackson (1973) used two empirical weighting techniques, a parallel forms keying based on Davis and Fifer's (1959) procedure, and an internal consistency keying similar to that used by Hendrickson (1971). Reliability increments ranging from .03 to .06 occurred for both methods over the conventional scoring method for the four forms of the test. However, validity, as measured by the correlation with undergraduate grade point average, decreased. Again, the factor structure of the test changed with the weighting procedures.

A different approach to item scoring requires additional input from the examinee. Dressel and Schmid (1953), for example, instructed examinees to cross out options of five-option multiple-choice items until they were certain that the correct response had been crossed out. Each incorrect mark was scored as -1/4 point. Reliability for this method was .67 as compared with .70 for the conventional scoring method.

Coombs, Milholland, and Womer (1956) required examinees to eliminate the incorrect alternatives, rather than the correct alternative. When r out of k alternatives were eliminated, the score was +r when the correct response was not eliminated and r-k when it was. The average gain in reliability with this method was equivalent to that achieved by lengthening the test by 20%. Greater gains in reliability occurred for difficult tests than for easy

tests. Validity, in terms of the correlation with aptitude test score, did not differ for this method from the conventional method of scoring.

Willey (1960) proposed that information about the examinee's ability to eliminate distractors could be incorporated into the test score by requiring the examinee not only to choose the correct alternative but also to eliminate two definitely incorrect alternatives from five options. However, this procedure was shown to raise the expected chance score from 20% for five-alternative items to 46.7% of the maximum score without increasing predictive validity, as measured by academic achievement (Bernhardson, 1966, 1967).

Still others have proposed confidence testing, which requires examinees to indicate their degree of confidence in the correctness of the response alternatives. Dressel and Schmid (1953) investigated confidence weighting with multiple-choice items, having subjects assign a confidence weight from one to four to the option they thought was correct. The weight was scored positive when the choice was correct and negative when incorrect. Reliability was .73 for this method versus .70 for conventional scoring, while validity was the same for the two methods. Also, degree of certainty was shown to be related to the difficulty of the item. Increases in reliability with confidence weighting were also found by Michael (1968) and

Hopkins, Hakstian, and Hopkins (1970). A decrease in validity, however, was found in the latter study.

Another type of confidence weighting is known as "admissible probability measurement." This type of weighting requires the examinee to assign subjective probabilities to each option of a multiple-choice item (Shuford, Albert, & Massingill, 1966). The scoring system is devised so that the examinee can maximize his expected item score only if he reports as accurately as possible the distribution of his subjective probabilities over the options. Shuford, Albert, and Massingill's (1966) admissible probability scoring systems have been investigated in two studies. Rippey (1968) reported erratic changes in the reliability of the test with these systems. Hambleton, Roberts, and Traub (1970) found no statistically significant difference in reliability or validity with Shuford's logarithmic scoring system, which depends only on the probability assigned to the correct option. These and other differential weighting methods were reviewed by Wang and Stanley (1968).

Alternatives To Differential Option Weighting

In discussing the reasons for the general lack of a significant increase in reliability and/or validity found in the above studies of differential weighting of item options, Hambleton and Cook (1977) noted that a major

problem with these studies was the use of a group statistic to assess the effectiveness of the scoring system rather than a criterion which reflects the change in precision of ability estimation at different ability levels. Namely, by producing a group statistic to reflect the improvements through the new scoring system "any gains at the low end of the ability continuum will be washed out when combined with the lack of gain in information at other places on the ability continuum" (Hambleton & Cook, 1977, p. 92).

In an attempt to take examinee ability level into account, Levine and Drasgow (1983) divided examinees into ability strata on the basis of LOGIST three-parameter maximum likelihood estimates of ability. They then plotted the proportion of people choosing each option at each ability level for selected items on the Graduate Record Examination (GRE) Verbal and SAT Verbal tests. The resulting curve was called the empirical option characteristic curve. Chi square statistics were used to examine the independence of ability and option choice. Levine and Drasgow concluded that the pattern of incorrect option choice was related to estimated ability.

Recently, a method of item parameter estimation has become available which takes into account examinees' wrong responses while providing a measure of precision of ability estimation at each ability level. This is Thissen and Steinberg's (1984) computer program MULTILOG (Thissen,

1984). The advantage which the multiple-category scoring procedure has over the three-parameter estimation used by Levine and Drasgow is its inclusion of the wrong response data in the actual estimation of the item and ability parameters.

Item Response Theory (IRT) Models

Binary IRT Models

To understand the multiple-category model requires an understanding of latent trait or item response theory (IRT). A latent trait model specifies the relationship between examinee test performance, which is observable, and the trait underlying the test, which is unobservable. This relationship is represented by an item characteristic curve (ICC), which is a mathematical function that relates the probability of success on the item to the ability measured by the test.

The ICC is a graph that depicts the relationship between the underlying latent trait or ability measured by a collection of test items and the probability that an individual at a given ability level will make a particular response to an item. The graph is usually in the form of an S-shaped curve, with ability (denoted by θ) measured on the horizontal axis and the probability of a correct response measured on the vertical axis (See Figure 1).

In binary IRT models each item is represented by one item characteristic curve only, since only the probability of making a correct response is considered. Early uses of ICCs for item analysis in the 1940s and 1950s were based primarily on the two-parameter normal ogive model (Baker, 1977).

With this binary-scoring model, the probability of a correct response to item g from a person with ability level θ is given by

$$P_g(\theta) = \int_{-L(\theta)}^{\infty} f(t)dt$$

where f(t) is the standard normal frequency function and

$$L(\theta) = a_g(\theta - b_g).$$

In the preceding equation b_g represents the difficulty of item g and a_g the discrimination of item g, as described by Birnbaum (1968). The value of b_g is the point on the θ scale at which the examinee has .50 probability of answering the item correctly. In the normal ogive model, this coincides with the point of inflection of the curve. The value of parameter a_g is proportional to the slope of the tangent to the curve at that point.

The logistic model, a function which coincides closely with the normal ogive (Birnbaum, 1968) is more commonly used in computer programs for convenience in computation.

According to this model

$$P_g(\theta) = \frac{e^z}{1 + e^z}$$

where, in the two-parameter model, $z = a_g(\theta - b_g)$.

In the equation above z may also be expressed as

$$z = a_g\theta + c_g$$

where a_g is the discrimination and $c_g = -a_g b_g$.

The Multiple-Category Scoring Model

Until recently, a method for estimating parameters and creating characteristic curves for each item response through a single calibration was not readily available. The multiple-category scoring model (Thissen & Steinberg, 1984) implemented in the computer program MULTILOG provides this option. Specifically, it provides a latent trait method which allows information in wrong responses to be used to improve the accuracy of ability estimation with procedures that yield efficient estimates of ability and large-sample standard errors of those estimates.

Thissen's model is based on Bock's (1972) nominal response model, which specifies the probability that an individual of a particular latent ability will respond in each of several categories for each item. It is a generalization of the Birnbaum three-parameter model that utilizes information in the pattern of wrong responses as well as the correct responses to estimate ability. The two

models yield identical estimates when there are only two categories of response, right and wrong.

According to Bock's model, for m categorical responses, m response functions of the form $z_k = a_k\theta + c_k$ are specified. For an item response, $x_g = h$, in which $h = 1,2,\ldots,m_g$ for a multiple choice item g with m_g response alternatives

$$P(x_g = h|\theta;\underline{a},\underline{c}) = \frac{\exp(z_h)}{\sum_{k=1}^{m} \exp(z_k)}$$

where $\underline{a}$ and $\underline{c}$ are vectors of the a_k and c_k parameters for the m response categories.

Bock's model restricts the parameters so that the $\Sigma c_k = 0$ and $\Sigma a_k = 0$. This model results in a characteristic curve for the correct response which is monotonically increasing and curves for the incorrect responses, one of which must be monotonically decreasing. To correct this condition, which is rarely found in actual test data, Samejima (1979) added an additional completely latent response category labelled "zero," which Thissen and Steinberg refer to as the "don't know" (DK) category. This adds parameters a_0 and c_0, so that

$$P(x_g = h|\theta;\underline{a},\underline{c},\underline{d}) = \frac{\exp(z_h) + d_h\exp(z_0)}{\sum_{k=0}^{m} \exp(z_k)}$$

in which h takes the values $1,2,\ldots,m_g$.

In Samejima's model, the d_h were fixed and equalled $1/m_g$, which represents the hypothesis that those of sufficiently low ability assign their responses randomly with equal probability to each of the response alternatives. Thissen and Steinberg (1984) found this unlikely and extended the model to allow d_h to vary over the m options. In their model the set of free parameters for an item consists of m_g α_k's (a-contrasts), m_g τ_k's (c-contrasts), and (m_g-1) δ_k's (d-contrasts), for a total of $3m_g-1$ parameters (11 for a four-option multiple-choice item).

Thissen and Steinberg's multiple-category scoring model is implemented through the MULTILOG computer program. Options within MULTILOG also include estimation for the one-, two- and three-parameter models.

Various estimation procedures are available for estimating item and ability parameters for IRT models (Swaminathan, 1985). The procedure used by MULTILOG is marginal maximum likelihood estimation. Estimates of a_g and c_g are obtained from repeated applications of a two-step solution, known as the E-M algorithm. Basically, in the E-step the likelihood of the occurrence of a particular response pattern is estimated at different ability levels, using provisional estimates of item parameters a_g and c_g. In the M-step, refined estimates of a_g and c_g are computed using expected frequencies of correct response to each item

at different ability levels that were calculated during the E-step. The sequence of E-step and M-step is repeated until either the parameters stabilize or a fixed number of cycles, usually 15 or 20, is reached.

Research utilizing the multiple-category model has been extremely limited. Thissen (1976) used responses of junior high students to 20 items of Raven's Progressive Matrices to compare information for the three-parameter and multiple-category models. He found that multiple-category scoring yielded from one-third more to nearly twice the information of the binary model for the lower half of the ability range. However, there was no substantial difference between test information curves for the upper half of the ability range. This increase in accuracy of ability estimation comes from the information available in different choices of incorrect responses made by subjects of differing ability. Thissen concluded that the multiple-category model is primarily useful for low ability subjects or difficult tests. Thissen also showed that the "category characteristic curves" produced by the multiple category model were meaningfully interpretable for Raven's Progressive Matrices.

Thissen and Steinberg (1984) illustrated the use of the multiple-category scoring model to provide a better fit to data from a military accession test. The authors based their interpretation of responses to these items on the

option characteristic curves, rather than item parameter estimates, since many apparently different sets of parameters can result in very similar option characteristic curves.

Crocker (1987) explored options of the MULTILOG program for multiple-category scoring on a state teacher certification examination. She examined the variability of item foil trace-lines over two samples by plotting option characteristic curves and found them to be "relatively stable" for two sets of four items each with samples of 1000 and 2000 examinees. Unfortunately, Crocker did not suggest any objective way of quantifying the differences observed in the OCCs. Nor did she consider how other factors in the item calibration situation might have affected the variability of the OCCs. Research with binary-scored IRT models, however, has indicated that a variety of factors may influence item parameter estimation of the ICC for the correct response. Thus, it is reasonable to suggest that these same factors may affect item parameter estimates when curves for incorrect options are also estimated. To date no study has been conducted to examine the variability of option characteristic curves when factors differ in the estimation procedure.

Factors Affecting IRT Parameter Estimation

Factors that may affect item parameter estimation in binary-scored latent trait models include the following:

1. sample size;
2. test length;
3. choice of IRT model (Hulin, Drasgow & Parsons, 1983);
4. ability distribution;
5. homogeneity or heterogeneity of the item set within which the items were calibrated (Cook, Eignor, & Petersen, 1982).

A number of researchers have explored the effect of the number of items and number of examinees on item parameter estimation for the binary scoring IRT models. The number of items and examinees needed for accurate estimation varies with the estimation procedure. Methods for item and ability parameter estimation include a) joint maximum likelihood estimation (JML), b) conditional maximum likelihood estimation (CML), c) marginal maximum likelihood estimation (MML), and d) Bayesian estimation (BE). Explanations of these procedures can be found in Hambleton and Swaminathan (1983) and Swaminathan (1985).

Most research on parameter estimation has dealt with the joint maximum likelihood estimation procedure, particularly as implemented by the LOGIST (Wood, Wingersky,

& Lord, 1976) computer program. The JML estimators are inconsistent when the number of items is finite; however, the estimators have been shown to be consistent as the number of items and examinees approach infinity for the one-parameter logistic model (Haberman, 1975). Empirical results (Lord, 1975; Swaminathan & Gifford, 1979, 1983) suggest that this result holds for the two- and three-parameter models.

Lord (1968) concluded that with the joint maximum likelihood estimation procedure at least 1000 examinees and 50 items are needed for adequate estimation of the discrimination parameter of the three-parameter model. Swaminathan and Gifford (1979) found that discrimination parameters were poorly estimated for item sets of 10, 15, and 20 items and samples of 50 or 200 examinees. Correlations between the estimated and true discrimination parameters ranged from -.02 for a 15-item test with 50 examinees to .88 for an 80-item test with 1000 examinees. The largest correlations for each test length occurred with the greatest number of examinees. Ree and Jensen (1980) found large estimation errors for the discrimination parameter even with 80 items and 1000 examinees.

For estimating difficulty, Swaminathan and Gifford found that smaller numbers of items and examinees were needed than for estimating discrimination. A test with only 10 items and 50 examinees produced a correlation

between the estimated and true difficulty of .95, with larger correlations for the longer tests and larger sample sizes.

Swaminathan and Gifford (1979) also investigated the effect of ability distribution on item parameter estimation. Correlations for the estimated and true discrimination parameters were near zero for a 10-item subset for all sample sizes with a skewed distribution of ability. With 1000 examinees and 80 items, correlations of true and estimated discrimination parameters were similar for the skewed ability distribution (.82) and the normal distribution (.88). For a uniform distribution, the correlation was even lower, .73. For this distribution, correlations of true and estimated discrimination parameters were generally poor for all sample sizes with the 10-, 15-, and 20-item tests.

Ability distribution had less effect on the estimation of the difficulty parameter. Results for the skewed distribution were similar to those for the normal distribution, except that extreme values of b_g sometimes occurred. There was more effect with the uniform distribution, with a correlation of .80 between estimated and true difficulty parameters for the smallest number of items and sample size, as compared to a correlation of .95 for the normal ability distribution.

Swaminathan and Gifford (1983) examined the relationship between number of items, number of examinees, and ability distribution on the accuracy of estimation of item and ability parameters for the LOGIST maximum likelihood estimation procedure and Urry's (1978) ANCILLES procedure, which is based on classical theory approximations. Again, discrimination parameters were poorly estimated for both procedures, with estimates for the 10-item test worst. Difficulty estimates correlated highly with the true values except for the test with 10 items and 50 examinees. Number of items seemed to have more effect on the estimation of discrimination parameters than did number of examinees.

Estimates for the uniform distribution of ability were similar to those for the normal distribution. However, the negatively skewed distribution affected estimates of a_g and c_g for the ANCILLES procedure and estimates of a_g for the LOGIST procedure. Swaminathan and Gifford concluded that it was not departure from normality but departures from symmetry and the unavailability of examinees in the lower tail of the θ distribution that affected the estimation procedure.

Wingersky and Lord (1984) showed that the standard errors of the item parameter estimates varied inversely as the square root of the number of examinees but were only slightly affected by the number of items. They also found that using a rectangular distribution of ability rather

than a normal distribution gave smaller standard errors for the item parameters than doubling the number of items. In fact, for low a_g's and for c_g's for items with b_g-2/a_g less than -1, the standard errors computed with a rectangular distribution of ability were nearly as low as the standard errors computed with a normal distribution and quadruple the number of examinees.

Ree (1979) compared three procedures, ANCILLES, LOGIST and OGIVIA (Urry, 1977) for the three-parameter model with different ability distributions, using 80 items and 2000 examinees. Correlations of estimated difficulty with true difficulty were high for all procedures, but correlations of discrimination were low for all. Correlations of estimated and true discrimination were worse for the negatively skewed distribution of ability than for the rectangular or normal distributions. LOGIST estimation was very poor for the negatively skewed distribution, with correlations of .45 for difficulty, .57 for discrimination and .23 for the lower asymptote. In addition, all of the ICCs resembled the true ICCs very poorly for this distribution of ability.

Cook, Eignor and Petersen (1982) investigated the temporal stability of item parameter estimates, using LOGIST and the three-parameter model with the SAT Verbal (85 items), SAT Quantitative (60 items), and three 100-item Achievement tests. They found a strong relationship

between variability of parameter estimates and discrepancies between ability levels of samples of examinees from earlier and more recent administrations. The researchers concluded that the stability of item parameter estimates was influenced more by differences in group ability than by the length of time between administrations.

Swaminathan and Gifford (1985) investigated Bayesian versus LOGIST maximum likelihood estimation procedures with tests of 15, 25, and 35 items and 50, 100, 200, and 500 examinees, with normally distributed ability and difficulty. The Bayesian procedure yielded larger correlations of estimated with true item parameters, particularly for the discrimination parameter. However, as the number of items and examinees increased, the procedures yielded similar results.

A few researchers have investigated effects of varying factors on the estimation of parameters by marginal maximum likelihood estimation. Bock and Aitken (1981) showed that parameter estimates obtained with rectangular and empirical priors are all but indistinguishable from those obtained with normal priors, which suggests that MML procedures can be freed from assumptions about the population of subjects by roughly estimating the distribution in the sample for some provisional estimate of item parameters. Although they maintained that MML estimation is insensitive to the form of the prior distribution, they stated that it is

preferable to employ empirical priors when the calibration sample is selected arbitrarily.

Swaminathan and Gifford (1986) examined the effect of test length (20, 40, and 60 items) and number of examinees (250, 500, and 1000) for the JML procedure of LOGIST versus the MML procedure of BILOG (Mislevy & Bock, 1984) on the accuracy of item and ability parameter estimation, for the one-, two- and three-parameter models. With the three-parameter model the MML procedure resulted in higher correlations between estimated and true item parameters for the combination of 20 items and 250 examinees for both the a_g and b_g parameters. Both procedures estimated c_g poorly (correlations less than .56 for true and estimated values). For a given number of items, estimation of a_g improved as the number of examinees increased for both estimation procedures.

With the one- and two-parameter models, the MML procedure of BILOG produced better estimates of the item parameters than the JML procedure of LOGIST, particularly with small item sets. For a given test length, increasing the number of examinees greatly increased the accuracy of estimation for both procedures, but for a fixed sample size, increasing the test length had no noticeable effect on BILOG but a substantial effect on LOGIST. Although item difficulty was not a factor varied in the study, Swaminathan and Gifford found that the MML procedure of BILOG

showed poorer estimation of the difficulty parameter at the lower end of the difficulty scale.

Qualls and Ansley (1985) compared item and ability parameter estimates derived from the LOGIST JML procedure and BILOG's MML estimation procedure, for nine combinations of test length (10, 20, and 30 items) and sample size (200, 500, and 1000 examinees). Ability was normally distributed. For the 20-item and 30-item subsets, item difficulties were uniformly distributed from -2 to +2; however, for the 10-item subsets, easy tests, with difficulty centered at -.90, were simulated. The authors indicated that this was done because both BILOG and LOGIST had great difficulty in deriving reasonable estimates for the 10-item subsets with difficulties centered at zero. Thus, although item difficulty was not specifically investigated, it appears as if there were an effect of item difficulty distribution on the estimation procedure.

Both estimation procedures produced higher correlations of true and estimated item parameters as sample size increased; however, BILOG estimates were, in general, more similar to the true parameters than were the LOGIST estimates. The superiority of the MML estimation procedure of BILOG was most apparent for the small sample size (200 examinees) and number of items (10).

Yen (1987) compared LOGIST and BILOG for item sets of 10, 20 and 40 items and a sample size of 1000, with

different ability distributions (normal, negatively skewed, positively skewed, and symmetric/platykurtic). For the 20- and 40-item subsets, the author also varied item difficulty, simulating easy, moderate and difficult tests, with respective mean proportion-correct scores of .75, .65, and .55. In almost every case, BILOG estimates of item parameters were more accurate than LOGIST estimates. The superiority of the MML procedure of BILOG over LOGIST's JML estimation in estimating the item characteristic curves was particularly apparent for the 10-item subset. However, BILOG estimates for a_g and c_g with the negatively skewed distribution were not as accurate as the LOGIST estimates, with the 20-item subset. For LOGIST, correlations of estimated discrimination parameters with the true values were somewhat higher for the negatively skewed ability distribution than with the symmetric or positively skewed distribution, for the 20-item subset only. Ability distribution had no noticeable effect upon the estimation of the ICCS. Item difficulty affected the JML estimation of the c_g parameter, resulting in a positive correlation between true difficulty and estimated c_g.

Buhr and Algina (1986) investigated the effect of sample size and prior ability distribution on item and ability parameter estimation for four methods of estimation within BILOG--MML, MML with updating of the ability distribution each estimation cycle, which is similar to the

JML procedure, and Bayesian estimation both with and without updating of the prior distributions. Samples of 250, 500, 750 and 1000 examinees were used with a 39-item examination. The authors found that with MML estimation item parameter estimates based on normal and empirical prior ability distributions were quite similar, but both differed somewhat from estimates based on the uniform distribution. Bayesian procedures, however, resulted in similar item parameter estimates for the three ability distributions. Sample size had an effect on estimates of item parameters, particularly discrimination, with substantial increases in between-method correlations and decreases in between-method differences for an increase in sample size from 250 to 500. Increasing the sample size further had little effect on correlations.

In the majority of estimation studies the stability of item parameters has been investigated. A study which offered important implications for the present research was Hulin, Lissak, and Drasgow's (1982) investigation of the stability of item characteristic curves. The researchers examined the effect of number of items (15, 30, and 60) and number of examinees (200, 500, 1000, and 2000) on the recovery of ICCs, using the JML procedure of LOGIST, for both the two- and three-parameter models. For their criterion, they used the root mean square error (RMSE), an index which measures the difference in area between ICCs.

Their results showed very accurate recovery of the ICCs for the two-parameter model for 30 and 60 items, but less accurate estimation for 15 items with samples of 1000 and 2000 examinees. For 15 items and 200 examinees, the error of estimation (RMSE) was very large.

With the three-parameter model, however, ICCs were not estimated as accurately as with the two-parameter. For 30 and 60 items with 1000 or 2000 examinees, the average RMSE was greater than .05; for 60 items and 200 or 500 examinees, the average RMSE was greater than .06. All other combinations of items and examinees had much larger RMSE indices.

For the three-parameter model, effects of sample size on the accuracy of estimation were substantial. Decreasing sample size from 2000 to 200 resulted in RMSE indices greater than 4.0. Hulin et al. concluded that the accuracy required by the estimation of item and ability parameters obviously depends on the questions being studied by the investigator. For example, ICCs are important in item bias; if ICCs are important, a large number of examinees are needed but not a large number of items. Their results led them to conclude that for accurate recovery of ICCs with the two-parameter model, 30 items with 500 examinees is acceptable. However, for the three-parameter model, their recommendation was for at least 30 items and 1000 examinees to accurately recover ICCs.

Thissen and Steinberg (1984) noted that the MML estimation procedure used in MULTILOG should be consistent considering the number of examinees alone if the model is correct, unlike procedures which make use of point estimates of θ in the estimation of item parameters and therefore require both large numbers of examinees and large numbers of items. Thus, as the number of examinees becomes large, the estimates for four items or 12 or 35 should all converge to the true values. They confirmed that OCCs for the four-item subset were very similar in appearance to those for the 35-item subset. No measure of the difference between OCCs was calculated, however, and differences were apparent, especially for regions of the OCC below $\theta = -2$. One possible explanation for this difference was that the model may be incorrect in that the latent dimension theta may be defined slightly differently in the set of four difficult items than in the entire test. Hence, the effect of the difficulty distribution of the item set appeared to be an additional factor worthy of investigation.

The factors which might be expected to affect the variability of the option characteristic curves, based on the literature reviewed, were identified as number of items, number of examinees, difficulty distribution of the item set and ability distribution of the sample of examinees (see Table 1). In addition, the OCCs were

Table 1

Factors Affecting Item Parameter Estimation

Estimation Procedure	Study	Sample Size	Test Length	Item Diff.	Abil. Dist.
JML	Cook et al. (1982)	---	---	---	Y
	Hulin et al. (1982)	Y	Y	---	---
	Qualls & Ansley (1985)	Y	Y	Y	---
	Ree (1979)	---	---	---	Y
	Ree & Jensen (1980)	Y	Y	---	---
	Swaminathan & Gifford (1979)	Y	Y	---	Y
	Swaminathan & Gifford (1983)	Y	Y	---	Y
	Swaminathan & Gifford (1985)	Y	Y	---	---
	Swaminathan & Gifford (1986)	Y	Y	---	---
	Wingersky & Lord (1984)	Y	Y	---	Y
	Yen (1987)	---	Y	Y	Y

Note. Y indicates that an effect was found for this factor; N that no effect was found.

Table 1 (cont.)

Estimation Procedure	Study	Sample Size	Test Length	Item Diff.	Abil. Dist.
MML	Bock & Aitken (1981)	---	---	---	Y
	Buhr & Algina (1986)	---	Y	---	---
	Qualls & Ansley (1985)	Y	Y	Y	---
	Swaminathan & Gifford (1986)	Y	N	---	---
	Thissen & Steinberg (1984)	---	N	---	---
	Yen (1987)	---	N	N	Y
BE	Buhr & Algina (1986)	---	Y	---	---
	Swaminathan & Gifford (1985)	Y	Y	---	---

expected to vary with the response option. The investigation of the effect of these factors on the variability of the OCCs produced by the multiple-category scoring model implemented by MULTILOG was the focus of this study.

CHAPTER III
METHOD

Purpose

The purpose of this study was to examine the variability of the item option characteristic curves (OCCs) produced by the multiple-category scoring model (Thissen & Steinberg, 1984). The effects of number of items, number of examinees, difficulty distribution of the item set, ability distribution of the sample of examinees, and response option on the variability of the option characteristic curves were investigated.

This chapter describes the methods and procedures used to investigate the above effects. The following areas are discussed: (a) description of the data sets, (b) calibration of the item parameters, (c) comparison of the option characteristic curves, (d) selection of the item sets and samples, and (e) data analyses.

Description of the Data Sets

Data for the study were drawn from two recent administrations of the Florida College Level Academic Skills Test (CLAST) Computation subtest. This test is a minimum competency examination required for college

sophomores to advance to upper class status or for students in community colleges to receive an Associate of Arts degree in state-supported institutions. There are 50 scored items on the Computation test. Raw score mean and standard deviation for the 21,331 first-time examinees on the first administration were 37.13 and 6.68, respectively. For the second administration, raw score mean and standard deviation for the 11,385 examinees were 35.02 and 7.65. Reliability (Kuder-Richardson Formula 20) coefficients for the two administrations were .83 and .85. Raw score distributions for the two administrations are given in Table 2.

The distribution of traditional item difficulties (p-values) for the two administrations is given in Table 3. This test was chosen for the study because of the availability of large numbers of examinees and the relative difficulty of the Computation subtest as compared to the other subtests of the CLAST.

The Computation subtest is divided into five broad content areas--algebra, arithmetic, geometry and measurement, logical reasoning, and statistics, including probability. Each of these content areas was considered to constitute a homogeneous item set.

Table 2

Raw Scores on the Computation Subtest

Score Range	Administration 1		Administration 2	
	Number	Percent	Number	Percent
0- 5	1	0.0	0	0.0
6-10	10	0.0	6	0.0
11-15	68	0.3	100	0.9
16-20	309	1.4	397	3.5
21-25	944	4.4	943	8.3
26-30	2151	10.1	1631	14.3
31-35	4252	19.9	2398	21.1
36-40	6154	28.9	2865	25.2
41-45	5729	26.9	2293	20.1
46-50	1713	8.1	752	6.6
Total	21331	100.0	11385	100.0

Table 3

Item p-Values For the CLAST Computation Tests

P-Value	Number of Items For Administration One	Two
.90 -1.00	11	10
.80 - .89	9	12
.70 - .79	15	7
.60 - .69	4	9
.50 - .59	7	6
.40 - .49	2	3
.30 - .39	1	2
.20 - .29	1	1
Total	50	50

Calibration of Item Parameters

MULTILOG, Thissen and Steinberg's (1984) multiple-category scoring model, was used to estimate the item parameters. According to this model

$$P(x_g = h|\theta;\underline{a},\underline{c},\underline{d}) = \frac{\exp(z_h) + d_h\exp(z_0)}{\sum_{k=0}^{m} \exp(z_k)}$$

where there are m response functions $z_k = a_k\theta + c_k$ for the m categorical responses, and h takes the values $1,2,\ldots,m_g$. Use of this model by MULTILOG produces fourteen parameters for each four-alternative item, an a, c, and DK (don't know) parameter for each of the four response options and an a and c parameter for the DK or zero category (z_0).

Two options were exercised in the estimation procedure. For the first option, the number of iterations (cycles) to achieve convergence was set at 25, as recommended by Crocker (1987) for use with data from minimum competency examinations. In addition, the maximum acceleration parameter was set at -1.0 to increase the speed of convergence, as suggested in Thissen and Steinberg (1984).

MULTILOG does not produce option characteristic curves. Thus, option characteristic curves were produced by substituting the item parameter estimates from MULTILOG into the equation above and calculating $P_g(\theta)$ for theta values at intervals of .05 from $\theta = -3$ to $\theta = +3$.

The Effect of Number of Examinees and Items

Selection of Item Sets and Samples

To study the effect of number of items and number of examinees apart from the difficulty or discrimination of the items, an item set of ten algebra items which were relatively homogeneous in difficulty and discrimination was selected from the first administration. For the regular test administration, BICAL (Wright, Mead, & Bell, 1980) is always used to estimate Rasch one-parameter item difficulties. For this study, however, three-parameter item difficulty and discrimination estimates were calculated for the entire test by using the "pseudo-Bayesian" estimation procedure ASCAL (Assessment Systems Corporation, 1987) with a random sample of 2000 examinees. These item parameter estimates were employed to select items similar in difficulty and discrimination for the subsets of varying lengths.

To investigate the effect of differing numbers of items on the estimation of the OCCs, eight-, six- and four-item subsets were formed from the ten-item subset, with four items common to the four subsets. Because it was possible that observed effects would differ depending on the characteristics of the particular set of items chosen, this part of the study was replicated with a different set of eight-, six- and four-item subsets formed from the ten

items. As shown in Table 4, items 1, 4, 8, and 10 were the four common items in the first replication; items 2, 5, 6 and 9 were the four common items in the second replication. Table 4 also contains the item statistics computed from test data for the complete test for the algebra item sets of various lengths.

To investigate the effect of number of examinees on the variability of the OCCs, ten random samples of 500 examinees were drawn. This procedure was then repeated to select ten random samples of 1000 examinees each. Samples of 500 and 1000 were chosen because 1000 is generally the recommended minimum number of examinees for calibration studies, and 500 is a realistic maximum number that might be expected for examinees of minority subgroups in a single large-scale test administration. Because the detection of differential item functioning for examinees of different ethnic groups is a potential future application of multiple-category IRT, the variability among OCCs for these sample sizes was of interest.

Responses of these examinees to each of the item subsets above were used to estimate parameters for each of four options (A, B, C, and D) for each item. Parameters were estimated through the multiple-category scoring model of MULTILOG.

Table 4

Item Statistics for Subsets of Varying Lengths

					Rep. 1			Rep. 2		
			3-Parameter		No. of Items			No. of Items		
Item	pVal	PBis	Diff	Disc	8	6	4	8	6	4
01	.57	.33	0.70	1.20	X	X	X	X	X	
02	.65	.38	0.44	1.40				X	X	X
03	.65	.40	0.22	1.03	X	X				
04	.69	.48	0.16	1.40	X	X	X	X	X	
05	.71	.32	0.08	1.30	X			X	X	X
06	.72	.37	0.04	1.21	X			X	X	X
07	.70	.49	-0.15	1.25	X	X		X	X	
08	.78	.46	-0.26	1.36	X	X	X			
09	.83	.50	-0.36	1.63				X	X	X
10	.83	.44	-0.40	1.37	X	X	X	X		
Mean pVal					.71	.70	.72	.71	.72	.73
Mean PBis					.41	.43	.43	.41	.42	.39
Mean Diff					.05	.05	.05	.06	.04	.05
Mean Disc					1.26	1.26	1.33	1.34	1.37	1.39

Note. X indicates the item is included in the specified subset.

Design of the Study

The design for the study, with four levels of test length, two levels of sample size, and ten samples for each combination of test length and sample size, resulted in 80 separate calibrations of item parameters through the multiple-category scoring procedure of MULTILOG (see Table 5) for each replication above. For each of these calibrations, 14 parameters were estimated for each of the four-option items.

Table 5

Design for Investigating the Effect of Number of Items and Number of Examinees

	Number of Examinees	
Number of Items	500	1000
4	$s_1 \cdots\cdots s_{10}$	$s_{11} \cdots s_{20}$
6	$s_1 \cdots\cdots s_{10}$	$s_{11} \cdots s_{20}$
8	$s_1 \cdots\cdots s_{10}$	$s_{11} \cdots s_{20}$
10	$s_1 \cdots\cdots s_{10}$	$s_{11} \cdots s_{20}$

Note. S indicates the sample; for example, s_1 is sample 1.

For the four common items in each subset, option characteristic curves were generated from the MULTILOG

examinees under each of the eight combinations of sample size and test length. The 80 different OCCs were obtained for each option, and the question of interest was whether there were differences in variability among the ten OCCs generated under the eight different treatment conditions.

Analyses of Data

Differences between option characteristic curves for a particular set of conditions were calculated through use of a procedure suggested by Brown and Forsythe (1974). Although a number of procedures are available for testing the equality of variances, the Brown-Forsythe procedure was selected because it has been shown to be robust under conditions of nonnormality (Olejnik & Algina, 1987).

With the Brown-Forsythe procedure, the absolute value of the difference between each observation on the dependent variable and its group median, $d_{ij} = |x_{ij} - m_j|$, is used as the data in an analysis of variance. In this study, the Brown-Forsythe procedure was adapted to create an index for ten OCCs as follows.

$X_{ij} = P_g(\theta)$ for a given option for a given θ for sample i under condition j, where each condition was one combination of test length and sample size. For a given level of an experimental variable (i), there were ten x_{ij} values; i.e., one for each of the ten samples. Thus, d_{ij} was the absolute value of the distance from a sample OCC to

the median of the ten sample OCCs at a particular value of θ. For each OCC, 601 d_{ij} values (at θ values from -3.00 to +3.00) were computed. These 601 d_{ij} for each OCC were then averaged, and these averages were used as observations on the dependent variable in an ANOVA to test for equality of variability in the OCCs under the different experimental conditions.

These average differences were the dependent variables in a repeated measures multifactor analysis of variance (Biomedical Computer Program, 1988, p. 1055), where the independent factors to be investigated were sample size (500 or 1000 examinees), test length (4, 6, 8, or 10 items), and option (1 to 4), with sample size specified as the between group factor and test length and option specified as repeated measures or within group factors. The model for the analysis was

$$Y_{jkm(i)} = \mu + \alpha_i + \beta_j + \delta_k + \tau_{m(i)} + \alpha\beta_{ij} + \alpha\delta_{ik} + \beta\delta_{jk} + \beta\tau_{jm(i)} + \delta\tau_{km(i)} + \alpha\beta\delta_{ijk} + \epsilon_{jkm(i)}$$

where μ = grand mean

α_i = effect of number of examinees, i = 1, 2

β_j = effect of number of items, j = 1 to 4

δ_k = effect of option, k = 1 to 4, and

$\tau_{m(i)}$ = effect of samples nested within number of examinees, m = 1 to 10.

Four separate analyses of variance, one for each of the four items, were conducted for each of the two

Four separate analyses of variance, one for each of the four items, were conducted for each of the two replications, with the overall significance level for each set of four tests set at .10, or an alpha level of .025 for each analysis. For these analyses, the degrees of freedom for the within group factors were reduced by means of the Huynh-Feldt (1976) method to correct for possible failure to meet the sphericity assumptions.

The Effect of Item Difficulty and Examinee Ability

Selection of Item Sets and Samples

For the study of the effect of item difficulty on the option characteristic curves, two five-item subsets were formed from the algebra subtest from the second administration, with three items common to both subsets (Table 6).

For subset 1, items were selected so that the average difficulty was approximately equal to the mean ability estimate for the test. For subset two, two easy items replaced two moderately difficult items, so that mean difficulty was approximately zero.

To investigate the effect of ability distribution on the option characteristic curves, two sets of ten samples of 1050 examinees each were selected, with replacement, from the testing population. The first ten samples were stratified random samples. Examinees were selected proportionately from each of twelve score groups to form an

approximately normal distribution, centered at an ability estimate of zero. Table 7 presents the descriptive statistics for the ten normal samples.

Table 6

Item Statistics for Item Subsets Differing in Difficulty

		Traditional		Three-Parameter	
Subset	Item	p-Value	Pt.Bis.	Diff.	Disc.
1	01*	.48	.35	0.780	1.086
	02*	.51	.44	0.594	1.452
	03	.54	.43	0.146	0.842
	04*	.57	.37	0.101	0.800
	05	.66	.41	-0.100	1.054
Mean		.55	.40	0.304	1.047
2	01*	.48	.35	0.780	1.086
	02*	.51	.44	0.594	1.452
	03*	.57	.37	0.101	0.800
	04	.76	.46	-0.629	1.011
	05	.79	.48	-0.765	1.066
Mean		.62	.42	0.016	1.083

*Items common to the two subsets

Table 7

Descriptive Statistics for the Normal Distributions

Sample[a]	Raw Score Mean	Standard Deviation	Skewness	Kurtosis
1	33.83	6.56	-0.11	-0.30
2	33.80	6.54	-0.10	-0.32
3	33.77	6.55	-0.07	-0.33
4	33.79	6.53	-0.09	-0.31
5	33.81	6.56	-0.09	-0.30
6	33.82	6.56	-0.13	-0.29
7	33.80	6.55	-0.09	-0.30
8	33.85	6.56	-0.12	-0.31
9	33.82	6.56	-0.12	-0.30
10	33.81	6.57	-0.10	-0.34

[a] n = 1050 examinees in each sample.

In the second set, ten simple random samples were selected. For this relatively easy examination given to very able examinees, the ability distribution was negatively skewed. With random sampling, the ability distribution was also negatively skewed in each of these ten samples. Thus, mean difficulty in the first set of more difficult items would approximate mean ability in these skewed

distributions. Table 8 presents the descriptive statistics for the ten random samples.

Table 8

Descriptive Statistics for the Random Samples

Sample[a]	Raw Score Mean	Standard Deviation	Skewness	Kurtosis
11	35.29	7.44	-0.55	-0.04
12	35.02	7.58	-0.46	-0.40
13	34.95	7.42	-0.41	-0.35
14	35.38	7.49	-0.47	-0.24
15	35.20	7.74	-0.56	-0.18
16	35.29	7.71	-0.56	-0.13
17	34.50	7.48	-0.40	-0.43
18	34.44	7.87	-0.40	-0.54
19	35.02	7.48	-0.44	-0.46
20	35.02	7.47	-0.38	-0.46

[a] n = 1050 examinees in each sample.

Design of the Study

The crossing of difficulty distribution by ability distribution resulted in 40 separate calibrations with MULTILOG as shown in Table 9. Within each of the cells in the design, difference measures between OCCs for each of

the options for each of the items were calculated by means of the Brown-Forsythe procedure, as in the preceding portion of the study. The differences were averaged for each option for each item and used as the dependent variables in the analysis of variance.

Table 9

Design for the Study of the Effect of Item Difficulty and Examinee Ability Distribution

	Difficulty Distribution	
Ability Distribution	Mean Difficulty Level 0.30	0.01
Normal	$s_1 \cdots\cdots s_{10}$	$s_1 \cdots\cdots s_{10}$
Negatively Skewed	$s_{11} \cdots\cdots s_{20}$	$s_{11} \cdots\cdots s_{20}$

Note. S indicates sample; i.e., s_1 is sample 1.

Analyses of Data

As in the previous part of the study, a repeated measures multifactor analysis of variance was performed for each item, with difficulty of the item set (easy and difficult), ability distribution of the examinees (normal and skewed), and option (1 to 4) the factors to be investigated. Ability distribution served as the between group

factor; item difficulty and option were repeated measures or within group factors.

The model for the analysis was

$$Y_{jkm(i)} = \mu + \alpha_i + \beta_j + \delta_k + \tau_{m(i)} + \alpha\beta_{ij} + \alpha\delta_{ik} + \beta\delta_{jk} + \beta\tau_{jm(i)} + \delta\tau_{km(i)} + \alpha\beta\delta_{ijk} + \epsilon_{jkm(i)}$$

where μ = grand mean

α_i = effect of ability distribution, i = 1, 2

β_j = effect of difficulty distribution, j = 1, 2

δ_k = effect of option, k = 1 to 4

$\tau_{m(i)}$ = effect of samples nested within number of examinees, m = 1 to 10.

Three separate analyses of variance, one for each of the three items, were conducted, with the overall significance level for the family of tests set at .10. Thus, the significance level for each of the three tests was .033. The degrees of freedom for the within group factor main effects and interactions were reduced by means of the Huyhn-Feldt procedure to correct for violations of the sphericity assumption.

CHAPTER IV
RESULTS

The purpose of this study was to investigate the variability of the option characteristic curves produced by the multiple-category scoring procedure of MULTILOG under a variety of controlled conditions. The specific questions to be addressed were

1. Is the variability of the option characteristic curves affected by the number of examinees in the sample?
2. Is the variability of the option characteristic curves affected by the number of items in the item set?
3. Does the difficulty distribution of the item set affect the variability of the option characteristic curves?
4. Does the ability distribution of the sample of examinees affect the variability of the option characteristic curves?
5. Does the variability of the option characteristic curves differ depending upon the option?

The multifactor repeated measures analysis of variance procedure (Biomedical Computer Program, 1988, p. 1055) was

used to investigate the above questions. The design and model for the analyses were given in Chapter III. The dependent variable in the analysis was the average Brown-Forsythe statistic ($\bar{d}_{ij}$) for each OCC estimated under each of the conditions. For each OCC, this "average difference" represents the amount of variance between the OCC and the median of all ten OCCs estimated under the same set of conditions. Larger average differences indicate greater variability in the OCC from the other OCCs.

The data for each item were analyzed separately, with an overall significance level for the set of four items in each replication set at .10. Huynh-Feldt adjusted p-values were used for the repeated measures factors. The results of the analyses of data, grouped by category to answer each of the above questions, are presented in this chapter.

Effect of Number of Items and Examinees for Replication 1

Results for the analyses of variance for the four common items in replication 1 are given in Tables 10 through 13. There were a number of significant main effects, one significant two-way interaction, and no significant three-way interactions. These results are described in the following sections. Because the effects of number of examinees and number of items were of primary interest in this study, the effects of these experimental conditions are considered first.

Table 10

Analysis of Variance Table for Item 1 in Part I

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0401	.0401	19.53*
Samples within NE/S(NE)	18	.0369	.0020	
Number of Items/NI	2.95	.0074	.0025	1.87
NE x NI Interaction	2.95	.0029	.0010	0.73
NI x S(NE) Interaction	53.05	.0712	.0013	
Option/OP	2.91	.0101	.0034	2.29
NE x OP Interaction	2.91	.0019	.0006	0.43
OP x S(NE) Interaction	52.30	.0796	.0015	
NI x OP Interaction	6.14	.0037	.0004	0.59
NE x NI x OP Interaction	6.14	.0053	.0006	0.86
NI x OP x S(NE)	110.53	.1123	.0007	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.025$.

Table 11

Analysis of Variance Table for Item 4 in Part I

Source of Variation	df	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0178	.0178	10.36*
Samples within NE/S(NE)	18	.0309	.0017	
Number of Items/NI	3	.0142	.0047	4.85*
NE x NI Interaction	3	.0017	.0006	0.58
NI x S(NE) Interaction	54	.0525	.0010	
Option/OP	2.35	.0474	.0158	13.11*
NE x OP Interaction	2.35	.0061	.0020	1.70
OP x S(NE) Interaction	42.30	.0650	.0012	
NI x OP Interaction	6.59	.0043	.0005	0.78
NE x NI x OP Interaction	6.59	.0079	.0009	1.44
NI x OP x S(NE)	118.61	.0995	.0006	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.025$.

Table 12

Analysis of Variance Table for Item 8 in Part I

Source of Variation	df	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0286	.0286	13.98*
Samples within NE/S(NE)	18	.0368	.0020	
Number of Items/NI	3	.0103	.0034	3.23
NE x NI Interaction	3	.0012	.0004	0.37
NI x S(NE) Interaction	54	.0577	.0011	
Option/OP	2.87	.0314	.0105	5.48*
NE x OP Interaction	2.87	.0043	.0014	0.76
OP x S(NE) Interaction	51.57	.1030	.0019	
NI x OP Interaction	5.53	.0073	.0008	1.29
NE x NI x OP Interaction	5.53	.0085	.0009	1.50
NI x OP x S(NE)	99.62	.1015	.0006	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.025$.

Table 13

Analysis of Variance Table for Item 10 in Part I

Source of Variation	df	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0436	.0436	23.77*
Samples within NE/S(NE)	18	.0330	.0018	
Number of Items/NI	2.76	.0069	.0023	1.74
NE x NI Interaction	2.76	.0045	.0015	1.15
NI x S(NE) Interaction	49.60	.0711	.0013	
Option/OP	2.87	.0794	.0265	27.33*
NE x OP Interaction	2.87	.0135	.0045	4.65*
OP x S(NE) Interaction	51.62	.0523	.0010	
NI x OP Interaction	5.01	.0052	.0006	0.89
NE x NI x OP Interaction	5.01	.0043	.0005	0.74
NI x OP x S(NE)	90.24	.1056	.0006	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.025$.

The significant interaction involved options and number of examinees. The followup analysis for this interaction is described in the section on number of examinees.

Effect of Number of Examinees

For each of the four items in replication 1, the mean "average difference" ($\bar{d}_{ij}$) was significantly different for samples of different numbers of examinees, with the average difference for OCCs estimated with samples of 500 examinees significantly larger than the average difference for OCCs estimated with samples of 1000 examinees (see Table 14).

Table 14

Mean Average Differences for Number of Examinees

	Number of Examinees	
Item	500	1000
1	.0580[a]	.0356[b]
4	.0546[a]	.0397[b]
8	.0570[a]	.0368[b]
10	.0554[a]	.0321[b]

Note. For each item, means with different superscripts differ significantly at $p<.025$.

As an example, Table 14 shows that the mean average difference for the OCCs for Item 1 estimated with samples of 500 examinees was .0580, while that for the OCCs estimated with samples of 1000 examinees was .0356. This indicates that the variability among OCCs estimated with samples of 500 examinees was significantly greater than the variability among OCCs estimated with samples of 1000 examinees.

For item 10, there was a significant interaction between number of examinees and option, $F(2.87,51.62)=4.65$, $p<.025$. The mean average differences for each sample size by option are given in Table 15.

Table 15

Mean Differences for Number of Examinees by Option Interaction for Item 10

	Option			
Number of Examinees	A	B	C	D
500	.0622	.0520	<u>.0842</u>	.0234
1000	.0309	.0313	<u>.0464</u>	.0183

Note. Means for the correct response are underlined.

To investigate the interaction between number of examinees and option, a test for simple main effects was utilized (Kirk, 1968, pp. 263-266). First the analysis of variance procedure was used to investigate the simple main effect of number of examinees for options A, B, C, and D. Then the simple main effect of option for samples of 500 and 1000 examinees was investigated. The error terms for the analysis were formed by pooling the appropriate error terms from the original analysis of variance for item 10. The test for simple main effects showed a significant effect of number of examinees for all options except option D (see Table 16). Additionally, the simple main effect of option was significant for sample sizes of both 500 and 1000 examinees.

As Figure 3 illustrates, although there is a significant interaction for item 10, the differences between means for the 500 and 1000 examinee samples for all options are in the same direction. Thus, it seems reasonable to interpret the main effects for this item.

Effect of Number of Items

The mean average differences for number of items are presented in Table 17. There was a significant effect due to the number of items for just one of the items, item 4, at an alpha level of .10 for the four tests. For the other three items, the average difference did not differ

Table 16

Analysis of Variance Table for Simple Main Effects For Item 10

Source of Variation	df	Sum of Squares	Mean Square	F
Number of Examinees - Op A	1	.0196	.0196	16.60*
Number of Examinees - Op B	1	.0086	.0086	7.25*
Number of Examinees - Op C	1	.0286	.0286	24.22*
Number of Examinees - Op D	1	.0005	.0005	0.45
Pooled Error Term	72	.0853	.0012	
Option - 500 examinees	3	.0646	.0215	21.50*
Option - 100 examinees	3	.0325	.0108	10.82*
Error	54	.0523	.0010	

*Significant at p<.01.

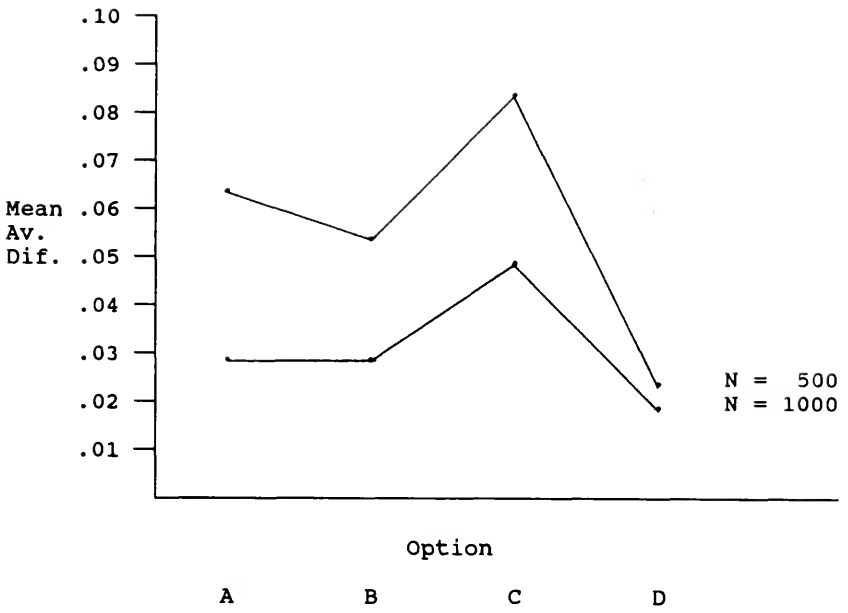

Figure 3. Interaction Between Option and Number of Examinees for Item 10

Table 17

Mean Average Differences for Number of Items

	Number of Items			
Item	4	6	8	10
1	.0510	.0522	.0423	.0418
4	.0540[a]	.0516[a]	.0462[ab]	.0367[b]
8	.0467	.0552	.0465	.0393
10	.0509	.0441	.0420	.0380

Note. For each item, means with different superscripts differ significantly at $p<.004$.

depending upon the number of items used to estimate the OCCs.

For item 4, followup tests were done by means of the Bonferroni procedure, with an overall significance level of .025 for the set of six tests, to make pairwise comparisons of means for subsets of different numbers of items. The mean difference measure was significantly larger for OCCs estimated with four items and with six items than for those estimated with 10 items. There was no significant difference between any of the other item subsets (see Table 17).

Effect of Option

For item 10, there was a significant interaction between option and number of examinees, as previously discussed. For this item, the test of the simple main effect of option was significant for both conditions (500 and 1000 examinees). For two other items, item 4 and 8, there were significant effects of option. Pairwise comparisons of means showed that, for items 4, 8, and 10, the mean average difference for the correct option was significantly greater than the means for the incorrect options (see Table 18).

Table 18

Mean Average Differences for Option

	Option			
Item	A	B	C	D
1	.0424	.0565	.0444	.0440
4	.0661[a]	.0493[b]	.0349[b]	.0382[b]
8	.0381[a]	.0465[a]	.0619[b]	.0412[a]
10	.0456[a]	.0434[a]	.0653[b]	.0208[c]

Note. For each item, means with different superscripts differ significantly at $p<.004$. Means for the correct response are underlined.

For these items, then, variability among OCCs was greater for the correct options than for the incorrect options.

However, the main interest for users of MULTILOG is in the estimation of the OCCs for the incorrect options. For this set of items, only item 10 showed a significant difference between means for incorrect options, when pairwise comparisons were conducted by means of the Bonferroni procedure. The mean average difference for option D, the option chosen by the smallest percentage of examinees, differed significantly from those for the other two incorrect options (see Table 18).

Effect of Number of Items and Examinees for Replication 2

Tables 19 through 22 contain the results of the analyses of variance for the four items in the second replication. There were no significant interactions. The results for sample size and option were similar to those for the first set of four items. However, the results for test length differed for this set of items.

Effect of Number of Examinees

For three of the four items, the effect of number of examinees was significant, at an alpha level of .10 for the set of four tests. Table 23 shows the mean average differences for each item for samples of 500 and 1000 examinees.

Table 19

Analysis of Variance Table for Item 2 in Part I

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0167	.0167	8.31*
Samples within NE/S(NE)	18	.0361	.0020	
Number of Items/NI	2.47	.0124	.0041	4.35*
NE x NI Interaction	2.47	.0040	.0013	1.40
NI x S(NE) Interaction	44.46	.0515	.0010	
Option/OP	3	.0380	.0127	9.78*
NE x OP Interaction	3	.0022	.0007	0.57
OP x S(NE) Interaction	54	.0700	.0013	
NI x OP Interaction	7.55	.0037	.0004	0.80
NE x NI x OP Interaction	7.55	.0057	.0006	1.22
NI x OP x S(NE)	135.96	.0838	.0005	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p < .025$.

Table 20

Analysis of Variance Table for Item 5 in Part I

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0201	.0201	11.00*
Samples within NE/S(NE)	18	.0330	.0018	
Number of Items/NI	2.34	.0109	.0036	3.64
NE x NI Interaction	2.34	.0020	.0007	0.68
NI x S(NE) Interaction	42.19	.0541	.0010	
Option/OP	2.25	.0671	.0224	18.84*
NE x OP Interaction	2.25	.0106	.0035	2.97
OP x S(NE) Interaction	40.46	.0641	.0012	
NI x OP Interaction	6.56	.0034	.0004	0.65
NE x NI x OP Interaction	6.56	.0037	.0004	0.70
NI x OP x S(NE)	118.14	.0948	.0006	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.025$.

Table 21

Analysis of Variance Table for Item 6 in Part I

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0226	.0226	5.48
Samples within NE/S(NE)	18	.0742	.0041	
Number of Items/NI	2.21	.0048	.0016	1.36
NE x NI Interaction	2.21	.0006	.0002	0.17
NI x S(NE) Interaction	39.79	.0638	.0012	
Option - OP	2.89	.1431	.0477	26.65*
NE x OP Interaction	2.89	.0034	.0011	0.63
OP x S(NE) Interaction	52.03	.0967	.0018	
NI x OP Interaction	5.50	.0060	.0007	1.12
NE x NI x OP Interaction	5.50	.0028	.0003	0.52
NI x OP x S(NE)	99.01	.0972	.0006	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.025$.

Table 22

Analysis of Variance Table for Item 9 in Part I

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Number of Examinees/NE	1	.0277	.0277	18.11*
Samples within NE/S(NE)	18	.0275	.0015	
Number of Items/NI	3	.0062	.0021	3.40*
NE x NI Interaction	3	.0048	.0016	2.64
NI x S(NE) Interaction	54	.0331	.0006	
Option - OP	2.72	.0373	.0124	9.83*
NE x OP Interaction	2.72	.0023	.0008	0.61
OP x S(NE) Interaction	49.03	.0683	.0013	
NI x OP Interaction	6.97	.0028	.0003	0.67
NE x NI x OP Interaction	6.97	.0020	.0002	0.48
NI x OP x S(NE)	121.93	.0762	.0005	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.025$.

Table 23

Mean Average Differences for Number of Examinees

Item	500 Examinees	1000 Examinees
2	.0558[a]	.0414[b]
5	.0553[a]	.0395[b]
6	.0544	.0376
9	.0529[a]	.0343[b]

Note. For each item, means with different superscripts differ significantly at $p<.025$.

In all cases, the mean average differences for OCCs estimated with samples of 500 examinees were larger than those for samples of 1000 examinees, with three of the four significantly larger. This indicates greater variability among OCCs estimated with samples of 500 examinees than among those estimated with samples of 1000 examinees.

Effect of Number of Items

The mean average differences for number of items are given in Table 24. For two of the four items, there were significant differences due to number of items in the subset, with larger average differences for the smaller item sets. Thus, greater variability among OCCs was associated with the shorter test lengths. Bonferroni tests

of comparison were used to determine for which test lengths the means were significantly different, at an overall alpha level of .025 for the set of six comparisons.

For item 2, the mean average differences for four and ten items were significantly different. However, the means for six and eight items did not differ from that for ten items or the mean for four items (see Table 24).

For item 9, the mean average difference for four items was significantly larger than that for eight items but not that for six and ten items. Mean average differences for six, eight and ten items did not differ significantly (see Table 24).

Table 24

Mean Average Differences for Number of Items

	Number of Items			
Item	4	6	8	10
2	.0579[a]	.0477[ab]	.0484[ab]	.0404[b]
5	.0575	.0445	.0442	.0435
6	.0524	.0444	.0452	.0420
9	.0501[a]	.0440[ab]	.0377[b]	.0424[ab]

Note. For each item, means with different superscripts differ significantly at $p<.004$.

Effect of Option

The mean average difference for each option for each of the four items is given in Table 25. There were significant differences among options for all four of the items. The Bonferroni procedure was used to compare means for each of the four options, with the overall significance level set at .025 for the set of six tests. For item 2, the average difference for the correct option (C) was significantly larger than that for incorrect options B and D, but not A. For item 5, the mean for the correct option (A) was significantly larger than that for incorrect options B and D, but not C. For item 6, the mean for the correct option (B) was significantly larger than the means for all three incorrect options. For item 9, the mean for the correct option (A) was significantly larger than the mean for incorrect option C only. The significantly larger means for the correct options indicate greater variability among the OCCs for the correct response than among those for the incorrect responses.

Item 6 was the only item for which there were significant differences between means for the incorrect options. For item 6, the means for option A and D were significantly smaller than that for option C.

Table 25

Mean Average Differences for Option

	Option			
Item	A	B	C	D
2	.0484^{ab}	.0439^{a}	<u>.0659</u>^{b}	.0363^{a}
5	<u>.0669</u>^{a}	.0319^{b}	.0498^{ab}	.0381^{b}
6	.0231^{a}	<u>.0754</u>^{b}	.0565^{c}	.0289^{a}
9	<u>.0597</u>^{a}	.0431^{ab}	.0293^{b}	.0422^{ab}

Note. For each item, means with different superscripts differ significantly at $p<.004$. Means for the correct response are underlined.

To interpret the differences across options, the options for all eight items were arranged in order of the percentage of examinees choosing each response. Table 26 shows the percentage of examinees choosing each option, averaged over the ten samples of 1000 examinees each. Table 27 presents the significant differences among means, ordered by number of examinees choosing each option. Overall, average differences were greatest for the correct option, with the average difference decreasing in direct relation to the number of examinees choosing the option.

Table 26

Percentage of Examinees Choosing Each Option

	Option			
Item	A	B	C	D
1	.147	.567	.124	.156
4	.677	.155	.046	.118
8	.082	.070	.765	.082
10	.045	.098	.817	.034
2	.149	.125	.643	.080
5	.707	.090	.053	.146
6	.026	.710	.224	.037
9	.824	.047	.028	.101

Note. Percentages for the correct response are underlined.

Table 27

Means for Option Arranged in Order of Percent of Examinees Choosing Each Option (With Option 1 the Largest Percentage)

	Option			
Item	1	2	3	4
1	.0565	.0440	.0424	.0444
4	.0661[a]	.0493[b]	.0382[b]	.0349[b]
8	.0619[a]	.0412[b]	.0381[b]	.0465[b]
10	.0653[a]	.0434[b]	.0456[b]	.0208[c]
2	.0659[a]	.0484[ab]	.0439[b]	.0363[b]
5	.0669[a]	.0381[b]	.0319[b]	.0498[ab]
6	.0754[a]	.0565[b]	.0289[c]	.0231[c]
9	.0597[a]	.0422[ab]	.0431[ab]	.0293[b]
Overall Mean	.0651	.0452	.0392	.0356

Note. For each item, means with different superscripts differ significantly at $p<.004$. Means for the correct response are underlined.

Effect of Ability and Difficulty Distribution

Separate multifactor repeated measures analyses of variance were conducted for each of the three common items in Part II, with the overall significance level set at .10 for the family of tests. The results for the analyses of variance are given in Tables 28 through 30. The factors investigated were ability distribution of the sample of examinees, difficulty distribution of the item set, and option. There were no significant interactions for any of the factors.

Effect of Ability Distribution

There was a significant effect of ability distribution (normal versus negatively skewed) for one of the three items only. For item 4, the mean average difference for the normal distribution was significantly higher than that for the skewed distribution, $F(1,18)=6.75, p<.033$. This indicates greater variability among the option characteristic curves for the normally distributed samples than for the negatively skewed samples. The mean average differences for the normal and skewed distributions are given in Table 31.

Table 28

Analysis of Variance Table for Item 1 in Part II

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Ability Distribution/AD	1	.0008	.0008	0.42
Samples within AD/S(AD)	18	.0354	.0019	
Difficulty Distrib/DD	1	.0005	.0005	0.23
AD x DD Interaction	1	.0002	.0002	0.10
DD x S(AD) Interaction	18	.0397	.0022	
Option/OP	2.85	.0099	.0033	2.85
AD x OP Interaction	2.85	.0029	.0009	0.83
OP x S(AD) Interaction	50.80	.0628	.0012	
DD x OP Interaction	3	.0004	.0001	0.15
AD x DD x OP Interaction	3	.0043	.0014	1.57
DD x OP x S(AD)	54	.0491	.0009	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.033$.

Table 29

Analysis of Variance Table for Item 2 in Part II

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Ability Distribution/AD	1	.0048	.0048	4.11
Samples within AD/S(AD)	18	.0212	.0012	
Difficulty Distrib/DD	1	.0000	.0000	0.05
AD x DD Interaction	1	.0000	.0000	0.02
DD x S(AD) Interaction	18	.0067	.0004	
Option/OP	2.39	.0088	.0029	3.32
AD x OP Interaction	2.39	.0008	.0003	0.30
OP x S(AD) Interaction	42.99	.0481	.0009	
DD x OP Interaction	3	.0011	.0004	1.21
AD x DD x OP Interaction	3	.0002	.0000	0.24
DD x OP x S(AD)	54	.0164	.0003	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.033$.

Table 30

Analysis of Variance Table for Item 4 in Part II

Source of Variation	df[a]	Sum of Squares	Mean Square	F
Ability Distribution/AD	1	.0055	.0055	6.75*
Samples within AD/S(AD)	18	.0147	.0008	
Difficulty Distrib/DD	1	.0001	.0001	0.04
AD x DD Interaction	1	.0000	.0000	0.00
DD x S(AD) Interaction	18	.0299	.0016	
Option/OP	2.98	.0262	.0087	11.33*
AD x OP Interaction	2.98	.0007	.0002	0.32
OP x S(AD) Interaction	53.58	.0417	.0008	
DD x OP Interaction	2.68	.0011	.0004	0.52
AD x DD x OP Interaction	2.68	.0016	.0005	0.76
DD x OP x S(AD)	48.28	.0372	.0007	

[a]Degrees of freedom for repeated measures factors are adjusted by means of the Huynh-Feldt method.
*$p<.033$.

Table 31

Mean Average Differences for Ability Distribution

	Ability Distribution	
Item	Normal	Skewed
1	.0506	.0551
2	.0474	.0365
4	.0486[a]	.0368[b]

Note. For each item, means with different superscripts differ significantly at $p<.033$.

Effect of Difficulty Distribution

There was no significant effect for type of difficulty distribution (easy versus difficult items) for any of the three items in Part II, at an overall significance level of .10 for the three analyses. Table 32 gives the mean average differences for difficulty distribution for the three items.

Effect of Option

For just one of the three items, item 4, there was a significant effect of option, at an overall significance level of .10 for the set of three tests (see Table 33).

Table 32

Mean Average Differences for Difficulty Distribution

	Difficulty Distribution	
Item	Easy	Difficult
1	.0511	.0547
2	.0416	.0422
4	.0433	.0421

Table 33

Mean Average Differences for Option

	Option			
Item	A	B	C	D
1	.0491	<u>.0664</u>	.0495	.0465
2	<u>.0540</u>	.0382	.0414	.0341
4	$.0250^{a}$	$.0451^{b}$	<u>$.0608^{c}$</u>	$.0398^{b}$

Note. For each item, means with different superscripts differ significantly at $p<.005$. Means for the correct response are underlined.

Bonferroni followup tests showed a significant difference between means for the correct option (C) and all three incorrect options, with the OCCs for the correct option exhibiting significantly more variability than those for the incorrect options. In addition, the mean for incorrect option A was significantly smaller than those for incorrect options B and D, at a significance level of .005.

Differences in the variability of the OCCs for the options were related to the percentage of examinees choosing each response, as in Part I. The incorrect option for which a significant difference occurred was one chosen by only five percent of the examinees (see Table 34). For this option - option A of item 4 - the variability among option characteristic curves was significantly less than for the other incorrect options.

Table 34

Percentage of Examinees Choosing Each Option

	Option			
Item	A	B	C	D
1	.2431	.5068	.1282	.1200
2	.5140	.1707	.1364	.1718
4	.0547	.2214	.6223	.1009

Note. Means for the correct response are underlined.

CHAPTER V
DISCUSSION AND CONCLUSIONS

Discussion

The results of this study indicated that the variability of the option characteristic curves produced by the MULTILOG estimation procedure is affected by some of the same factors that affect estimation of item parameters in the binary latent trait models. Sample size was identified as a factor affecting estimation in a number of studies (Hulin et al, 1982; Qualls & Ansley, 1985; Ree & Jensen, 1980; Swaminathan & Gifford, 1979, 1983, 1985, 1986; and Wingersky & Lord, 1984). The number of examinees needed for stable estimation of item parameters was related to the complexity of the latent trait model. For the one-parameter model, 500 examinees is considered adequate for item parameter estimation. However, for the three-parameter model, the recommended sample size is generally at least 1000 examinees.

Although the complexity of the multiple-category model would seem to require sample sizes of at least 1000 examinees, in practice the sample sizes available are often less. For example, if the multiple-category model is to be used for the study of differential response to options by

examinees of different ethnic groups, it is unlikely that subgroups will be as large as 1000 examinees.

This would appear to be a serious problem in the use of the multiple-category model to study differential option performance. Significant differences in the OCC variability index were found due to the number of examinees in the calibration sample for seven of the eight items studied. The average differences between the option characteristic curves for the ten samples and the curve of median values were significantly larger for samples of 500 examinees than for samples of 1000 examinees.

Figures 4 through 11 illustrate the greater variability for the OCCs estimated with samples of 500 examinees and 10 items than for those estimated with 1000 examinees and 10 items. For example, for Option A, the correct option, nine of the ten curves estimated with 1000 examinees follow the median curve very closely, except for the area below an ability of -2.00, where insufficient examinee data are available for this minimum competency examination (see Figure 4). However, the OCCs estimated with samples of 500 examinees show much more spread across the range of ability and variability in the area above an ability of 2.00 also (see Figure 5).

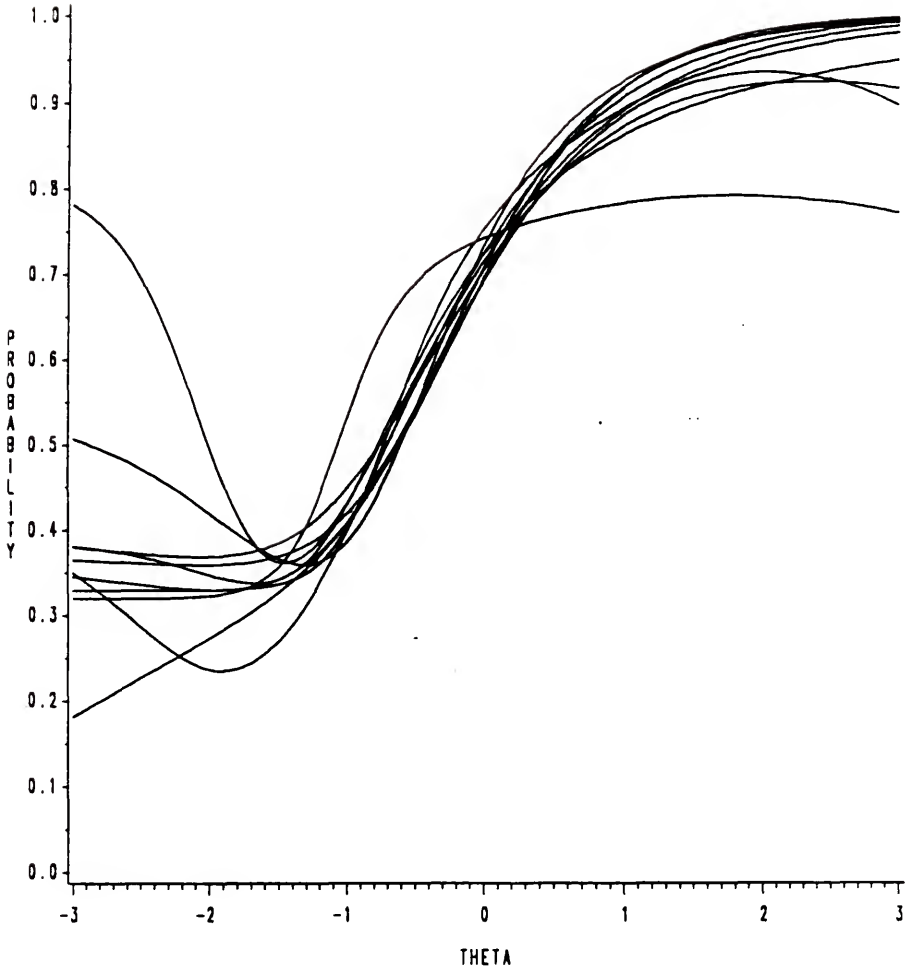

Figure 4. Item 4: Option A OCCs estimated with 1000 examinees and 10 items

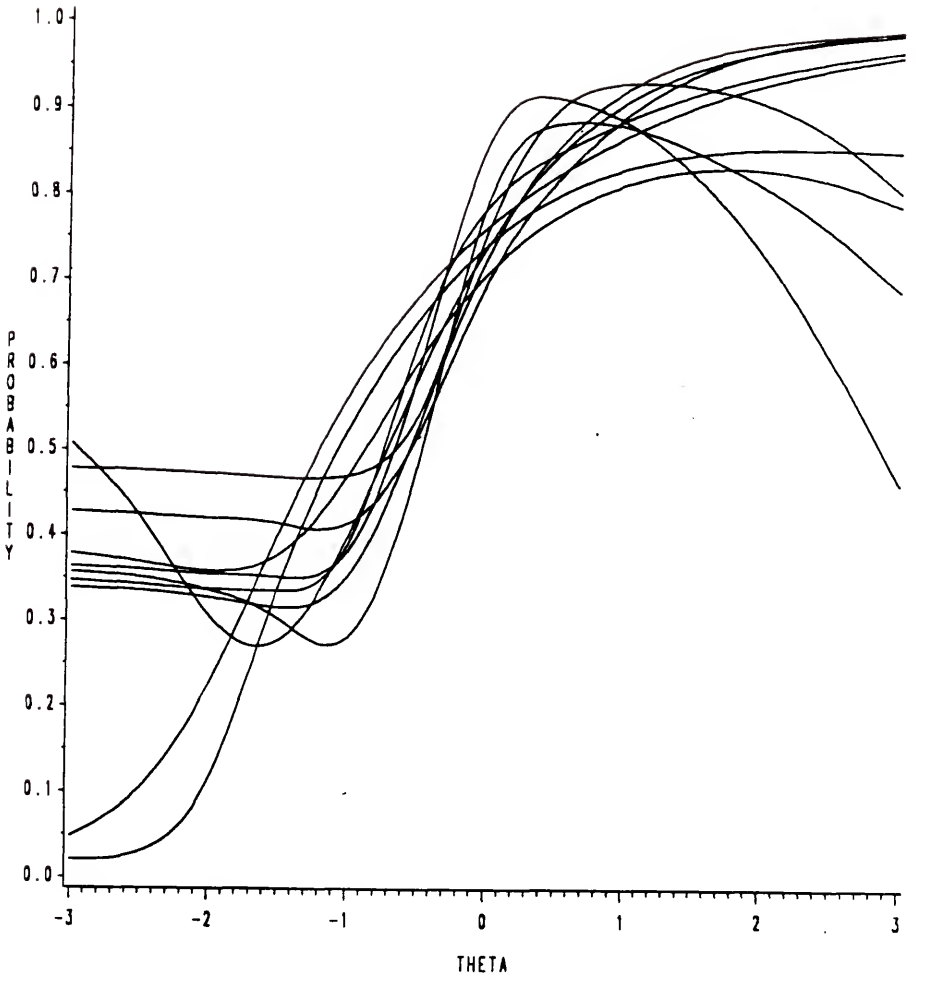

Figure 5. Item 4: Option A OCCs estimated with 500 examinees and 10 items

This is illustrated for incorrect option B for item 4 (Figures 6 and 7). Although there is variability in the ability range of -2 to -3 for the OCCs estimated with 1000 examinees, for the remainder of the ability range the curves follow the same pattern, with the OCCs converging toward a probability of zero at the highest ability. However, the OCCs estimated with 500 examinees also show greater variability in the higher ability ranges. The same pattern of variability in the lower ability ranges occurs for incorrect option C, with greater variability throughout the range of ability for the OCCs estimated with samples of 500 examinees (Figures 8 and 9).

Because Thissen (1976) specifically recommended the multiple-category scoring procedure for low-scoring examinees, the variability in the lower ability ranges is potentially troublesome. However, the effect of this variability in the estimation of the OCCs on the estimation of the ability parameters needs to be examined.

Test length or number of items in the item set was another factor which was identified through the binary-scoring model literature as having an effect on item parameter estimation, with greater variability occurring with smaller numbers of items (Buhr & Algina, 1986; Hulin et al, 1982; Qualls & Ansley, 1985; Ree & Jensen, 1980; Swaminathan & Gifford, 1979, 1983, 1985; Wingersky & Lord, 1984; and Yen, 1987). Since the MULTILOG procedure is

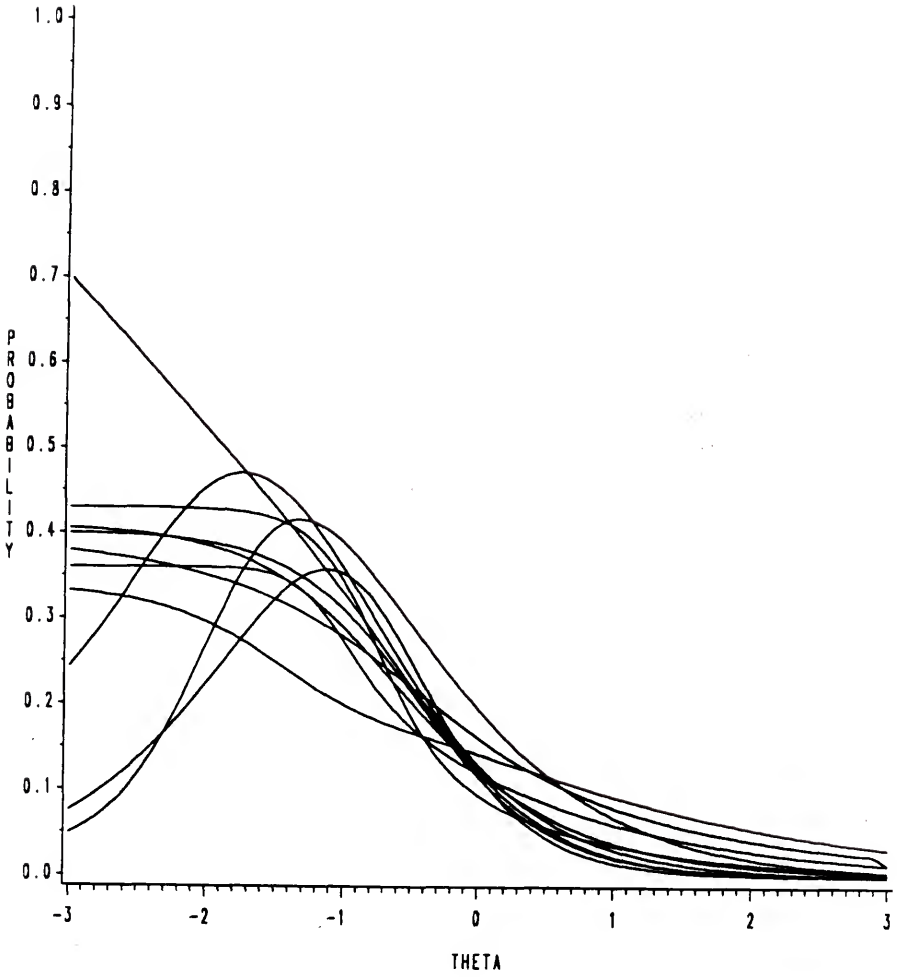

Figure 6. Item 4: Option B OCCs estimated with 1000 examinees and 10 items

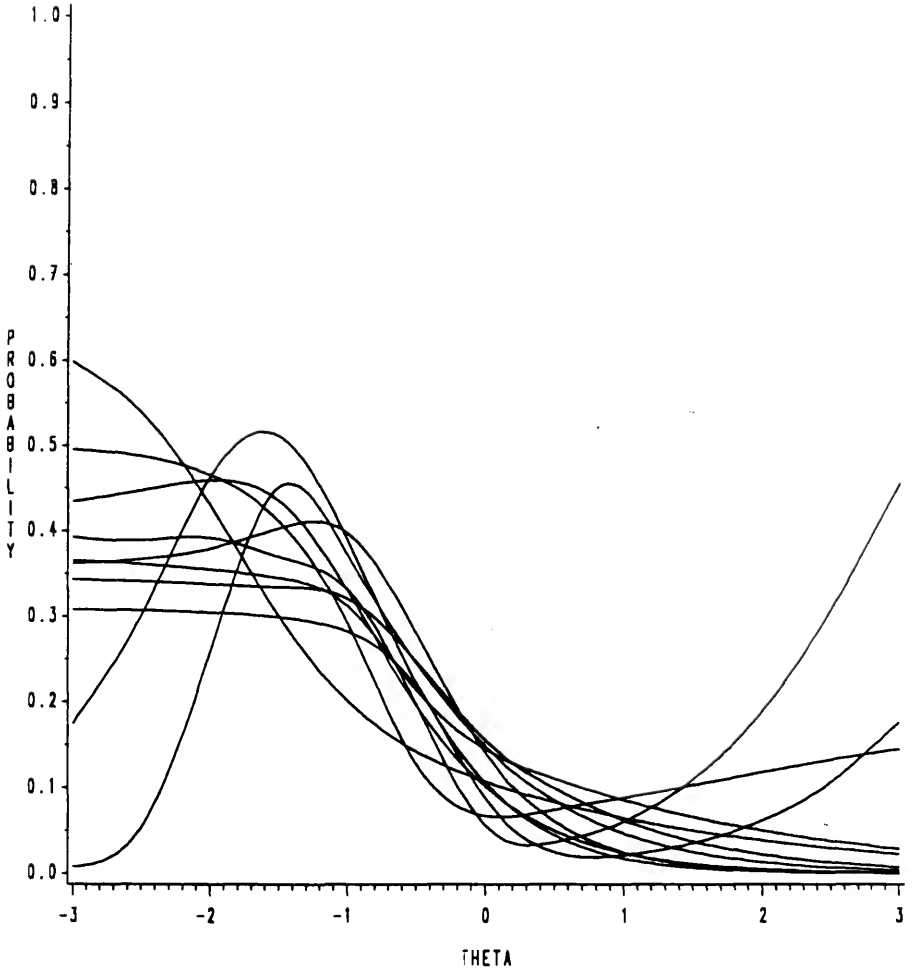

Figure 7. Item 4: Option B OCCs estimated with 500 examinees and 10 items

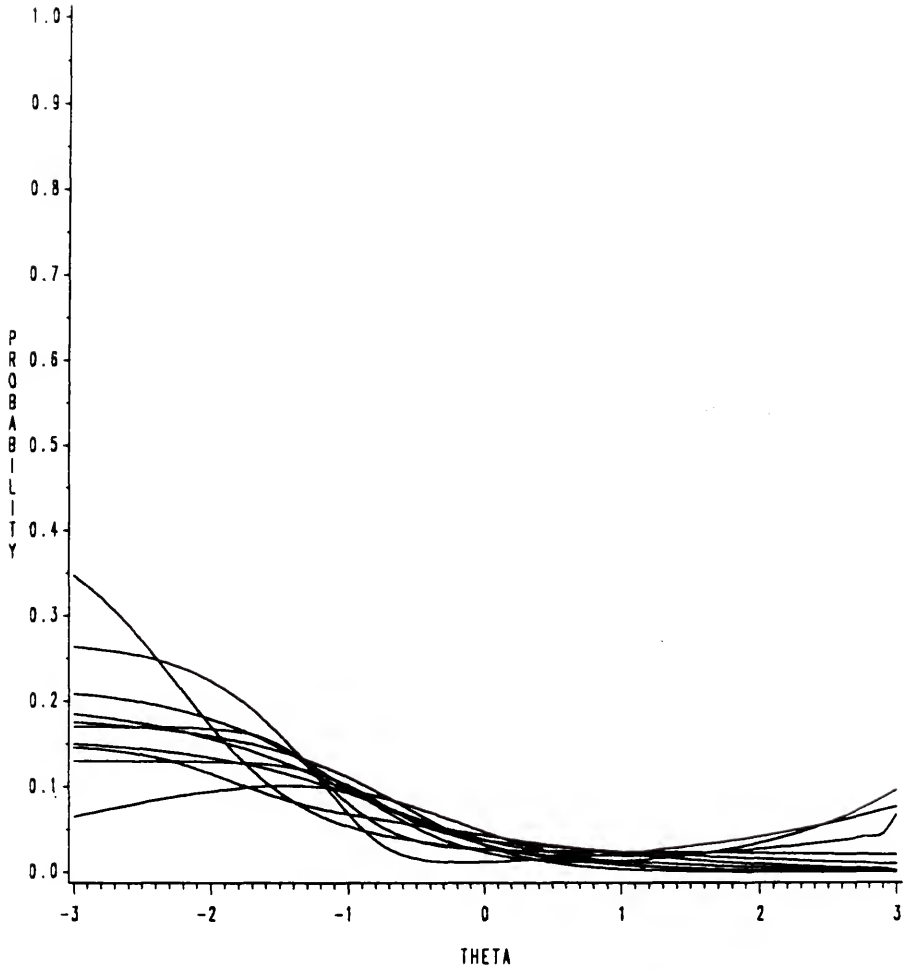

Figure 8. Item 4: Option C OCCs estimated with 1000 examinees and 10 items

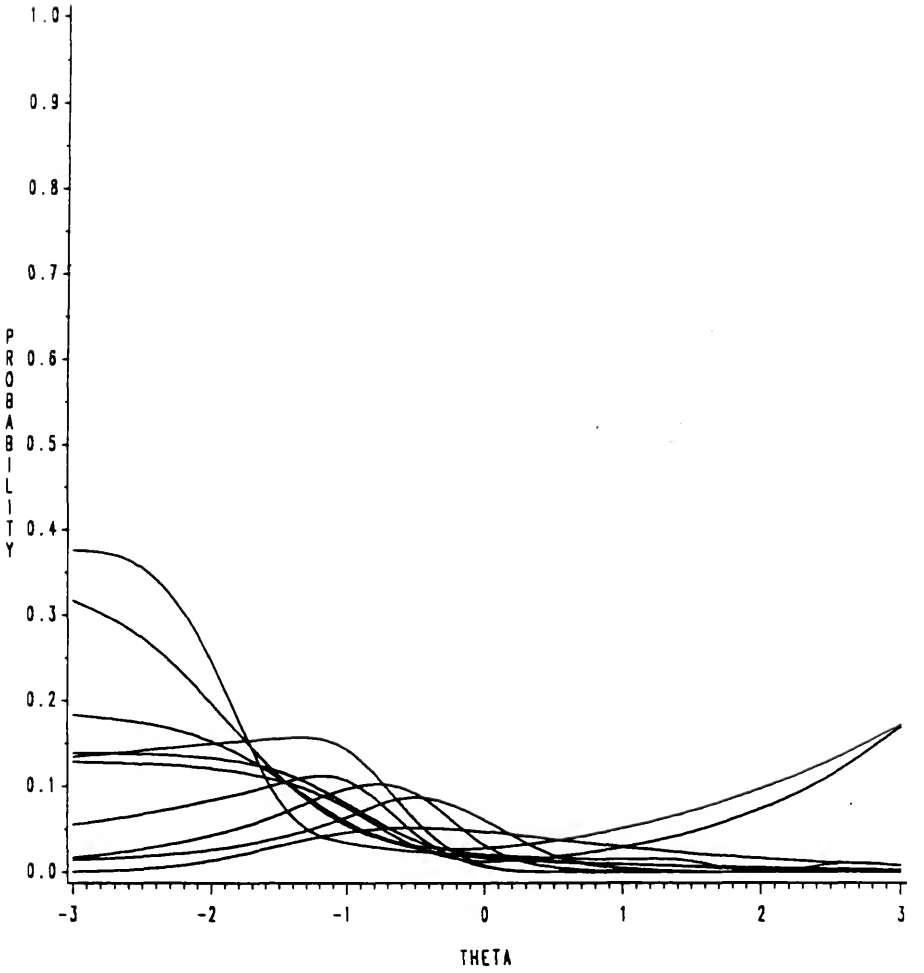

Figure 9. Item 4: Option C OCCs estimated with 500 examinees and 10 items

recommended for use with small item sets (Thissen & Steinberg, 1984), test lengths of four, six, eight, and ten items were used.

Results for the investigation of the effect of test length were not as clearcut as those for sample size. Significant differences in the average difference among OCCs due to the test length were found for three of the eight items. Table 35 contains a summary of the significant differences for number of items. Four-item subsets appeared to produce the most variable results, differing for two of the eight items from the 10-item subsets and for one of the eight items from the 8-item subset.

Table 35

Number of Significant Differences Between Common 4-Item Sets Estimated in Tests of Different Lengths

Base Test Length	Contrast Test Length		
	6 Items	8 Items	10 Items
4 Items	0	1	2
6 Items	---	0	1
8 Items	---	---	0

According to the size of the average differences, for these items, the OCCs estimated with a sample size of four items showed the most variability. This is illustrated by the OCCs for the incorrect responses to item 4, estimated with 1000 examinees and four- versus ten-item subsets. There is some instability in the lower ability range for the OCCs for option B of item 4 estimated with 1000 examinees and 10 items (see Figure 6 on page 96).

However, the OCCs estimated with 1000 examinees and four items show extreme differences in shape of the curve and lower asymptote, and also exhibit differences in the upper range of ability (Figure 10). This same effect can be seen by comparing the OCCs for option C of item 4 estimated with ten items (Figure 8 on page 98) and with four items (Figure 11) and by comparing the OCCs for option D estimated with four and ten items (Figures 12 and 13).

The variability of the OCCs in the extremes is supported by Thissen's (1986) study, in which he showed that the curves are "more similar in the middle, where the data are; 95% of the population distribution lies between θ's of -2 and +2. The different parameter estimates seem to affect the curves primarily at the extremes" (p. 95).

Thissen himself has shown an example in which the curve for an incorrect response goes to unity for low ability and explains that this occurs because in the region below $\theta = -2$ there are essentially no data and

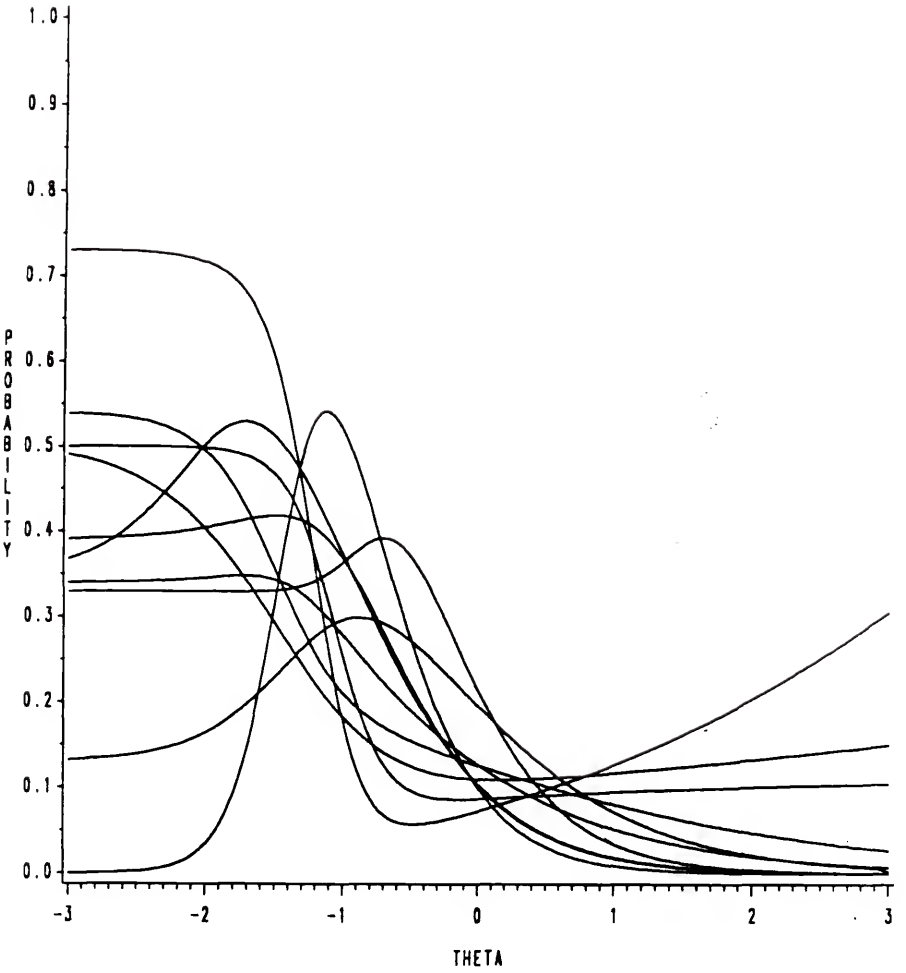

Figure 10. Item 4: Option B OCCs estimated with 1000 examinees and 4 items

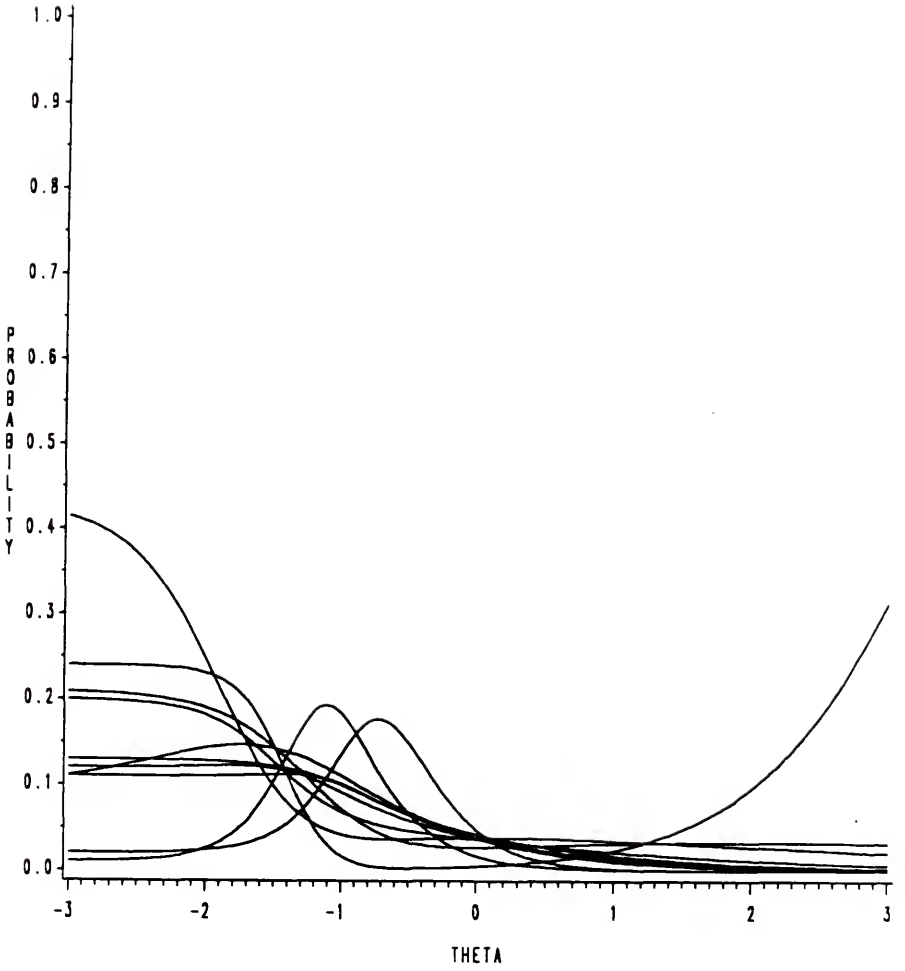

Figure 11. Item 4: Option C OCCs estimated with 1000 examinees and 4 items

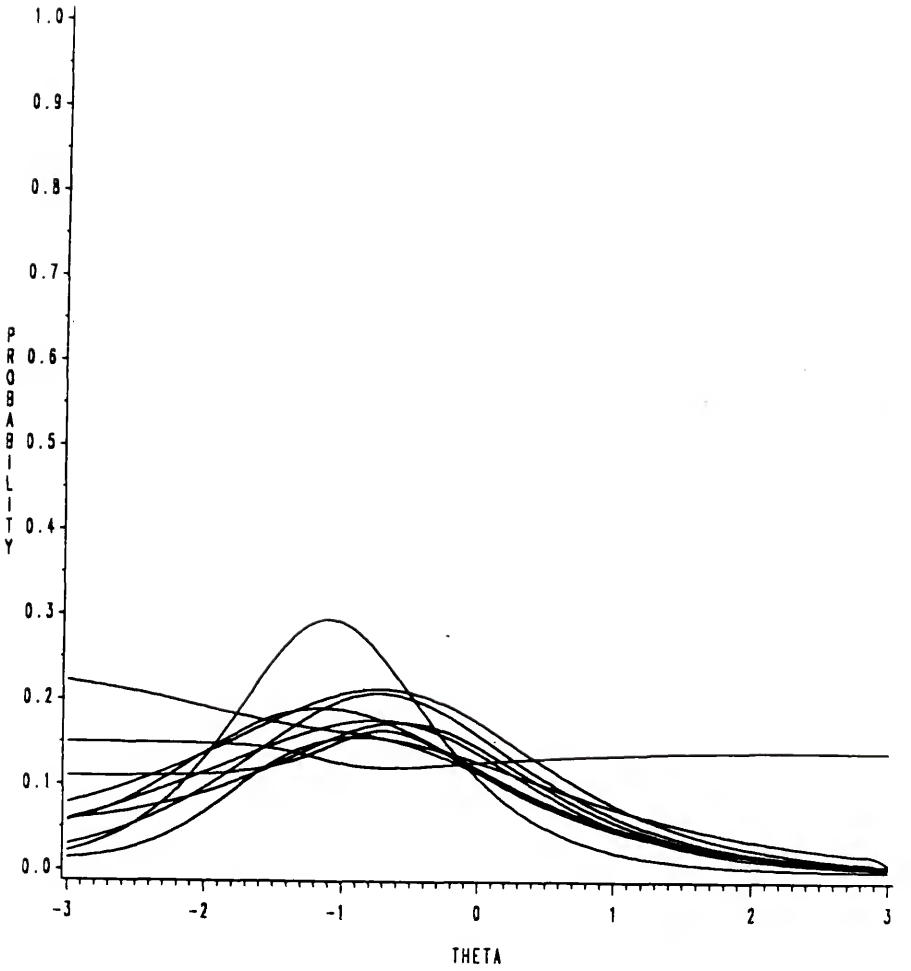

Figure 12. Item 4: Option D OCCs estimated with 1000 examinees and 10 items

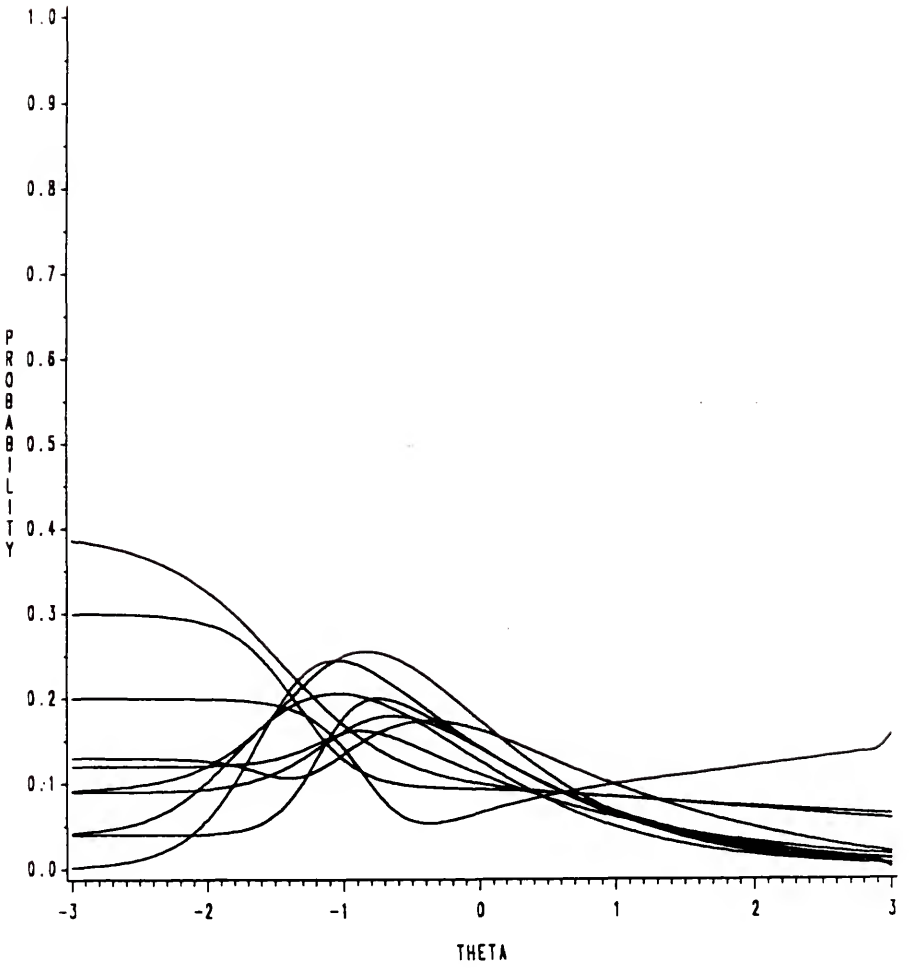

Figure 13. Item 4: Option D OCCs estimated with 1000 examinees and 4 items

extrapolation occurs. Thissen suggested three possible reasons for the differences between the curves for four items and those for the entire test of 35 items.

The first possible explanation was that 1000 examinees might not be "asymptotic" for a table with 256 cells. His second explanation was that one or the other solution may not have been completely converged, since E-M algorithms can be slow near the maximum. The third possibility, which Thissen stated was most likely, was that the model may have been incorrect in that the latent dimension θ may have been defined slightly differently in the set of four difficult items than in the entire test.

The results of the current study also indicated that the average differences for the correct options were significantly greater than those for at least one of the incorrect options for eight of the eleven items studied. However, just two of the eleven items showed significant differences between incorrect options.

These results differ from those expected at the beginning of the study. Because the correct responses are estimated with more data (i.e., a larger proportion of the examinees select that response), it was expected that OCCs for the correct response would show less variability. However, an explanation for these results can be found in the method used to measure the average difference. When an option is chosen by only a small percentage of the

examinees, the OCC produced by the estimation procedure is most likely to be a very flat low curve. Thus, the range of probabilities for these options is much narrower than that for an option with a greater range of probabilities of response across ability levels, which is typical of correct responses. Since the dependent variable used in this study measured the difference in probabilities between each OCC and the median OCC for a particular condition, differences in the dependent variable were more likely for higher incidence options.

When the incorrect options only were used, only two items showed significant differences due to option. Again, the differences were linked to the number of examinees responding to an item, with those options chosen by less than five percent of the examinees showing significantly less variability, in terms of differences across OCCs. Thus, when the differences between OCCs are measured by the difference between the median probability at a particular theta value and the probability for each OCC at that theta value, options for which the probabilities are very low show less variability.

In contrast to studies with binary models, which have shown differences in parameter estimates due to ability distribution (Ree, 1979; Swaminathan and Gifford, 1979, 1983; Wingersky and Lord, 1984; Yen, 1987), in this study

significant differences occurred due to ability distribution for just one of the three items. This difference was not in the direction expected. It was expected that the negatively skewed distributions would produce more instability in OCCs than the normally distributed samples. However, for item 4, the mean difference for the normal distribution was significantly higher than that for the skewed distribution.

Figures 14 through 17 present the OCCs for two of the incorrect options of item 4, estimated with the "easy" item set and normal versus negatively skewed ability distributions. For option A, the difference in OCCs for the normal samples occurs chiefly for abilities below -2.00 for two samples which have a discrepantly high probability at this level (Figure 15). Option B presents a different picture, with both the normal and negatively skewed samples showing large differences at the lower ability range (Figures 16 and 17). However, the normal samples appear to result in OCCs which differ more throughout the range of ability in comparison to the negatively skewed samples.

The ambiguity of the results in relation to the ability distribution may occur because the selection of the samples was based upon the distribution of raw score in the total set of 50 mathematics items. As Tables 8 and 9 illustrated, the raw score distributions for the randomly

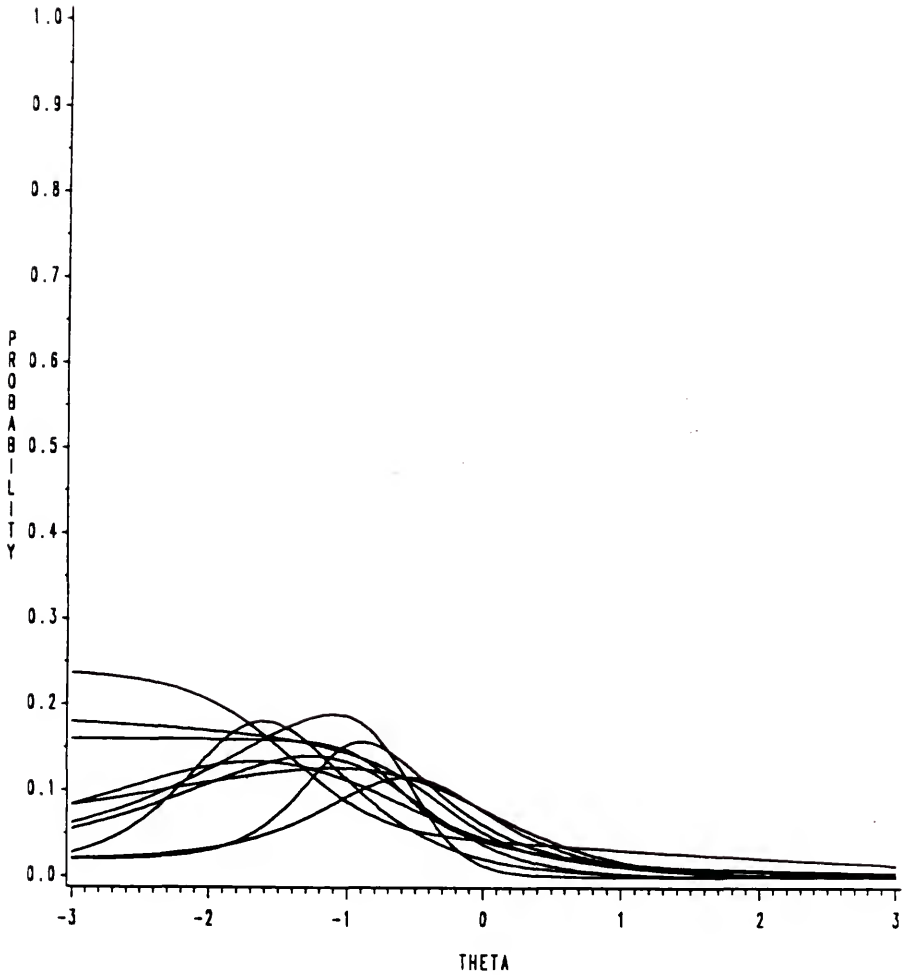

Figure 14. Item 4 in Part II: Option A OCCs estimated with negatively skewed ability and easy difficulty distributions

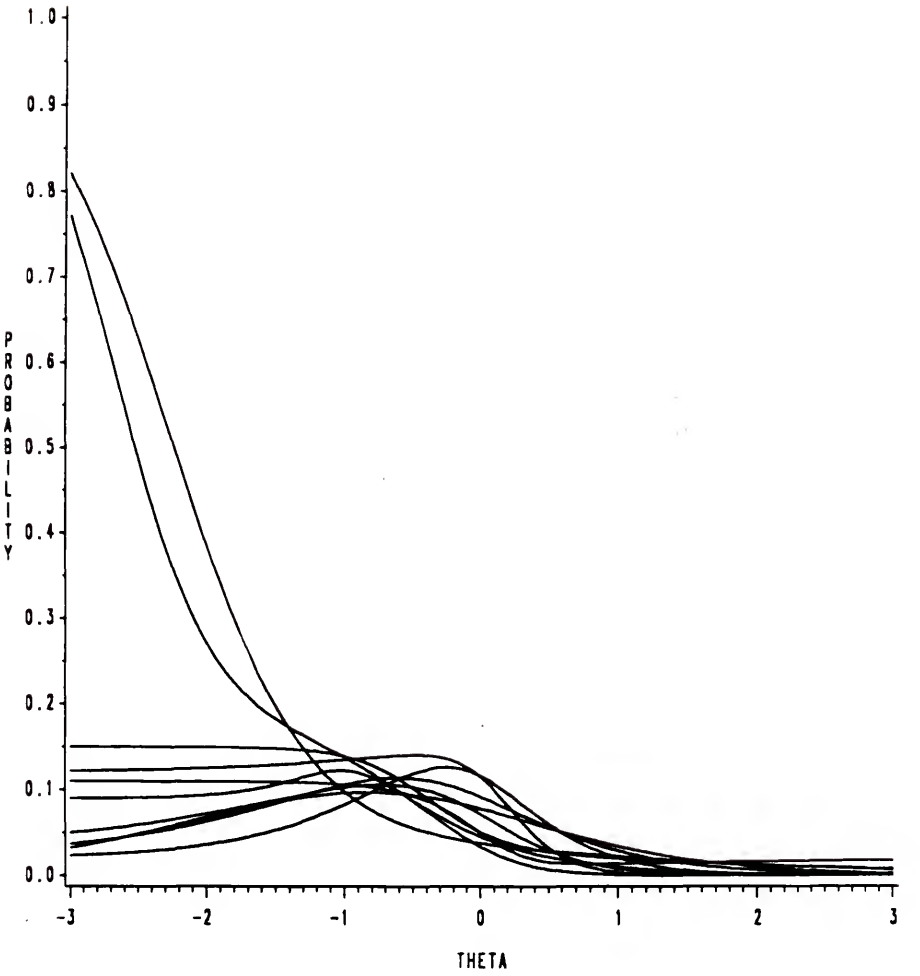

Figure 15. Item 4 in Part II: Option A OCCs estimated with normal ability and easy difficulty distributions

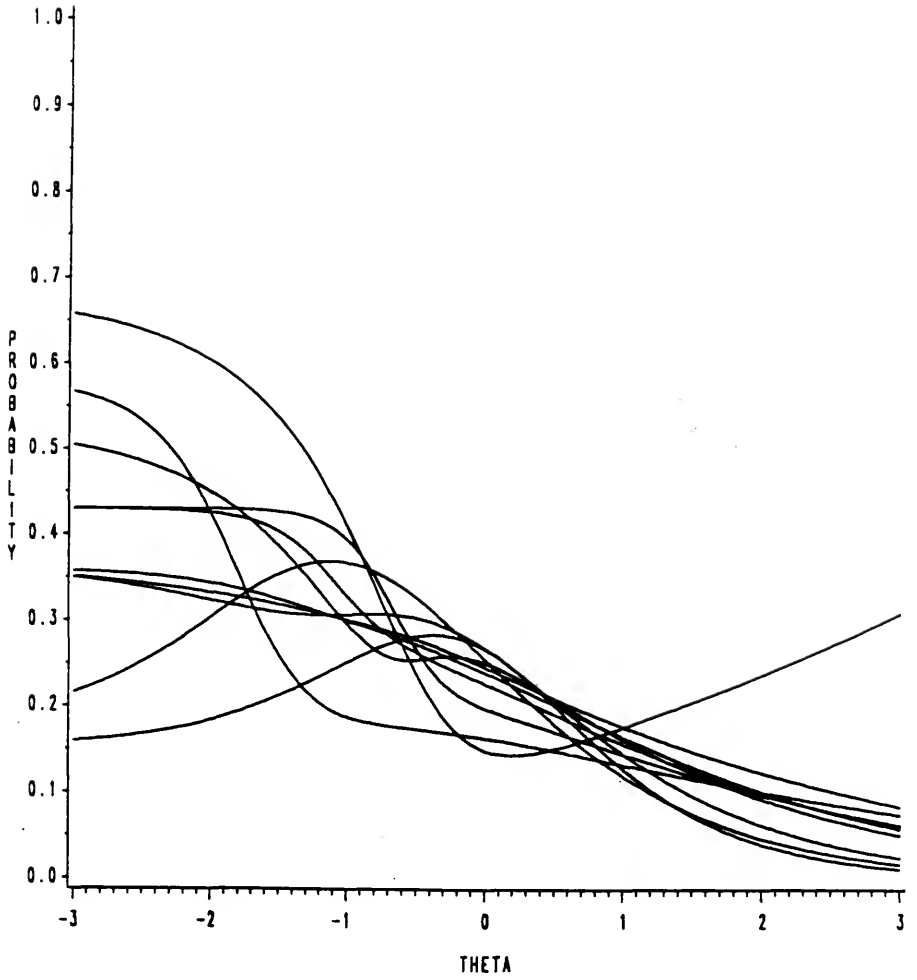

Figure 16. Item 4 in Part II: Option B OCCs estimated with negatively skewed ability and easy difficulty distributions

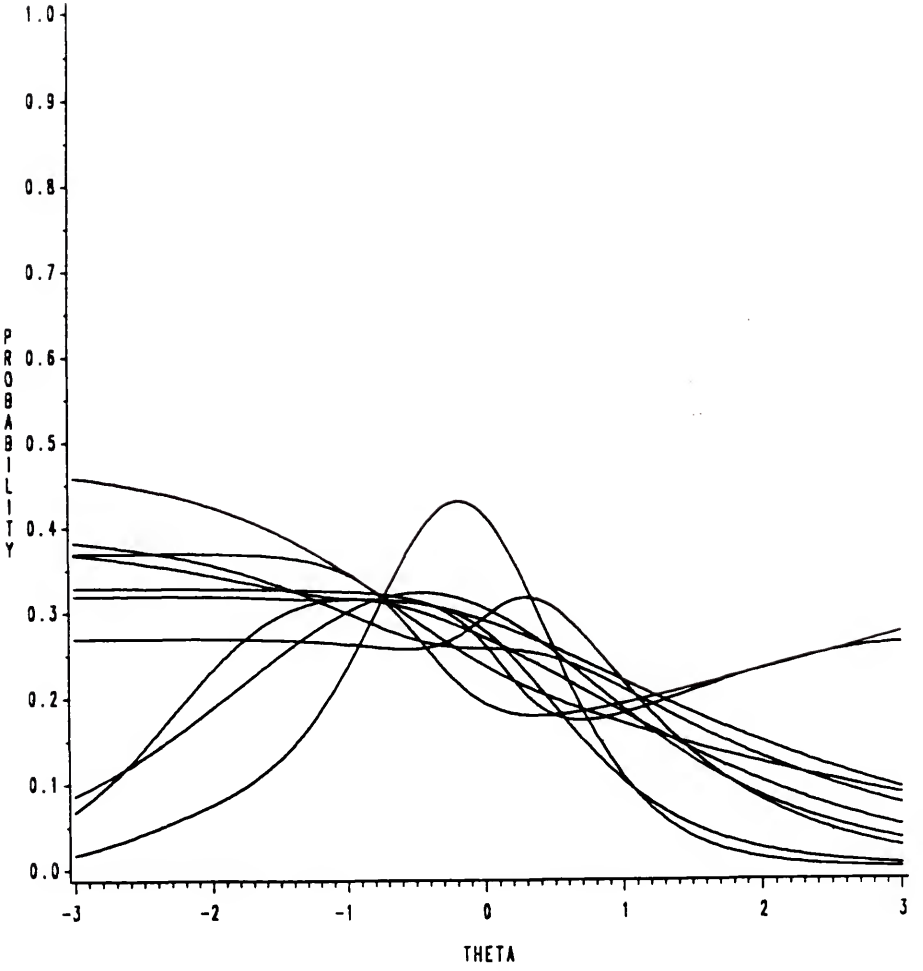

Figure 17. Item 4 in Part II: Option B OCCs estimated with normal ability and easy difficulty distributions

selected samples are negatively skewed in comparison to the normal samples. However, the distributions of ability produced by the MULTILOG estimation with the set of five items show less difference in skew for the random and normal samples, as illustrated in Table 36 by the descriptive statistics for sample 1 in each set. It is possible that larger differences in the ability distributions might result in differences in the variability of the OCCs across samples.

There was no significant effect for type of difficulty distribution (easy versus difficult items) for any of the three items. This is reassuring in that MULTILOG has been recommended for use with difficult items, but the majority of the items as the minimum competency examinations currently in use in statewide examination programs are relatively easy items. However, due to the restrictions of the test used in this study, the item sets did not differ greatly in difficulty. The study should be repeated with item sets which differ more in difficulty and with larger numbers of items.

Thissen (1986) suggested that analysis of curves for the correct response which are non-monotonic, such as that shown in Figure 18, can contribute much to item analysis. He explored reasons for the non-monotonicity on the left, such as "positive misinformation"; i.e., the correct response for the item differs from the distractors on

Table 36

Descriptive Statistics for the Ability Estimates For Sample 1

	Normal		Random (Skewed)	
	Easy	Difficult	Easy	Difficult
Mean	.0037	.0007	.0082	-.0041
Std. Dev.	.6591	.7463	.7488	.7972
Median	-.0400	-.0900	-.0200	-.0600
Mode	1.0300	1.1900	.9700	1.0100
Skewness	-.0808	.1093	-.3483	-.2030
Kurtosis	-.5585	-1.000	-.5885	-.9000

dimensions other than that which is intended and observable, given sufficient ability. Examinees of low ability respond to these "other features" and select the correct response. Another explanation suggested by Thissen for this "J-shaped" curve is cheating, so that low ability examinees show a high probability of choosing the correct response. The danger shown by this study in making interpretations such as these is that the non-monotonic trace line may be simply an artifact of estimation error.

This is illustrated by comparing the OCC for the correct response calculated for sample 9 only (Figure 18) and that for the median of the 10 samples, which is shown

in Figure 19. It is clear that a different interpretation from that presented by the "J-shaped" OCC for sample 9 would result in the majority of the cases.

Conclusions

The results of this study lead to the following conclusions:

1. The variability of the option charactistic curves produced by the estimation procedure of MULTILOG is affected by the number of examinees in the sample. The visual impression is of a marked decrease in variability as the sample size is increased from 500 to 1000 examinees. Thus, a sample size of 1000 is preferred, just as has been recommended for binary scored IRT models.

2. The variability of the option characteristic curves may be affected by the number of items in the item set. In particular, test lengths of four items do not allow as accurate estimation of OCCs as do test lengths of ten items. These results may be affected by differences in model fit among the different subsets of items.

3. No definite conclusion can be made regarding the effect of the ability distribution of the sample of examinees - normal or negatively skewed - on the variability of the OCCs, given the small differences in distributions in the study.

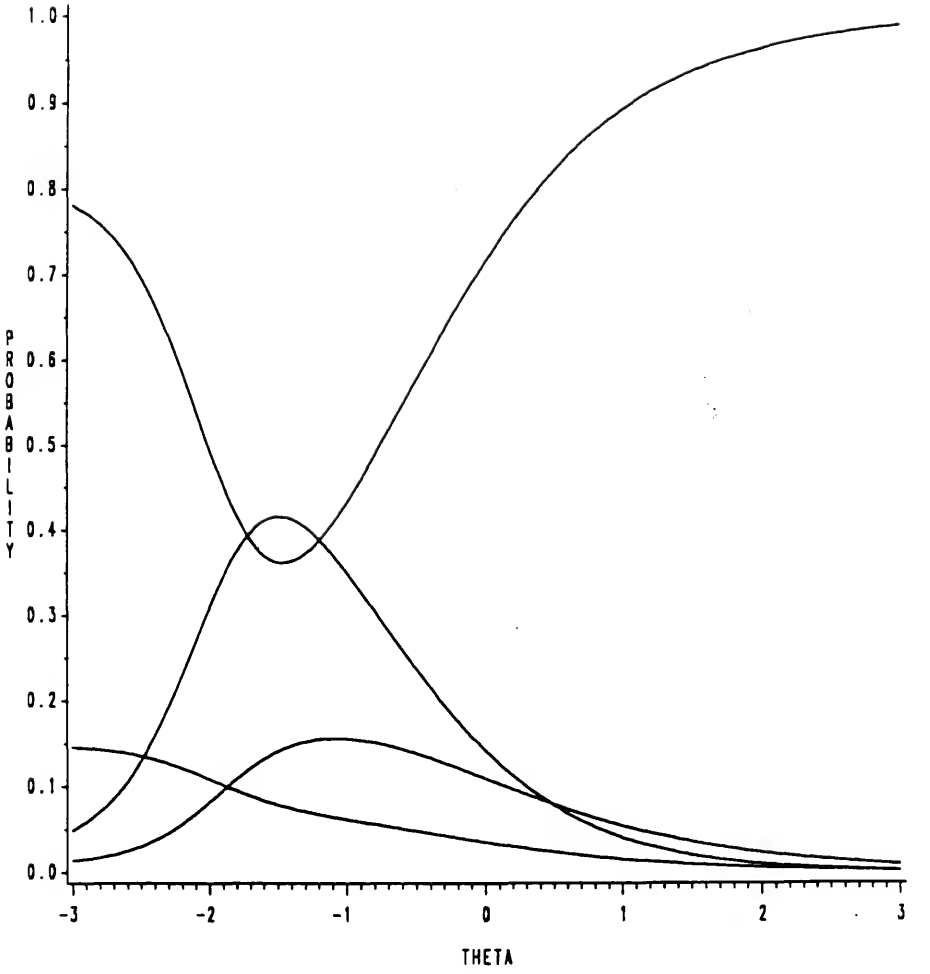

Figure 18. Sample 9 OCCs for Item 4 estimated with 1000 examinees and 10 items

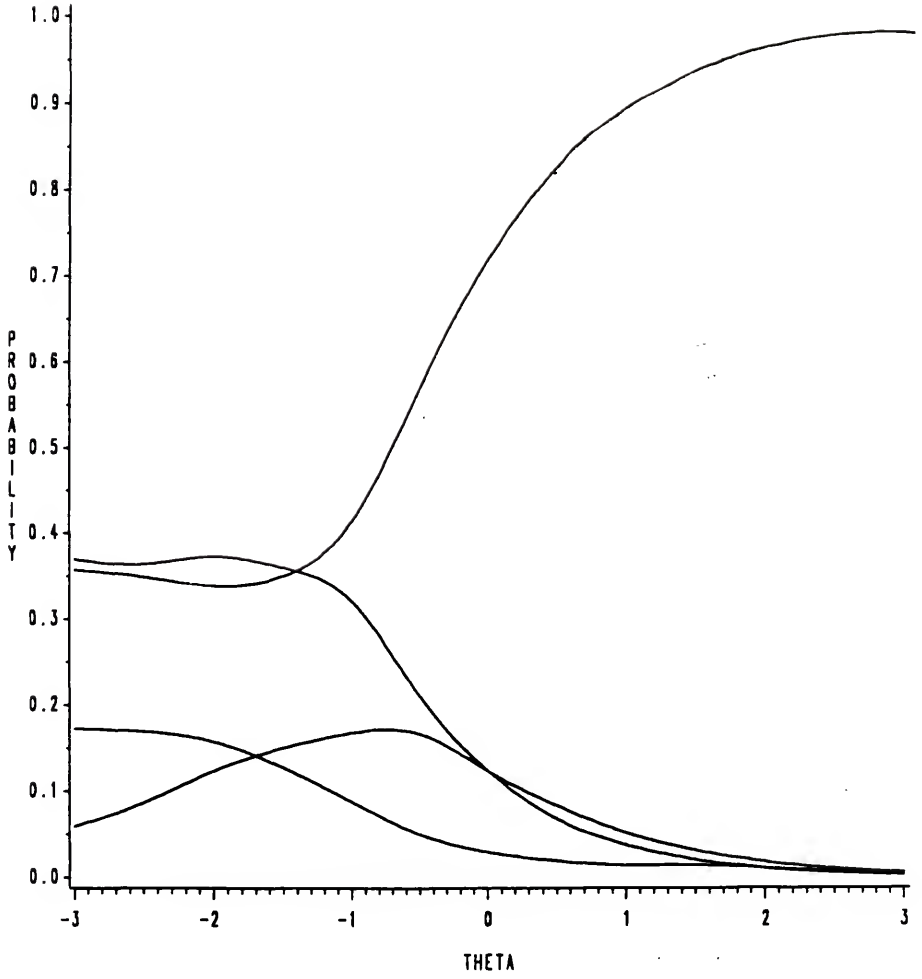

Figure 19. Median OCCs for Item 4 estimated with 1000 examinees and 10 items

4. The variability of the option characteristic curves is not affected by the difficulty distribution of the item set.

5. The variability of option characteristic curves differs over options. One factor that may contribute to this is variation in the number of examinees choosing each response.

Implications for Further Research

One of the limitations of this research was that only actual test data were used. Thus the study was limited in the type of item sets which could be studied. One recommendation would be to repeat the study using simulated data, in order to control more rigidly the parameters of the items within the sets, and to know their "true" values.

Another recommendation would be to include other sizes of samples and test lengths. For example, samples of 1000 examinees still show instability for low-ability examinees, for whom interpretations of wrong responses are most informative. A replication of the study should include samples of 2000 examinees each to investigate whether such sample sizes would stabilize the estimation at the lower end of the ability scale. Also, test lengths of more than ten items should be included.

Another recommendation would be to limit the comparison of OCCs to the ability range of -2.00 to +2.00,

rather than -3.00 to +3.00, since this is the range in which most of the data occur. The greatest variability in OCCs was observed for the area between -2.00 and -3.00, which included approximately one-third percent of the examinees for this minimum competency examination. Thus, a more reasonable measure of the differences between OCCs would be to calculate the Brown-Forsythe statistic for a smaller range of ability.

Furthermore, a procedure is needed to equate the option characteristic curves estimated on different examinee samples. Various procedures are available for equating with the binary item response models. When there are common items between two calibrations, the relationship between the item difficulties for these common items for the two subsets can be used to convert the parameters from the second subset to the scale of the first subset. Although item difficulty, the b_g parameter in the binary models, is not estimated specifically by MULTILOG, the relationship between item difficulty and the c_k and a_k parameters can be expressed as $c_k = -a_k b_g$ (Crocker, 1987). However, it does not appear that the converted b_g parameters from the multiple category scoring estimation procedure have the same relationship as those from the binary models. Thus, a procedure which allows equating across samples with MULTILOG is needed.

Finally, it is recommended that, in practical test development situations, interpretation of the option characteristic curves produced by MULTILOG proceed with caution until further determination of the stability of the option characteristic curves across samples is made.

REFERENCES

Assessment Systems Corporation. (1987). User's manual for the MicroCAT testing system. St. Paul, MN: Author.

Baker, F. B. (1977). Advances in item analysis. Review of Educational Research, 47, 151-178.

Bernhardson, C. S. (1966). Determination of the chance score on the three-decision multiple-choice test. Psychological Reports, 19, 559-562.

Bernhardson, C. S. (1967). Comparison of the three-decision and conventional multiple-choice tests. Psychological Reports, 20, 695-698.

Biomedical Computer Programs, P-Series, Health service computing facility. (1988). Los Angeles: University of California Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick. Statistical theories of mental test scores (pp. 397-479). Reading, MA.: Addison Wesley.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association, 69, 364-367.

Buhr, D. C., & Algina, J. (April, 1986). A comparison of item parameter estimates and ability parameter estimates obtained by different methods implemented by BILOG. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Cook, L., Eignor, D. R., & Petersen, N. S. (April, 1982). A study of the temporal stability of IRT item parameter estimates. Paper presented at the annual meeting of the American Educational Research Association, New York.

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. Educational and Psychological Measurement, 16, 13-37.

Crocker, L. (1987). Estimating multiple choice item parameters using information in wrong responses. Final report to the Institute for Student Assessment and Evaluation and the Florida State Department of Education, Grant No. 291-9095-86003.

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

Davis, F. B., & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 19, 159-170.

Donlon, T. F. (1984). Distractor analysis as evidence of test fairness in the Scholastic Aptitude Test. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Dressel, P. L., & Schmid, P. (1953). Some modifications of the multiple-choice item. Educational and Psychological Measurement, 13, 574-595.

Echternacht, G. J. (1972). The use of confidence testing in objective tests. Review of Educational Research, 42, 217-236.

Frary, R. B., & Giles, M. B. (1980). Multiple-choice test bias due to answering strategy variation. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Gullikson, H. (1950). Theory of mental tests. New York: John Wiley.

Guttman, L. (1941). The quantification of class attributes: A theory and method of scale construction. In P. Horst (Ed.). The prediction of personal adjustment. New York: Social Science Research Council.

Haberman, S. (1975). Maximum likelihood estimates in exponential response models. The Annals of Statistics, 5, 814-841.

Hambleton, R. K., & Cook, L. (1977). Latent trait models and their use in the analysis of educational data. Journal of Educational Measurement, 14, 76-96.

Hambleton, R. K., & Traub, R. E. (1971). Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hambleton, R. K., & Swaminathan, H. (1983). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing Co.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. Journal of Educational Measurement, 7, 75-82.

Hendrickson, G. F. (1971). The effect of differential option weighting on multiple choice objective tests. Journal of Educational Measurement, 8, 291-296.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. (1970). Validity and reliability consequences of confidence weighting. Paper presented at the meeting of the American Educational Research Association, Minneapolis, Minnesota.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item Response Theory: Application to Psychological Measurement. Homewood, IL: Dow Jones-Irwin.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement, 6, 249-260.

Huynh, T., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics, 1, 69-82.

Jacobs, P. L., & Vandeventer, M. (1970). Information in wrong responses. Psychological Reports, 26, 311-315.

Jensen, A. R. (1976). Test bias and construct validity. Phi Delta Kappan, 58, 340-346.

Kirk, R. Experimental design: Procedures for the behavioral sciences. (1968). New York: Wadsworth Publishing Co.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 43, 675-685.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F. M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. (Research Bulletin 73-33). Princeton, NJ: Educational Testing Service.

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

Michael, J. C. (1968). The reliability of a multiple-choice examination under various test-taking instructions. Journal of Educational Measurement, 5, 307-314.

Mislevy, R. J., & Bock, R. D. (1984). BILOG version 2.2: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software, Inc.

Nedelsky, L. (1954). Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 14, 459-472.

Olejnik, S.F., & Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. Journal of Educational Statistics, 12, 45-61.

Patnaik, D. & Traub, R. E. (1973). Differential weighting by judged degree of correctness. Journal of Educational Measurement, 10, 281-286.

Qualls, A. L., & Ansley, T. N. (1985). A comparison of item and ability parameter estimates derived from LOGIST and BILOG. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Raven, J. C. (1956). Guide to using the Coloured Progressive Matrices, Sets A, Ab, and B. London: Lewis.

Ree, J. M. (1979). Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-385.

Ree, M. J., & Jensen, H. E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D.J. Weiss (Ed.), Proceedings of the 1979 computerized adaptive testing conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Reilly, R. R., & Jackson, R. (1973). Effects of empirical option weighting on reliability and validity of an academic aptitude test. Journal of Educational Measurement, 10, 185-194.

Rippey, R. (1968). Probabilistic testing. Journal of Educational Measurement, 5, 211-215.

Sabers, D. L., & White, G. W. (1969). The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. Journal of Educational Measurement, 6, 93-96.

Samejima, F. (1979). A new family of models for the multiple choice item. (Research Report No. 79-4). Knoxville: University of Tennessee, Department of Psychology.

Schueneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed.), Handbook of methods for detecting item bias (pp. 180-198). Baltimore: Johns Hopkins University Press.

Shuford, E. H., Jr., Albert, A., & Massingill, H. E. (1966). Admissable probability measurement procedures. Psychometrika, 31, 125-145.

Swaminathan, H. (1985). Parameter estimation in item response models. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Swaminathan, H., & Gifford, J. A. (1979). Estimation of parameters in the three-parameter latent trait model (Report No. 90). Amherst, MA: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluation Research.

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

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. Psychometrika, 50, 349-364.

Swaminathan, H., & Gifford, J. A. (1986). A comparison of the joint and marginal maximum likelihood procedures for the estimation of parameters in item response models. Unpublished manuscript, University of Florida, The Institute for Student Assessment and Evaluation, Gainesville, FL.

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 13, 201-214.

Thissen, D. M. (1984). MULTILOG: A user's guide. Mooresville, IN: Scientific Software.

Thissen, D. M., & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 49, 501-519.

Urry, V. (1977). OGIVIA: Item parameter estimation program with normal ogive and logistic three-parameter model options. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Urry, V. (1978). ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Veale, J. R., & Foreman, D. I. (1983). Assessing cultural bias using foil response data: Cultural variation. Journal of Educational Measurement, 20, 249-258.

Wang, M. W., & Stanley, J. C. (1968). Differential weighting: A review of methods and empirical studies. Review of Educational Research, 40, 663-705.

Wendler, C. L. W., & Carleton, S. T. (1987). An examination of SAT verbal items for differential performance by women and men: an exploratory study. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Willey, C. F. (1960). The three-decision multiple-choice test: A method of increasing the sensitivity of the multiple-choice item. Psychological Reports, 7, 475-477.

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

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

Wright, B. D., Mead, R., & Bell, S. (1979). BICAL: Calibrating items with the Rasch model. (Research Memorandum 23b). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

BIOGRAPHICAL SKETCH

Dianne C. Buhr was born in Shelby, Iowa, on December 9, 1943. She attended the Shelby Community School for twelve years, graduating in 1961, as valedictorian.

In 1964, Dianne received a B.S. degree in English, with minors in education and psychology from Iowa State University. For the next two years, she and her husband, Ken, served as Peace Corps volunteers in a rural education project in Brazil.

In 1967, they returned to Ames, Iowa, where Dianne taught junior and senior high school English for three years.

Between 1970 and 1980, Dianne and her husband lived for several years in South America, where Dianne taught in a Panamerican middle school.

In 1981, Dianne received her master's degree in educational psychology from the University of Florida. In 1984, she began working for the Office of Instructional Resources at the University of Florida as Assistant Director of Testing and Evaluation.

Dianne and Ken, who is with the Department of Agronomy at the University of Florida, have two sons, Aaron and Joshua.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda Crocker

Linda Crocker, Chair
Professor of Foundations
of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James Algina

James Algina
Professor of Foundations
of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

William Hedges

William Hedges
Professor of Educational
Leadership

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

M. David Miller
Assistant Professor of Foundations of Education

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1989

Chairman, Foundations of Education

Dean, College of Education

Dean, Graduate School