To the Graduate Council:

We are submitting herewith a dissertation written by Megumi Asako entitled "An Application of Latent Trait Theory in Analyzing the Field Test Results of a Mathematics Proficiency Test." We have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosphy, with a major in Education.

Schuyler W. Huck

Schuyler W. Huck, Major Professor

Fumiko Samejima

Fumiko Samejima, Major Professor

We have read this dissertation and recommend its acceptence:

John R. Ray

Thomas C. Innes

Accepted for the Council:

CW Minkel

The Graduate School

AN APPLICATION OF LATENT TRAIT THEORY IN ANALYZING THE FIELD TEST RESULTS OF A MATHEMATICS PROFICIENCY TEST

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Megumi Asako

August 1984

DEDICATION

This study is dedicated to the author's husband, Yoji Takata. He has provided unselfish support, willingness to learn together, and the pleasant environment to continue this study under any circumstances.

ACKNOLEDGEMENT

The deepest gratitude is extended to Dr. Schuyler Huck for his guidance throughout the course of the study. He has always treated the author with warm care and respect. Also, the author is very grateful to Dr. Fumiko Samejima for giving instructions, letting the author use her computer programs, and granting the author to reproduce the figure from her work, *Final report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item*.

The author is greatly indebted to Dr. Thomas Innes for his constructive criticisms and willingness to assist the author to grow as a professional and also for helping her obtain the data for this study. Comments by Dr. John Ray is also very helpful.

It has been extremely fortunate for the author to have Dr. Sheldon Clark as her friend. He has provided innumerable invaluable information and enlightening comments, and has treated her in a warm manner whenever she lost her way since she entered the University of Tennessee. The author wishes to express her heartfelt appreciation to Dr. Peccolo and the other people in the Bureau of Educational Research and Service for kindly providing her with the environment to accomplish her study.

Encouragements and hearty supports given by Mary K. Cox and Michael Searcy since the author came to the United States are most gratifying. Allen Knight's friendship and valuable help to the author are very much appreciated.

The author's mother and the foster mother's insight about the author and their attitudes toward the life has motivated the author to continue this study. The author also makes a cordial acknowledgment to the University of Tennessee Computing Center for the use of the facilities in every stage and phase of this study.

ABSTRACT

The application of one of the latent trait models, the normal ogive model to a large scale test is not numerous. The purpose of present study is to find out information provided by the normal ogive model, using the results of the multiple-choice items of the field test of the Tennessee Proficiency test which have four alternatives. The results of 933 sixth- and seventh-grade students were used in this study. The principal factor solution using tetrachoric correlation was employed to ascertain the dimensions. Item characteristic function and information function following the normal ogive model were produced. Maximum likelihood estimates of abilities of 933 students were computed. The distractor analysis (plausibility function) was employed. Information provided by different types of objectives and test items were discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

(Continued)

FIGURE PAGE

(Continued)

FIGURE PAGE

(Continued)

INTRODUCTION

The popularity of latent trait theory is increasing year after year. However, the application of the normal ogive model (one of the latent trait models) to a large scale test has not been numerous. The purpose of the present study is to obtain information, using the latent trait theory, which is not available by traditional analysis. The additional information includes that obtained by the plausibility function (distractor analyses) which is an analysis of each distractor in the multiple-choice test items in addition to those of the correct answer.

Chapter I contains: backgrounds of latent trait theory, the purpose of the study, the research questions, and significance and justification of this study. In Chapter II, attractive concepts in latent trait theory applied in this study are reviewed and discussed along with comparisons with traditional test theory. Chapter III explains the procedures employed in this study and the reasons why selected methods were employed. The procedures include: (a) description of the field test; (b) selection of subjects; (c) production of frequency distributions of alternatives; (d) computation of tetrachoric correlation coefficients; (e) finding the dimensionality; (f) computation of item parameters;

(g) production of item characteristic functions; (h) computation of maximum likelihood estimates of ability and production of plausibility functions; (i) production of information functions; and (j) conventional methods of analysis. Chapter IV consists of descriptions of findings. Chapter V discusses the findings, and conclusions are presented. Also mentioned is possible further research.

CHAPTER I

STATEMENT OF PROBLEM

Backgrounds

Modern test theory (latent trait theory) was initiated by Lawley in 1943 and was further developed by Lazarsfeld in 1950 on the social attitude test (Lord, 1968). Lord (1952) refined this theory for aptitude and achievement tests as a mental test theory. Since then this mental test theory has been refined by many other psychometricians. Even though many advantages of latent trait theory have been pointed out (Hambleton and Cook, 1977; Lord and Novick, 1968; Samejima, 1977b), applications of this theory to large scale testing have not been numerous.

Purpose of the Study

The purpose of this study is to apply latent trait theory in analyzing a field test which measures mathematics proficiency. It is of primary interest to investigate the information obtainable from distractors as well as correct answers of these multiple-choice items, and to find out in what ways latent trait theory findings differ from conventional findings. In so doing, the frequency distributions of the alternatives and the estimated plausibility functions as suggested by Samejima (1981) will be examined.

Representative Research Questions

The following questions represent the foci of the research:

1. Do items meet the assumption of unidimensionality?
2. Is the normal ogive model suitable for the field test of Tennessee Proficiency Test?
3. Do plausibility functions provide information useful in ability estimation?
4. In what ways, do findings by latent trait theory differ from findings by traditional test theory?
5. What guiding principles may be found for constructing test items amenable to latent trait theory interpretation?

Significance and Justification of the Study

Among public school personnel and many scholars, latent trait theory is still considered too complex for practical use. While some published tests now utilize latent trait theory, most of them have been developed under traditional methodology. One reason for the limited use of latent trait methodology may be that it is difficult and cumbersome to write an item to fit some chosen models. It is hoped that this study will make some directions for constructing test items which the normal ogive model fits. Many researchers are interested in this problem. Interpretation of the information obtained from distractor analysis is focused in the present study. It is also hoped that this study will contribute to a better understanding of the way latent trait theory gives information compared to traditional test theory.

CHAPTER II

REVIEW OF LATENT TRAIT THEORY

In this chapter, new concepts in modern test theory and terminology relating to latent trait theory are reviewed and discussed. First, attractive concepts in modern test theory are discussed. Second, conmparisons of the new concept of ability in latent trait theory with the test score in traditional test theory are made. The method of obtaining the estimate of ability is mentioned. Third, assumptions using latent trait theory--dimensionality and local independence--are reviewed. A common technique to ascertain the dimensionality (principal factor solution) is mentioned. Fourth, differences between item parameters in latent trait theory and those in traditional test theory are discussed. Fifth, the item characteristic function and the plausibility function are reviewed. Sixth, available latent trait models are explained. Seventh, descriptions about the information function are given. Standard errors of estimation in traditional test theory and those in latent trait theory derived from the information function are compared. Finally, one of applications of latent trait theory--adaptive testing--is explained.

Notation

The following notations are used in this study:

a	person
T	true score
E	expectation
X	observed score of population
g	item (= 1. 2, 3. . . . , n)
X_g	observed score of gth item
$P(X_g)$	proportion of getting X_g
T_a	true score of Person a
X_a	observed score of Person a
$E(X_a)$	expectation of X_a
$E(X)$	expectation of X
x_g	graded item score of gth item (= 0, 1, . . ., m_g)
$P(x_g)$	proportion of getting x_g
θ	ability or latent trait theta
K	number of common factors
V	response pattern
p_g	proportion correct of gth item; for convenience p is sometimes used to denote p_g
a_g	discrimination parameter of gth item
b_g	difficulty parameter of gth item
$P_g(\theta)$	item characteristic function
u_g	binary item score (which assumes either 0 or 1, correct or incorrect)

D	scaling factor: 1.7
c_g	guessing parameter of gth item
$P^*_g(\theta)$	three parameter model
$P_{x_g}(\theta)$	operating characteristic of the graded score, or conditional probability of obtaining the score x_g, given θ
$P_V(\theta)$	response pattern operating characteristic (conditional probability of the occurence of a specific response pattern. given θ)
$I_{x_g}(\theta)$	item response information function
$A_{x_g}(\theta)$	the basic function which is the ratio of the first derivative of $P_{x_g}(\theta)$ with respect of θ to $P_{x_g}(\theta)$
$I_g(\theta)$	item information function
$P'_g(\theta)$	the first derivative of $P_g(\theta)$ with respect to θ
$I_V(\theta)$	response pattern information function
σ_e	standard error of measurement
σ_X	standard deviation of the observed test score X
$\rho_{XX'}$	reliability coefficient of the test
X'	parallel test of X
ρ_{XT}	test reliability index
$I(\theta)$	test information function
ρ_g	factor loading of the first principal factor, or θ for item g

r_{bis} biserial correlation coefficient

r_{pbis} point biserial correlation coefficient

y y ordinate

$\phi(u)$ standard normal distribution function

Attractive Features of Modern Test Theory

In traditional test theory, a set of test items, or a total test score, is the smallest unit used to represent the examinee's ability. Item analysis is a smaller unit but describes group ability. In contrast, latent trait theory postulates that each item (a question) in a test can provide discreet information leading to an estimate of the examinee's ability. Therefore, each item serves as an important source of information. In traditional test theory, one item usually has as much weight as any other in determining the subject's ability. Latent trait methodology yields weights which, when applied, could lead to different scores by subjects having identical raw scores. In latent trait theory, latent traits (abilities) which cannot be observed directly are estimated from the response pattern described by Samejima (1969, p. 1):

> . . . with n dichotomous items there are 2^n possible response patterns, which are represented by such vectors as (0, 1, 0, 0, 1 . . .), while these response patterns are classified into (n + 1) score categories, if we use the number of right answers as the test score.

Ability estimation, the concepts of the item parameters (item difficulty and item discrimination parameters), and error of estimation in latent trait are also considered to be attractive features of modern test theory. They will be discussed later in this cchapter.

True Score and Ability

In traditional test theory, an observed score is the starting point for obtaining a true score estimate. The true test score in traditional test theory is defined (Lord and Novick, 1968):

$$T_a = E(X_a), \text{ and } e_a = (X_a - T_a) \ . \tag{2.1}$$

True score is the expectation of the observed test score. Therefore, the true test score varies depending on the test. For the population, in the case of discrete random variable,

$$E(X) = \sum_{g=1}^{n} X_g \, P(X_g) \ . \tag{2.2}$$

In latent trait theory, the ability of the examinee should not vary even when different sets of items are used to estimate that ability. Once an initial estimate of ability is made, an estimate also can be made of the probability of a correct response to each item at the fixed ability level. Therefore, individual items can contribute more information in latent trait methodology than in traditional test theory methodology. In modern test theory, the concept of ability is introduced, and this is not affected by specific sets of items. Some method is required in estimating abilities. Two frequently used methods are discussed in the next section.

Maximum Likelihood and Bayesian Estimates of Ability

The maximum likelihood estimate and the Bayesian estimate are commonly used in estimating abilities. The biggest difference between the Bayesian estimate of ability and maximum likelihood estimate is the Bayesian's requirement of prior distribution of θ. The categorization of the prior distribution can be based on many different characteristics, such as race, sex or age groups. Therefore, the selected distribution makes a difference. The categorization may not have meaningful information about the subject's ability. A different prior distribution could be used in the estimation of ability for the individual.

The computation of the maximum likelihood estimate of ability requires only the response pattern of each subject. However, the Bayesian model requires prior distribution of θ in addition to the response pattern. Since the Bayesian estimate has additional information, it appears that it provides more accurate information regarding the subject's ability. According to Samejima (personal communication, 1983), however, the prior distribution of θ has nothing to do with the individual's performance. This additional information, therefore, gives unnecessary biases to ability estimates. The Bayesian estimation penalizes high ability examinees and gives unqualified advantages to low ability examinees within a single population. In the Bayesian estimation, the estimated abilities of the two examinees who have exactly the same ability but belong to different groups (with different prior distribution) will vary depending on the prior distribution.

Dimensionality and Local Independence

Dimensionality. It is very important to know the dimensionality of a set of items before any analysis is applied in latent trait theory. It is often the case that a test measures more than one dimension, such as both reading ability and mathematics ability. However, Lord and Novick (1968, p. 359) stated that latent trait theory is most useful when the dimensionality of latent space is small. Moreover, they stated:

> . . . in the complete latent space, the conditional distribution of item score for fixed θ is the same for all the population of interest. . . .
>
> . . . the nature and dimensionality of the complete latent space have been defined in such a way that they depend both on the particular traits and on the particular population of interest to the psychometrician. (pp.359-360)

In many cases, latent trait theory is used with unidimensional latent traits. In other words, a set of test items measures a single ability such as only reading ability or only mathematics ability, but not both at the same time. When a set of test items is not homogeneous, and measures more than a single ability, it is measuring multidimensional latent traits. In both the unidimensional and the multidimensional cases, factor analysis can be used in determining the dimensionality.

Phi-coefficient and tetrachoric correlation. For the use of factor analysis, phi-coefficient and tetrachoric correlation are the candidates for this study which used dichotomous test items. Tetrahoric correlation was used because phi-coefficient uses the item score which is considered to change depending on the subjects, while tetrachoric correlation uses the response tendency (hypothetical concept). It is hoped that the response tendency will be invariant over the groups of examinees (Lord and Novick, p. 348). The computation of tetrachoric correlation assumes an underlying bivariate normal distribution for the joint distribution of the two variables.

Factor analysis. Since one of the assumptions in the application of unidimensional latent trait theory is that a set of items measures a single latent trait, it is important to determine the dimensionality before any unidimensional latent trait model is applied. Factor analysis is useful at this stage. If factor analysis shows that the first common factor is very dominating among other factors, this indicates that this set of tests contains items which are more likely to measure one latent trait. However, if the analysis shows that the contributions of the first several factors are substantial, this means that the latent space is multidimensional. Factor analysis is also used to determine the cluster of items in the multidimensional latent space. After the discovery of clusters of items, each cluster can be analyzed separately using the unidimensional latent trait model, because each cluster usually has unidimensionality. Part of results from principal factor solution are also used to compute item parameters.

There are some major differences between the principal factor solution of the factor analysis and the principal component analysis. In principal component analysis each of n observed variables is explained as a linear combination of n variables, or n principal components. On the other hand, the principal factor solution is described as K common factors of which the number is reduced and smaller than n, and a unique factor. In unidimensional latent trait theory, only one specific ability (e.g. mathematical skill) is desired from a set of specific test items. This specific ability should emerge as a common factor.

Local independence. To satisfy local independence, no two items of a set should be correlated except through the traits of interest. Lord and Novick (1968, p. 361) stated:

> Local independence means that within any group of examinees all characterized by the same values, the conditional distributions of the item scores are all independent of each other. This in no way suggests that item scores are unrelated to each other for the total group of examinees. What it means is that item scores are related to each other only through the latent trait variables θ_1, $\theta_2, \ldots, \theta_m$.

Local independence is defined mathematically in such a way that the probability of success on all items is equal to the product of the separate probabilities of successes. Thus the following formulas are given by Lord and Novick (1968, p. 361):

$$P_{x_g}(\theta) = P(x_g | \theta) \text{ , and} \tag{2.3}$$

$$P_V(\theta) = \prod_{x \in g} P_{x_g}(\theta) \text{ .} \tag{2.4}$$

$P_V(\theta)$ plays the role of the likelihood function in the estimation of examinee's ability using the maximum likelihood estimation.

Item Difficulty and Item Discrimination Parameters

When the assumption of local independence is met and the dimensionality is determined, an appropriate latent trait model can be applied. Following the selection of the model, item dificulty and item discrimination parameters are determined. The concepts of item parameters in traditional test theory and latent trait theory differ. According to Thorndike (1971, p. 139), in traditional test theory, the simplest and the most common measure of item difficulty, p_g, is the proportion of correct answers on the item. It may be stated as:

$$p_g = \frac{\text{Number of subjects who have correct answer on the gth item}}{\text{Total number of subjects}} . \qquad (2.5)$$

In traditional test theory, item discrimination is found by one of the following correlations:

1. Point biserial correlation coefficient between a dichotomous variable and a nondichotomous variable, which is a special case of the product-moment correlation coefficient;
2. Biserial correlation coefficient between the dichotomous variable and the continuous variable: This coefficient is obtained by hypothesizing the

existence of a continuous latent variable imposed in scoring a dichotomous item (Lord and Novick, 1968);

3. Rank biserial correlation coefficient between one variable with untied ranks and a dichotomous item (Glass, 1966).

In latent trait theory. the concept of item difficulty parameter is different from that in traditional test theory. Lord (1980, p. 12) says "Item difficulty parameter is a location parameter." It determines the position of the curve along the ability scale.

In Figure II.1, the item characteristic curve indicates A is an easier item compared to the item of B, and Item B is easier than Item C.

Discrimination parameter is defined as being proportional to the slope of the curve at the inflexion point (Lord and Novick, 1968, p. 366). Furthermore, they stated:

> The normal ogive item characteristic curve has a point of inflection at $\theta = b_g$; at this point, the probability of a correct answer is 0.5, and the slope of the curve is $a_g/\sqrt{2\pi}$ (Lord and Novick, 1968, p. 366).

One of the most appealing features of latent trait theory is that item parameters theoretically do not change across any sub-groups of a population, whereas in traditional test theory, item parameters depend on a

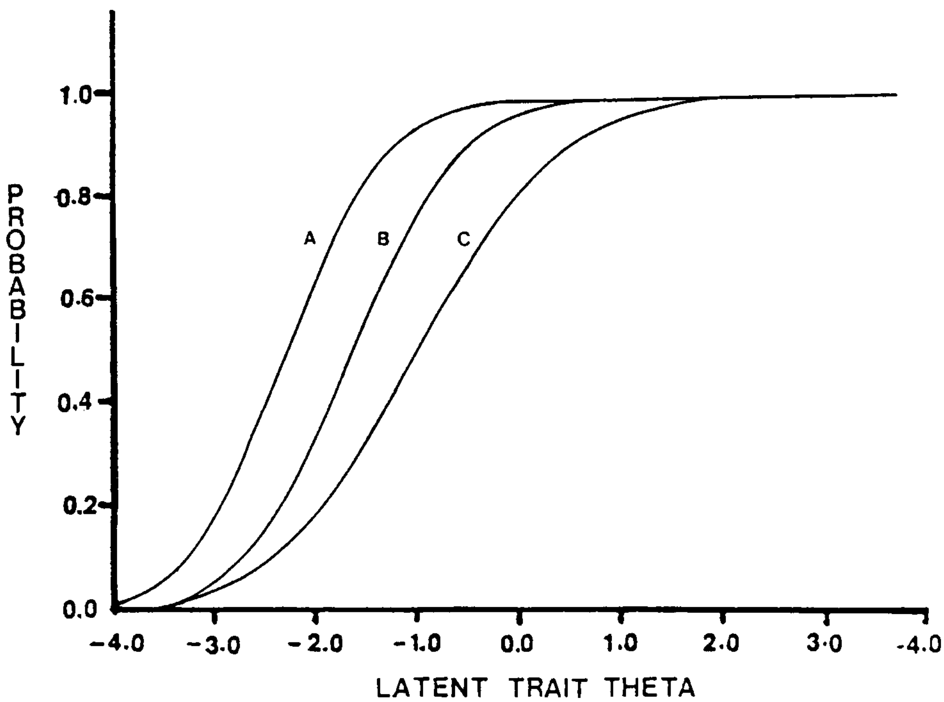


Figure II.1. Several item characteristic curves of the normal ogive model form: (A) $a_1 = 1.243$, $b_1 = -2.282$; (B) $a_2 = 1.165$, $b_2 = -1.620$; (C) $a_3 = 0.856$, $b_3 = -1.004$

specific group of examinees and, therefore, are not exactly the properties of the test item.

Item Characteristic Function

One of the most useful concepts in latent trait theory is its capacity to produce the item characteristic function, which is the regression of item score on ability. (Lord and Novick, 1968, p. 360). Lord and Novick stated that item characteristic is a key concept for making references about unobservable latent traits from the observed item responses. It is defined as follows (Lord and Novick, 1968, p. 362):

$$P_g(\theta) = \text{Prob}\ [u_g = 1 | \theta] \ , \tag{2.6}$$

$$Q_g(\theta) = 1 - P_g(\theta) \ . \tag{2.7}$$

If the model follows the normal ogive model (explained in the next section), the item characteristic function will be graphically depicted as shown in Figure II.2. The shape of this curve varies depending on each model and its item parameters; i. c., difficulty parameter and discrimination parameter. Examples of such variation are shown in Figure II.1.

Different Kinds of Latent Trait Models

In latent trait theory, at least four different categories of models are available for use:

1. One parameter model (Rasch model);
2. Two parameter models:

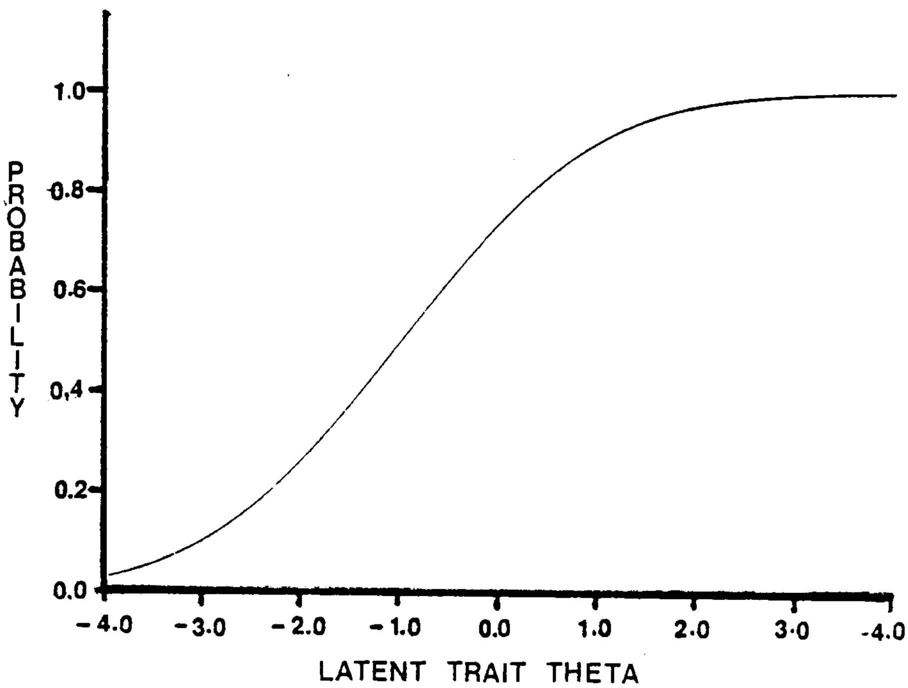


Figure II.2. Item characteristic function curve following the normal ogive model with $a_g = 0.641$, and $b_g = -0.983$.

2. Two parameter models:

   (a) Normal ogive model;

   (b) Logistic model;

3. Three parameter models:

   (a) Normal ogive model;

   (b) Logistic model;

4. Samejima's new family of models.

All these models will be explained in this section.

One parameter model. The one parameter model is also called the one parameter logistic model or Rasch model. This model was developed by George Rasch, a Danish mathematician in 1960. The equation of the item characteristic function for this model is:

$$P_g(\theta) = \frac{1}{1 + e^{-D(\theta - b_g)}} \quad . \qquad (2.8)$$

(Birnbaum, 1968, p.402)

All item discrimination parameters in the Rasch model are considered the same, which is unity. Therefore, the shapes of perfectly fitting items are identical. Since only one parameter, the difficulty parameter, is used in the Rasch model, it is simple to use, compared to the other models. This model is also a special case of the two or three parameter logistic model. Formula 2.8 can be obtained from the formula 2.10 (p. 22) by substituting unity for the item discrimination parameter a_g.

Two parameter models. There are two kinds of two parameter models: (a) normal ogive model; (b) logistic model. The normal ogive model was first developed by Lawley in 1943 and was primarily developed for binary items (Lord and Novick, 1968, p. 369). The item characteristic function is given by:

$$P_g(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_g(\theta - b_g)} e^{-\frac{u^2}{2}} \, du \, . \tag{2.9}$$

Samejima (personal communication, 1983) shows that as θ increases, $a_g(\theta - b_g)$ increases and so does the shaded area of the distribution in Figure II.3. Thus, $P_g(\theta)$ is strictly increasing in θ. Formulas for the item discrimination parameter and the item difficulty parameter for this model used in this study are shown in Chapter IV.

The logistic model was developed by Birnbaum for the mathematical convenience. The item characteristic function of this model is:

$$P_g(\theta) = \frac{1}{1 + e^{-Da_g(\theta - b_g)}} \, . \tag{2.10}$$

(Birnbaum. 1968, p.400)

Birnbaum (1968, p.400) stated that the scaling factor D is usually 1.7 to maximize the agreement between the logistic model and the normal ogive model.

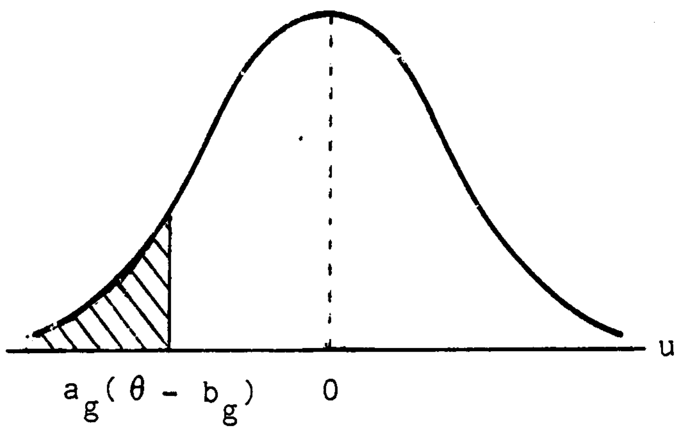


Figure II.3. Standard normal density function.

Three parameter models. Three parameter models involve the guessing parameter c_g developed by Birnbaum (1968), in addition to the two above-mentioned parameters. They were developed for the multiple-choice test items. The item characteristic function is defined by him as:

$$P_g^*(\theta) = P_g(\theta) + [1 - P_g(\theta)]\ c_g \quad (2.11)$$

(Birnbaum. 1968, p.404)

$P_g^*(\theta)$ is item characteristic function of either the logistic model or the normal ogive model. c_g was originally given as unity divided by the number of alternatives (Birnbaum. 1968, p. 404). This model was developed based on either knowledge or random guessing principle which presupposes the examinee's random guessing. Samejima (personal communication, 1982) draws its function compared to $P_g(\theta)$ as shown in Figure II.4.

Samejima's new family of models. Samejima's models are very different from the other models:

1. Each alternative as well as the correct answer of the multiple-choice item has important information about the examinee's ability.
2. The examinee's random guessing has an effect on the item characteristic function.

In other models for the multiple-choice test items, such as the three-parameter normal ogive model or the three-parameter logistic model, the examinee's responses are treated as either correct or incorrect. In other words, each wrong answer is not considered as a source to give any specific information about the examinee's ability, other than the fact that he/she does not know the correct answer. These models are called Equivalent Distractor Models. On the other hand, in Samejima's new family of models, it is assumed that there is some simple statistical relationship between each distractor and the ability (Samejima, 1979, 1981). Separate distractors, as well as the correct answer, can provide information regarding the examinee's ability. In Samejima's new family of models, the correct answer and the distractors

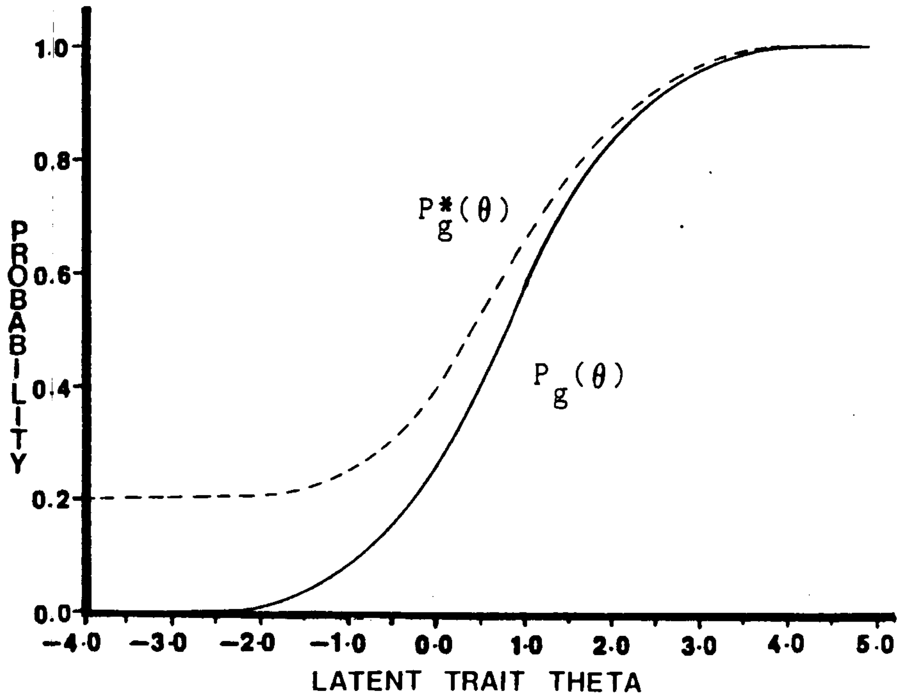


Figure II.4. Item characteristic function of three parameter models.

are treated as polychotomous responses. Samejima's models also deal with the random guessing effect. According to Samejima (1981), when the examinees do not know the correct answer, most of them take a guess from the strong pressure for success. Therefore, the item characteristic function for each alternative or the correct answer is affected by this guessing. Samejima's models deal with the polychotomous instead of the dichotomous response, and response examples of the operating characteristics are shown in Figure II.5. The curve of each alternative provides information regarding the subject's performance at the given ability level.

Information Functions

The information function is an important concept in latent trait theory. There are four different information functions:

1. Item response information function
2. Item information function
3. Response pattern inforamtion function
4. Test information function

Information functions provide measures of accuracy in estimating the ability at each ability level.

Item response information function. Samejima (1973) described the item response information function as a measure of accuracy estimating the examinee's θ from his

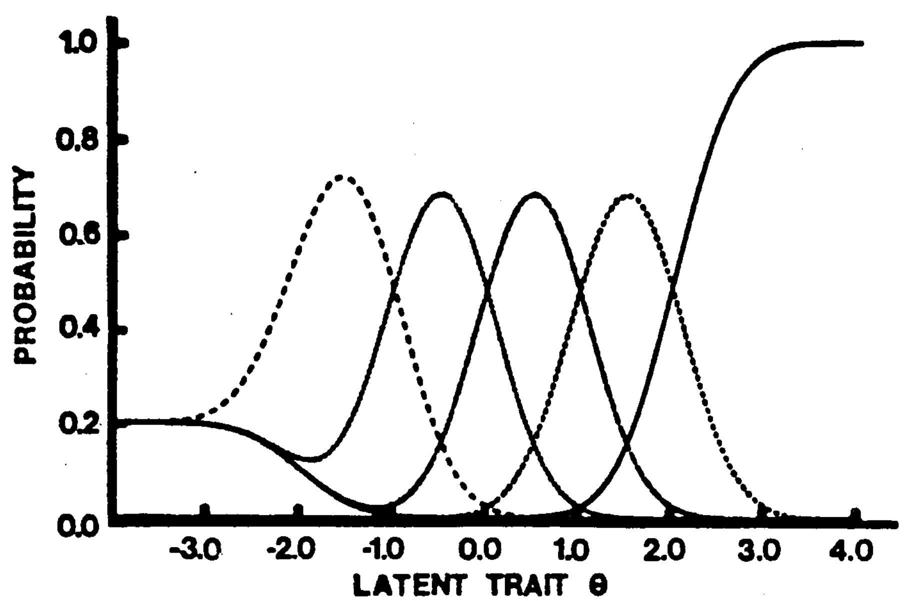


Figure II.5. Operating characteristic of Model A, Type 1 of Samejima's new family of models: the parameters are $a_g = 2.00$, $b_1 = -2.00$, $b_2 = -1.00$, $b_3 = 0.00$, $b_4 = 1.00$, and $b_5 = 2.00$ (reproduced with permission from Samejima; see Samejima, 1981, p.222 for more detail).

response pattern shared by each response. Item response information function $I_{x_g}(\theta)$ for a graded item score category x of item g is defined by Samejima (1969, ch. 6) as follows:

$$I_{x_g}(\theta) = -\frac{\partial}{\partial \theta} A_{x_g}(\theta) = -\frac{\partial^2}{\partial \theta^2} \log P_{x_g}(\theta) \quad . \qquad (2.12)$$

(Samejima, 1977a, p.234)

The item response information function is also defined for a continuous variable (Samejima, 1977a).

<u>Item information function</u>. Samejima (1969, p.38) defined the item information function for graded response as the conditional expectation of the item response information function, given θ:

$$I_g(\theta) = E[I_{x_g}(\theta)|\theta] = \sum_{x_g=0}^{m_g} I_{x_g}(\theta)\ P_{x_g}(\theta) \quad . \qquad (2.13)$$

The item information function $I_g(\theta)$ for the binary item is defined by Birnbaum (1968, p. 454) as below:

$$I_g(\theta) = \frac{[P_g'(\theta)]^2}{P_g(\theta)Q_g(\theta)} \quad . \qquad (2.14)$$

Information from a specific item is independent of that from all other items in the test.

Hambleton and Cook (1977, p.86) stated:

> The height of the item information curve at a particular ability level is a direct measure of the usefulness of the item for precisely measuring ability at that level.

They also said (p. 85):

> The higher the slope of the item characteristic curve and the smaller the conditional variance, the higher will be the information curve at that particular ability level.

It can be seen in Figure II.6. The item with $a_1 = 1.14$ is the most informative at the ability level of -0.4.

Response pattern information function. Samejima (1973) stated that the response pattern information function is a direct measure of accuracy in estimating the individual's ability assigned to the individual response pattern. Response pattern information function, $I_V(\theta)$ is defined mathematically by Samejima in 1969 as follows:

$$I_V(\theta) = - \frac{\partial^2}{\partial\theta^2} \log P_V(\theta) = \sum_{x_g \epsilon V} I_{x_g}(\theta) \ . \qquad (2.15)$$

(Samejima, 1977a, p.78)

The information functions 1 and 2 were developed by Samejima (1969). Further information regarding them can be found in her publications (1969, 1973, 1977b).

Test information function. The properties of the test information function are very similar to those of the item information function, except that the test

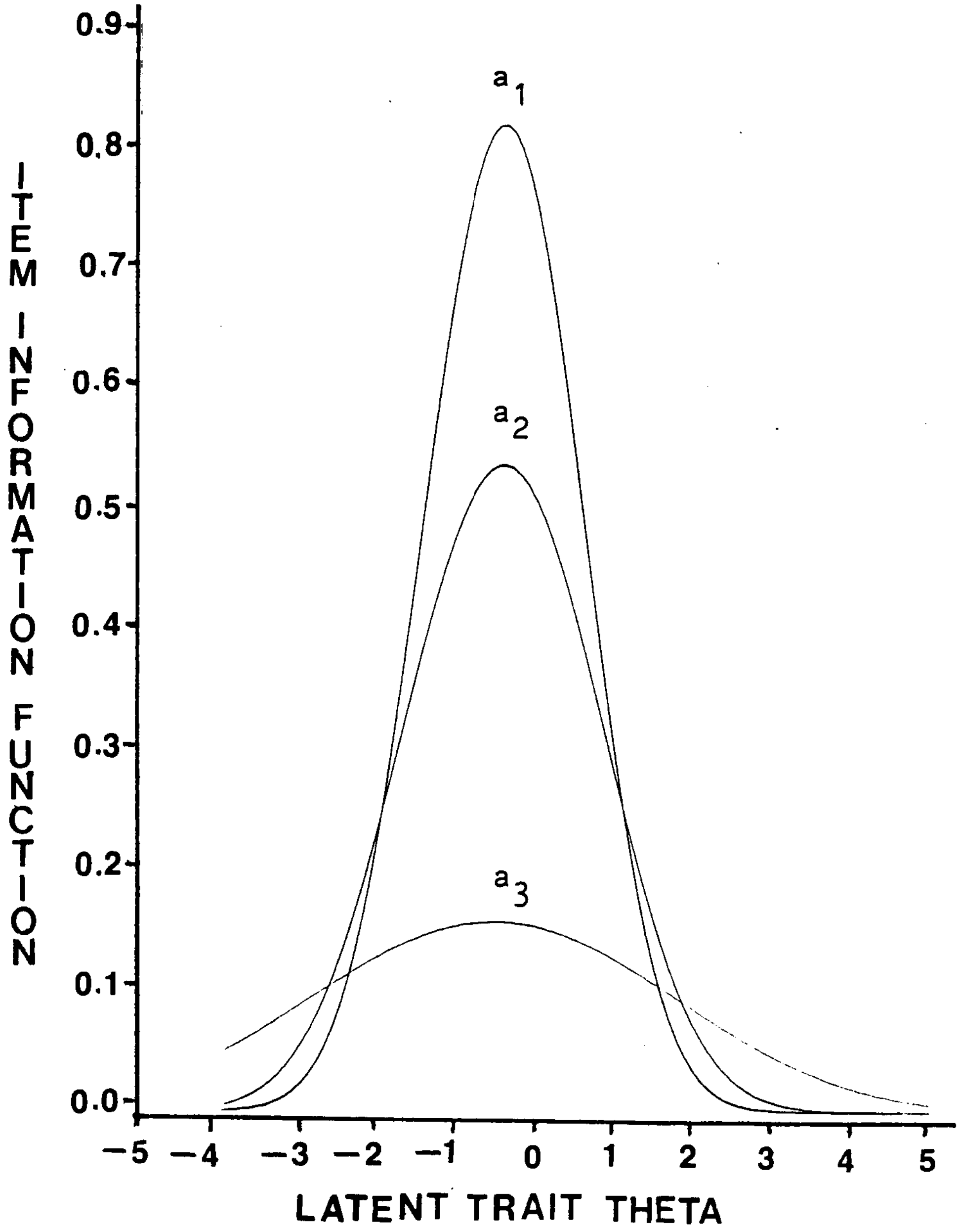


Figure II.6. Item information functions: $a_1 = 1.14$, $a_2 = 0.92$, and $a_3 = 0.79$

information function, $I(\theta)$, is the summation of individual item information functions. Samejima (1969, p.37) defined the test information function for the graded response as follows:

$$I(\theta) = \sum_{V} I_V(\theta)\ P_V(\theta) = \sum_{g=1}^{n} I_g(\theta)\ . \qquad (2.16)$$

The test information function for the binary item is defined by Birnbaum (1968, p.454):

$$I(\theta) = \sum_{g=1}^{n} \frac{[P_g(\theta)']^2}{P_g(\theta)\ Q_g(\theta)}\ . \qquad (2.17)$$

This is a special case of of the graded response. Each item contributes independently to the test information function.

Error of Estimation

There are significant differences regarding the error of estimation between latent trait theory and traditional test theory. Lord (1980, p.7) defined the standard error of measurement σ_e, in traditional test theory as follows:

$$\sigma_e = \sigma_X[1 - \rho_{XX'}]^{1/2}\ . \qquad (2.18)$$

The test reliability index, ρ_{XT}, is defined as the correlation of observed scores with true scores, and satisfied the following:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_X - \sigma^2_e}{\sigma^2_X} \quad ,$$

so,

$$\sigma^2_X - \sigma^2_e = \sigma^2_X \ \rho^2_{XT} \quad .$$

Therefore,

$$\sigma^2_e = \sigma^2_X \ (1 - \rho^2_{XT}) \quad . \tag{2.19}$$

$$\sigma_e = \sigma_X \ (1 - \rho^2_{XT})^{1/2} \quad .$$

Reliability coefficient $\rho_{XX'}$ is derived from a specific group of examinees (Samejima, 1977c, p.196). Errors of estimation using both the index of reliability and the reliability coefficient are also influenced by the nature of the total group of examinees. Unlike traditional test theory, latent trait theory provides the standard error of estimation, free from any specific examinee group (Samejima, 1977c, p.196). The standard error of estimation is given by the reciprocal of the square root of the test information function, $I(\theta)$ (Birnbaum, 1968).

The amount of the test information varies with the ability level of the examinee. Therefore, the standard error of estimation also automatically varies at each ability level. If the test information function is high. the error of estimation is low. With a low information function, the error of estimation is high. For example:

If the test information function = 18,

the standard error of estimation = $\frac{1}{\sqrt{18}}$ = 0.2357.

If the test information function = 8,

standard error of estimation = $\frac{1}{\sqrt{8}}$ = 0.3536.

In summary, traditional test theory provides us with only one standard error of estimation to all examinees. However, at each ability level, the standard error of estimation varies with the fixed level in latent trait theory. Therfore, each examinee possibly has a different error of estimation.

Adaptive Testing

One of the most valuable applications of latent trait theory, as compared to traditional test theory, is adaptive testing. The purposes of adaptive testing are: (1) to obtain precise information about an individual's ability regarding a specific subject area; and (2) to minimize the administration time. In order to use adaptive testing, a suitable item pool is needed. An item pool consisting of a large number of items with established item characteristics is needed so that appropriate items can be selected to estimate the examinee's ability.

In adaptive testing (also called branched testing, individualized testing, tailored testing or programmed testing), a subset of items is chosen to make the most efficient ability measurement for an individual examinee without spending too much testing time. While an examinee is taking a test, the choice of items to be presented to him/her is determined by the purpose of the test administration and his/her ability.

There are some important reasons why adaptive testing requires the use of latent trait rather than traditional methodology:

1. In traditional test theory, the sailent feature of a test is considered to be the reliability coefficient which depends not only on the test items but on a specific examinee group. In contrast, latent trait theory defines the standard error of estimation as a function of ability, and considers this to be the major index of dependabiity, instead of the reliability of a whole test. The standard error of estimation, defined as $1/\sqrt{I(\theta)}$, is free from any subject group. The standard error of estimation changes for different levels of ability. This information is not available in traditional test theory.
2. In traditional test theory, the main concern is a set of items (a test), whereas latent trait theory starts from an individual item. Therfore, we can easily find which item measures a certain level of examinees more efficiently than others.

CHAPTER III

PROCEDURES

The following procedures were employed in this study:

1. Description of the field test
2. Selection of subjects.
3. Production of frequency distributions of alternatives.
4. Computation of tetrachoric correlation coefficients.
5. Utilization of principal factor solution.
6. Computation of item parameters.
7. Production of item characteristic functions.
8. Computation of maximum likelihood estimates of ability and production of plausibility functions.
9. Production of information functions.
10. Conventional methods of analysis.

Description of the Field Test

A previously administered and analyzed field test of Tennessee Proficiency Test of Mathematics was chosen for reanalysis. The test consisted of 76 multiple-choice items, each having four response alternatives. The objective description for each item, as defined by the State Department of Education (1982), is as shown in Table III.1.

Table III.1. Objective descriptors of test items of the field test of Tennessee Proficiency Test of Mathematics

	Item #	# of items	Description
1	1 to 4	4	Adding whole numbers
2	5 to 8	4	Subtracting fractions
3	9 to 12	4	Dividing fractions
4	13 to 16	4	Adding decimal numbers
5	17 to 20	4	Subtracting decimal numbers
6	21 to 24	4	Dividing a decimal number by a whole number
7	25 to 28	4	Place value
8	29 to 32	4	Decimal number and percent equivalency
9	33 to 36	4	Percent of a number
10	37 to 40	4	Dividing whole numbers
11	41 to 44	4	Fraction and decimal number equivalency
12	45 to 48	4	Finding perimeter of rectangles
13	49 to 52	4	Finding area of rectangles
14	53 to 56	4	Linear measurement
15	57 to 60	4	Writing a number for a word name
16	61 to 64	4	Rounding off numbers
17	65 to 68	4	Equivalent measurements
18	69 to 72	4	Finding the largest or the smallest decimal fraction
19	73 to 76	4	Determining the arithmetic mean

According to the manual provided by the State Department of Education (1981), the following rules for administration were specified:

1. The principal was allowed to schedule the administration of the field test at a time during those two weeks that best fit into the local - school schedule.
2. The field test was untimed. Pupils were allowed to spend enough time to finish the test.
3. The following types of students were not to take the field test; or, if they did, their answer sheets should not be forwarded to the State Testing and Evaluation Center: (a) verified handicapped pupils who spend one-half or more of the school day in a special education program; and (b) pupils for whom English is a second language if English usage is a handicapping condition that would adversely affect test performance.
4. Make-up test: The State Department of Education did not request that the principal require make-up tests for pupils who were unable to take the field test during the scheduled testing session.

Results of a mathematics field test were chosen for reanalysis for the following reasons:

1. The test has curricular validity for Tennessee students. It covers most areas of mathematics usually taught through the sixth grade.
2. The field test was given to students in several grade levels. This is amenable to the purpose of this study, because it insures greater variability among the abilities of the subject pool.

Selection of Subjects

The field test was administered in 1981 to 408 sixth-, 542 seventh-, and 318 eigth- grade students in selected schools in Tennessee. Only sixth- and seventh-grade students were used in the present study, because the proportion of correct responses for each item for eighth-grade students was too high for the purpose of this study (see Table III.2).

The following procedures were taken in the subject elimination process:

1. Response patterns of seventh- and eighth-grade students who did not respond at least one item were listed.
2. Their response tendencies were observed.
3. Students who responded to fewer than 87 percent of 75 items were eliminated (see Appendix A).

As a result, 933 subjects were retained for this study.

Table III.2. Frequency distribution of the item with respect to their correct responses for each of the three grades.

Proportion of correct response	Grade		
	6	7	8
$p < .2$	4	0	0
$.2 < p < .3$	3	10	0
$.3 < p < .4$	11	2	0
$.4 < p < .5$	11	5	5
$.5 < p < .6$	10	12	15
$.6 < p < .7$	9	13	19
$.7 < p < .8$	11	11	9
$.8 < p < .9$	12	14	15
$.9 < p$	4	8	12
Total items*	75	75	75

Note: The proportion correct for each item in Table II.2 was computed based on the results provided by the State Evaluation Center, Knoxville, Tennessee.

*Item 26 was eliminated due to a scoring error at the beginning of this study. Therefore, 75 of 76 items were used.

Production of the Frequency Distribution of Alternatives

In order to investigate the frequency distributions of alternatives and the correct answer for each item, they were graphically depicted using programs available in the Statistical Analysis System (SAS) statistical software package. Some of these graphs are shown and discussed in Chapter IV. The rest can be found in Appendix B. Proportions of correct answers for each item were computed from the combined results of sixth- and seventh-grade students. These frequency distributions of alternatives and the correct answer, and proportion correct for each item are frequently used in the conventional methodology. Additional information obtained by latent trait theory compared to results by these conventional methods are discussed in Chapter IV.

Tetrachoric Correlation

Tetrachoric correlation coefficients among the 75 items were computed for use in the principal factor solution using Samejima's program (POLYCHO.FOR: FSCM-82-8). The assumption of tetrachoric correlation coefficients, a multi-normal distribution for the joint distribution of all possible pairs of items was considered to be met.

Utilization of Principal Factor Solution

A principal factor solution was employed to ascertain the number of dimensions or clusters of the 75 items as a set, and the factor loading of the first principal factor to compute the discrimination parameter. Factor analysis using a subprogram of the Statistical Package for the Social Sciences (SPSS) was performed on all 75 items. The forty-seven remaining items were again analyzed, twenty of which were subjected to a further analysis. These subsets were analyzed separately to determine dimensionality and to obtain estimates of parameters. The procedure to eliminate items from the original set of 75 was conducted as stated below.

All items with a high loading on the first factor only were retained. Then, some items with high loadings on both the first factor and other factors were eliminated. The items to be eliminated in each objective (subset) were determined by the number of factors with high loadings. When the high loading items were spread over two (or more) of the fourteen factors, either one or two items were eliminated. In cases where the high loading figures were very similar, only two were eliminated. When the figures were dissimilar, only one was eliminated. Proportion of correct answers, random guessing and the effectiveness of distractors were also considered within each objective in the elimination procedure. In Appendix C, the factor loadings of 75 items 47 items, and 20 items are listed.

Item Discrimination and Item Difficulty Parameters

The normal ogive model was used to generate item discrimination and item difficulty parameters. Samejima's computer program [FSCM-81-6] was used. In this program, the item discrimination parameter, a_g, is obtained by:

$$a_g = \rho_g (1 - \rho_g)^{-1/2} . \tag{3.1}$$

The difficulty parameter, b_g, is given by:

$$b_g = \phi^{-1} (1 - p_g) p_g^{-1} \tag{3.2}$$

(Samejima, 1981, pp. 205-206).

This program is written to compute item parameters from the proportion correct, and the correlation between item variable and theta on binary response which is obtained from the results of principal factor solution. In order to find out the relationship between discrimination and difficulty parameters, procedures available in the Statistical Analysis System (SAS) were used to plot pertinent values.

Production of Item Characteristic Curves

Two programs, ICF.FOR [FSPC-83-1] and PLOT.FOR [FSPC-80-18], by Samejima were used to produce the item characteristic curves for each item. The program ICF.FOR is to compute the probability of answering gth item correctly on each theta for each item.

In this study. values of the theta from -3.000 to 5.000 with the interval of 0.1000 were used to plot the item characteristic curves. Eighty-one possible values as a function of the ability were computed for each item. PLOT.FOR was used to actually plot the probability values against the theta (ability) on the graph. Some of the results are shown and discussed under Plausibility Functions in Chapter IV. The rest of the graphs are presented in Appendix D.

Maximum Likelihood Estimates of Ability

Maximum likelihood estimates of ability (MLE θ) for the 933 subjects were computed using Samejima's program MLETH.FOR (FSPC-80-7, FSCT-80-30). This program was written for: (a) the computation of the maximum likelihood estimate of latent trait theta (θ) for each subject which follows normal ogive model; and (b) the rearrangement of θ's in ascending order as well as their frequency distribution. In order to compute maximum likelihood estimates of ability using this program, the estimated discrimination parameter and difficulty parameter for each of the 75 items and individual subject's response pattern were used as a part of the input data.

The 933 subjects were categorized into 16 groups, based on maximum likelihood estimates of their abilities. For each response alternative for each item, the percentage of each subgroup choosing that response was plotted against the mean ability of the subgroup in order to determine the plausibility function. Plausibility function is the probability of choosing each distractor at a given ability. The lowest MLE θ, the highest MLE θ and the midpoint for each of the 16 groups are shown in Table III.3.

Production of Information Functions

Item information functions and the test information function were computed by using Samejima's program TAUTRN.FOR (FSCT-80-26. FSPL-78-63, FSPL-79-21, FSCT-80-24). The initial values of theta for the information function was -4.000 with each of the 90 intervals of theta, 0.1000 units wide. First, the highest value of the item information function among the ninety-one values was depicted graphically against its theta for each item using SAS program. Second, ninety-one values of the test information function were also plotted against their values of theta in the same manner. Third, the standard error of estimation for each theta was

Table III.3 Lowest and highest MLE θ's and midpoints for 16 groups.

GRP	Lowest MLE θ	Highest MLE θ	Mid point
1	-2.99719	-1.46087	-2.229030
2	-1.45692	-1.19841	-1.327665
3	-1.19562	-0.97843	-1.087025
4	-0.97695	-0.76086	-0.868905
5	-0.75869	-0.60451	-0.681600
6	-0.60137	-0.39613	-0.498750
7	-0.39485	-0.26574	-0.330295
8	-0.25414	-0.07129	-0.162715
9	-0.06961	0.10099	0.015690
10	0.10166	0.25964	0.180650
11	0.25982	0.47784	0.368825
12	0.47816	0.71066	0.594410
13	0.71403	0.96316	0.838595
14	0.97223	1.33935	0.155790
15	1.34549	1.91305	1.629270
16	1.91473	4.58566	3.250195

computed, using a portion of the results (square root of test information function) from Samejima's program. Values of the inverse square root of the test information function, which is the standard error of estimation, were depicted again, against their theta, using the SAS procedures. These graphs are shown and discussed under Test Information Function in Chapter IV.

Conventional Methods of Analysis

1. The biserial correlation coefficient for each item was computed to obtain the discrimination index in the conventional way. Biserial correlation coefficients, r_{bis}, (discrimination index in traditional test theory) for 75 items were computed using the following formula:

$$r_{bis} = \frac{r_{pbis}\ p(1 - p)}{y}\ . \qquad (3.3)$$

Since no statistical software package subprograms to compute the biserial correlation directly were available, the following procedures were employed: (a) computation of Pearson product-moments using SPSS: point biserial correlation coefficients and Pearson product-moments are the same; and therefore, (b) the values of r_{pbis} were appropriately substituted in the formula.

2. The frequency distribution of MLE θ and test score were depicted using SAS.

3. The relationship between point biserial correlation coefficient (discrimination index), biserial correlation coefficient, and p (proportion correct) for each item was depicted using SAS.

The graphs mentioned in this section are presented and discussed under Comparison of Latent Trait Theory Methods and Conventional Methods in Chapter IV.

CHAPTER IV

RESULTS

Frequency Distribution of Alternatives

The frequency distributions of alternative choices made by 933 students were categorized into four groups, based on the configurations of the frequency distributions of alternatives. Only one frequency distribution of alternatives for each category is shown in this section. This categorization was made to see if it was possible to predict the behaviors of distractors that are considered to be observed by using the plausibility functions. Some more examples are dicussed in a later section of this chapter. The rest of the distributions can be found in Appendix B.

Category 1. In this category the correct answer was obvious. Therefore, the proportion correct was distinct. Included in Category 1 were Items 1, 2, 3, 4, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 28, 37, 38, 49, 50, 51, 52, 53, 54, 56, 57, 58, 59, and 60. Response distributions for a typical example, Item 1, is shown in Figure IV.1.

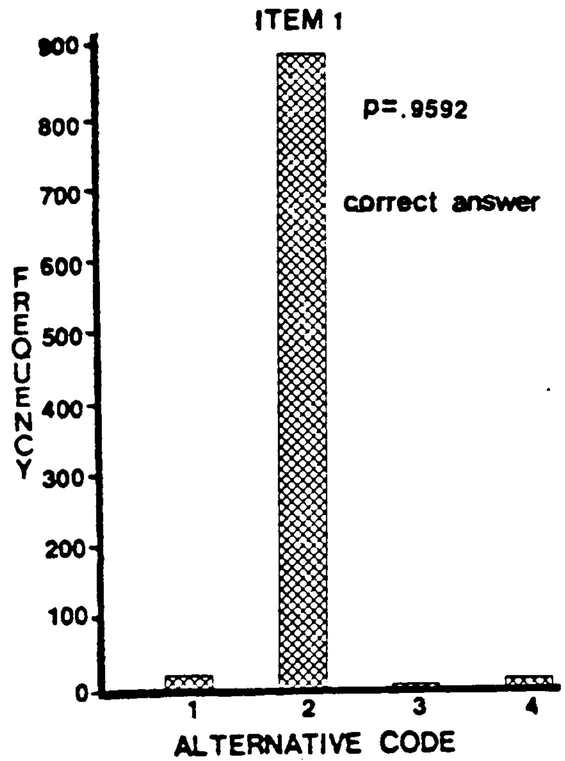


Figure IV.1. Frequency distribution of alternatives of Item 1.

Category 2. The correct answers of the items in this category are not as obvious as those in Category 1, but still very distinctive. A typical response distribution in Category 2, Item 7, is shown in Figure IV.2. The other items were 5, 6, 8, 9, 10, 11, 12, 24, 29, 30, 31, 32, 33, 35, 36, 39, 40, 55, 61, 62, 63, 64, 73, 74, 75, and 76.

Category 3. The proportion correct and the proportion of one distractor are almost the same. The item shown in Figure IV.3 (Item 66) is an example typical of this category. The other items in Category 3 are 34, 65, 67, 68, and 72.

Category 4. Figure IV.4 depicts representative response distribution for items in Category 4, Item 42. The proportion of one distractor was higher than the proportion correct. Other items included here were 41, 43, 44, 45, 46, 47, 48, 69, 70, and 71.

Results of Factor Analyses

The eigenvalues from the results of the three principal factor solutions are shown in Figure IV.5. Note that the eigenvalues of the first, second, third, and fourth factors for subsets of 75, 47, and 20 items are very similar. Therefore, only the results of 75 items were used in subsequent analyses. Since the results of the three principal factor solutions are very similar, 75

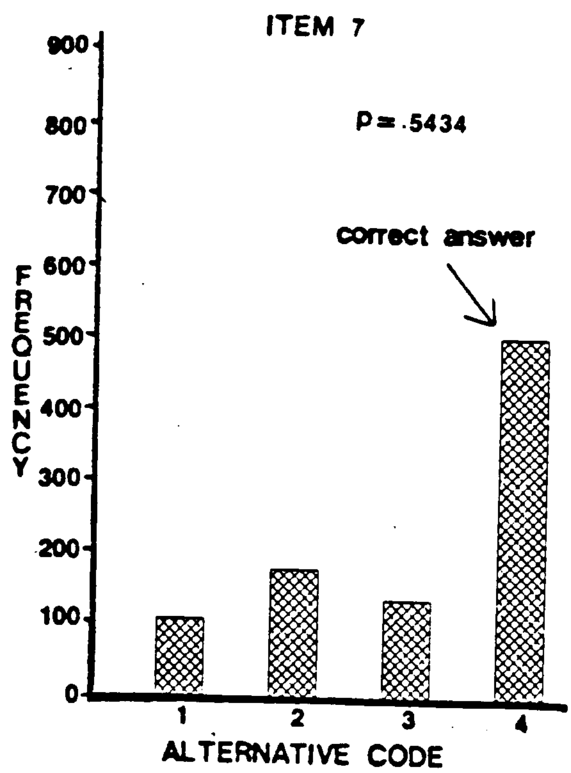


Figure IV.2. Frequency distribution of alternatives of Item 7.

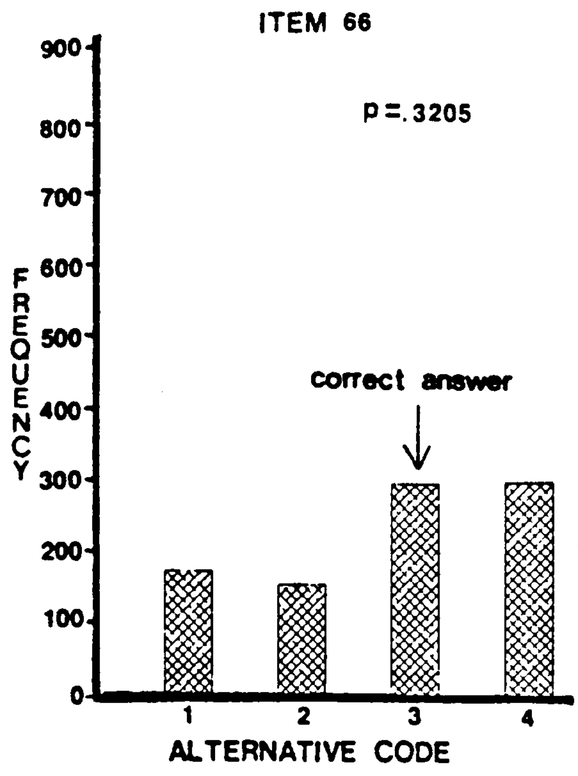


Figure IV.3. Frequency distribution of alternatives of Item 66.

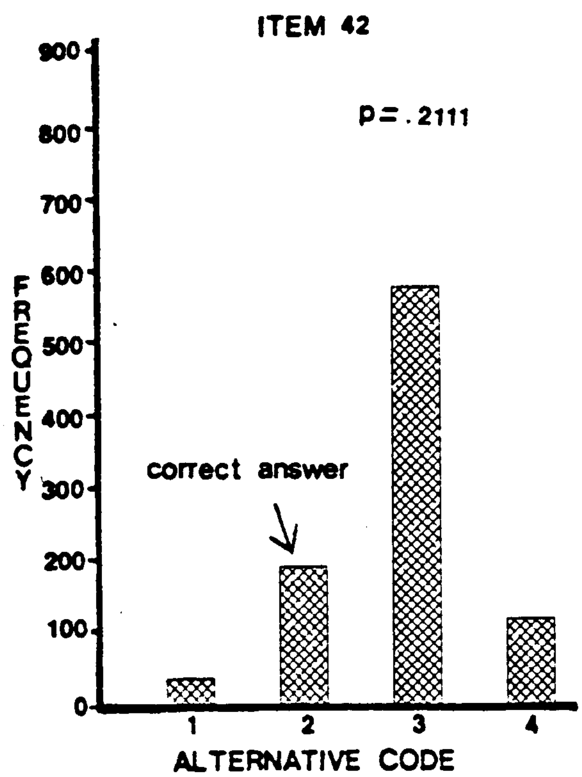


Figure IV.4. Frequency distribution of alternatives of Item 42.

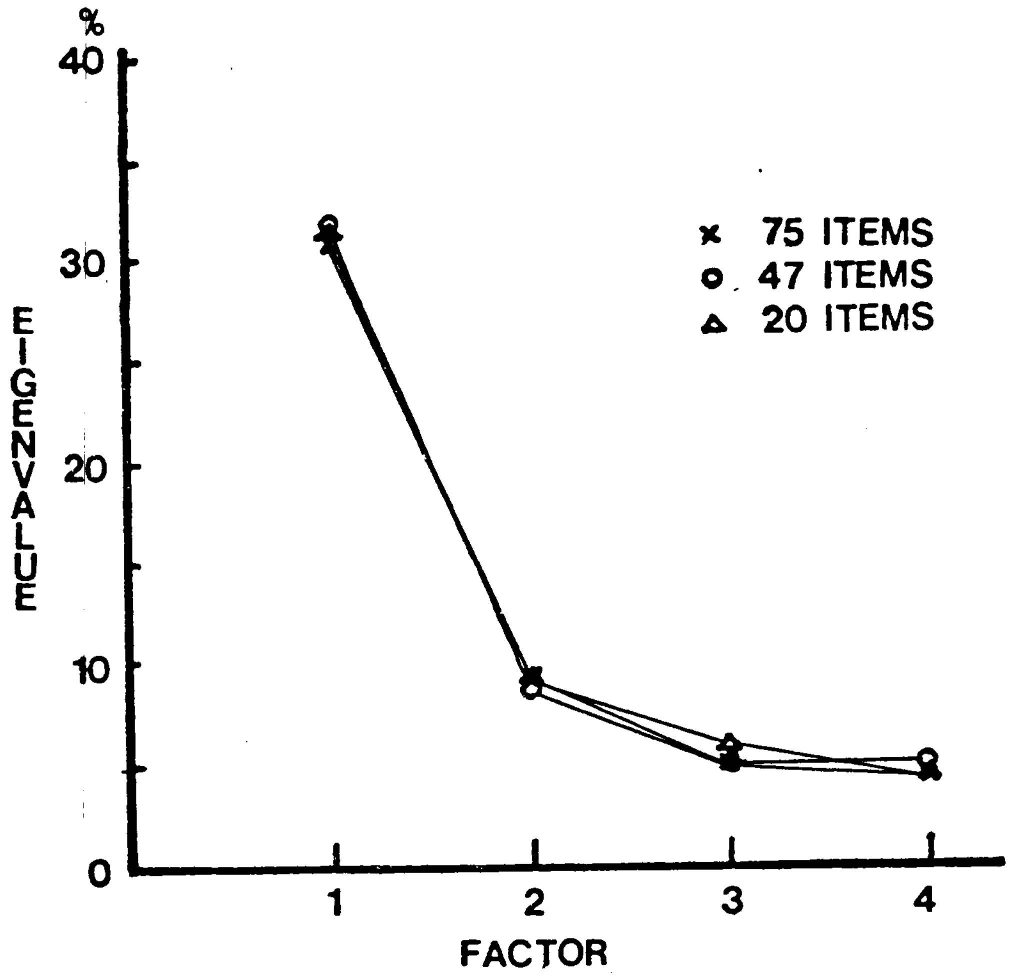


Figure IV.5. Eigenvalues from the results of the three principal factor solutions.

items would give more generalization than 47 items or 20 items. Even though eigenvalues of the second, third, and fouth factors were reasonably high, that of the first factor was much larger than the others. The first principal factor, therefore, was treated as the general factor, θ.

Item Paramters

Objectives could be grouped according to their item parameters as listed below. This classification was made to see possible relations between categories of frequency distributions of alternatives and item parameters. In so doing, it was hoped that some additional information would be found from item parameters.

High discrimination (i.e., discrimination parameter higher than .90000).

1. Subtraction of fraction--Category 2 of frequency distributions: Items 5, 6, 7, and 8.
2. Division of numbers with decimal points (e.g., 45.5 divided by 5)--Category 1 of frequency distributions: Items 21, 22, and 23.

These results showed that items with high discrimination parameters also had high proportions correct in this study.

Low discrimination (i.e., discrimination parameter less than .39000).

1. Linear measurement--Category 1 of frequency distributions: Items 53, 54, and 56.
2. Equivalent measurement--Category 3 of frequency distributions: Items 65, 66, 67, and 68.

Items with low discrimination either had very high proportion correct or had one attractive distractor (similar proportions both for the correct answer and the distractor).

High difficulty (i.e., difficulty parameter higher than 1.15000).

1. Decimal equivalency--Category 2 of frequency distributions: Items 29, 30, 31, and 32.
2. Equivalent measurement--Caytegory 3 of frequency distributions: Items 65, 66, 67, and 68.
3. Finding the perimeter of a rectangle--Category 4 of frequency distributions: Items 45, 46, 47, and 48.

Except (1), items with high difficulty parameters had one or more distractors which had considerable proportions of responses.

Low difficulty (i.e., difficulty parameter lower than -1.50000).

1. Addition of three digit numbers and numbers with decimal points: Items 13, 14, 15, and 16
2. Writing a number for a word: Items 57, 58, 59, and 60
3. Subtraction of numbers with decimal points: Items 17, 18, 19, and 20

These items all belong to Category 1 of frequency distributions of alternatives, which means the proportions correct are high. The relationship between the two item parameters showed that the values of the discrimination parameters are at the maximum when its difficulty level is about between -0.5 to 0.1 as shown below under Comparison of Latent Trait Theory Methods and Conventional Methods in this chapter.

Maximum Likelihood Estimates of Ability (MLE θ)

The range of MLE θ was from -2.99719 to 4.58566 as shown in Table III.3 (p. 45). The frequency distribution of MLE θ was graphically depicted and is shown later in this chapter.

Test Information Function

The ability (θ) at which this test was most informative was around -0.5, as shown in Figure IV.6. Also, the standard error of estimation is very small

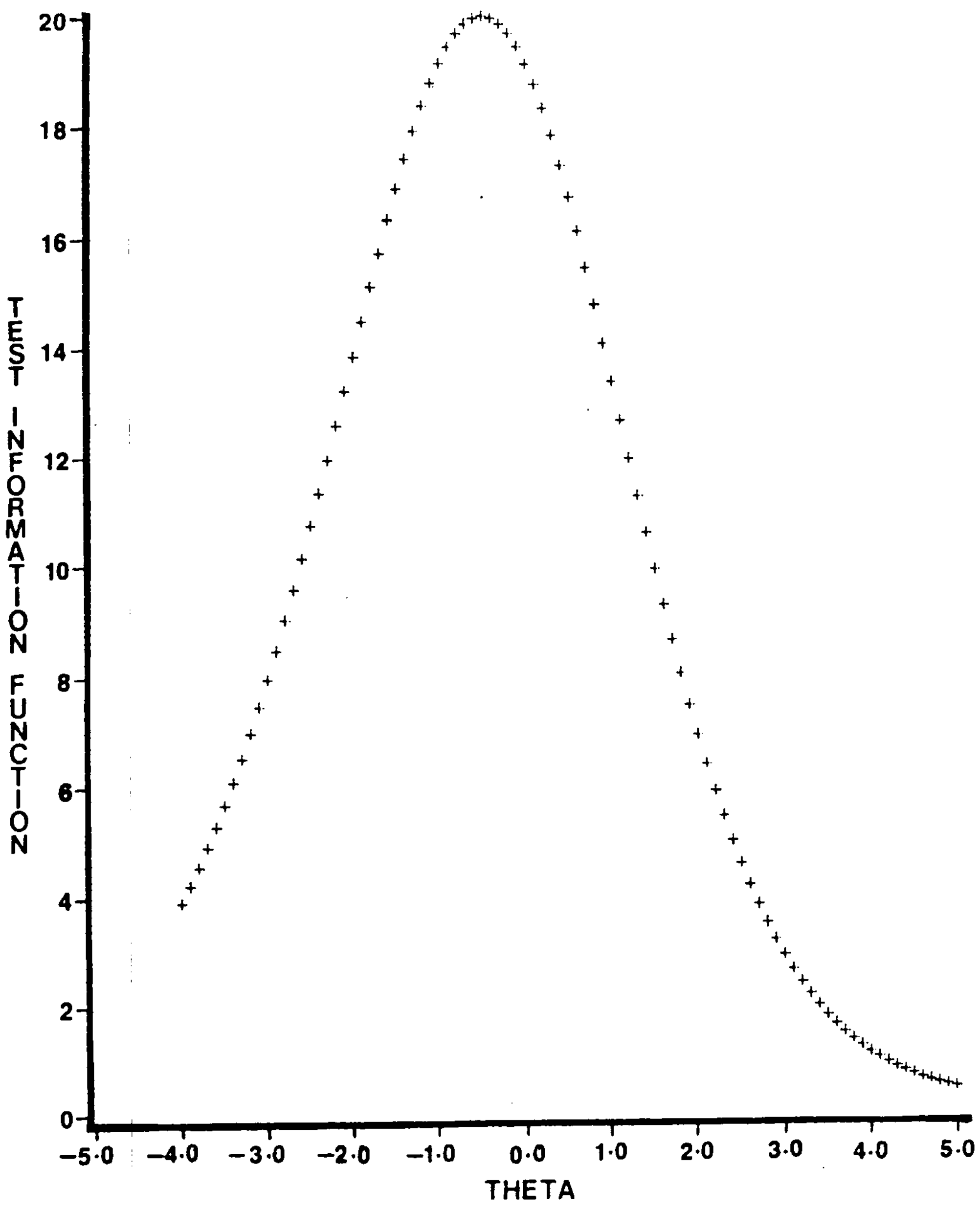


Figure IV.6. Test information function of 75 items around -0.5 in Figure IV.7, which is what one would expect when a test is most informative. Figure IV.8 presents the highest value of the item information function and the corresponding value of θ for each of 75 items. It indicates that those values tend to be large around $\theta = -0.5$. This finding confirms the findings from the test information function and the standard error of estimation. Figure IV.9 also shows that the relationship between the ability and the estimated ability was the closest at about the ability level. -0.5. In other words, the error of estimate of ability was minimized at around that level.

Plausibility Functions (distractor analyses)

The results of plausibility functions and item characteristic functions may be categorized into seven groups. This categorization was made to observe: (a) if the normal ogive model will fit each item; and (b) if not, what kind of additional information it may provide. The normal ogive model was considered not to fit an item, if: (a) the plot of the correct answer deviated from the item characteristic curve and formed a U-shape; or (b) the plot of the correct answer was erratic throughout the length of the item characteristic curve. The relations among groups were schematically depicted in Figure IV. 10.

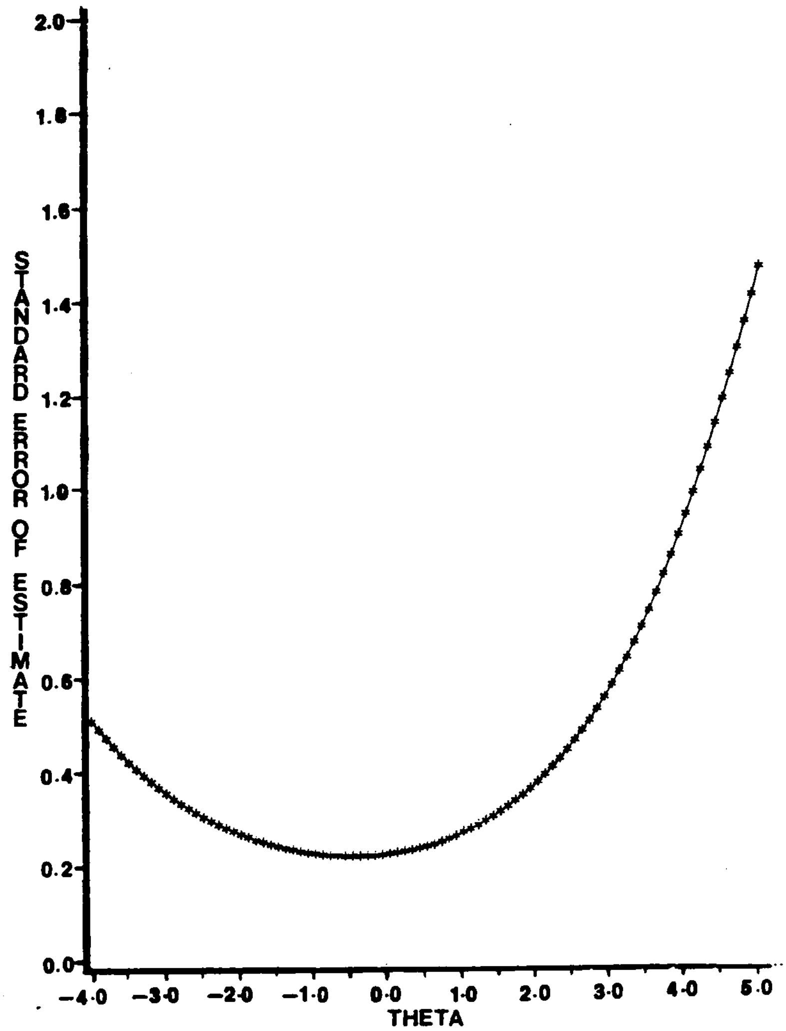


Figure IV.7. Standard error of estimation of 75 items

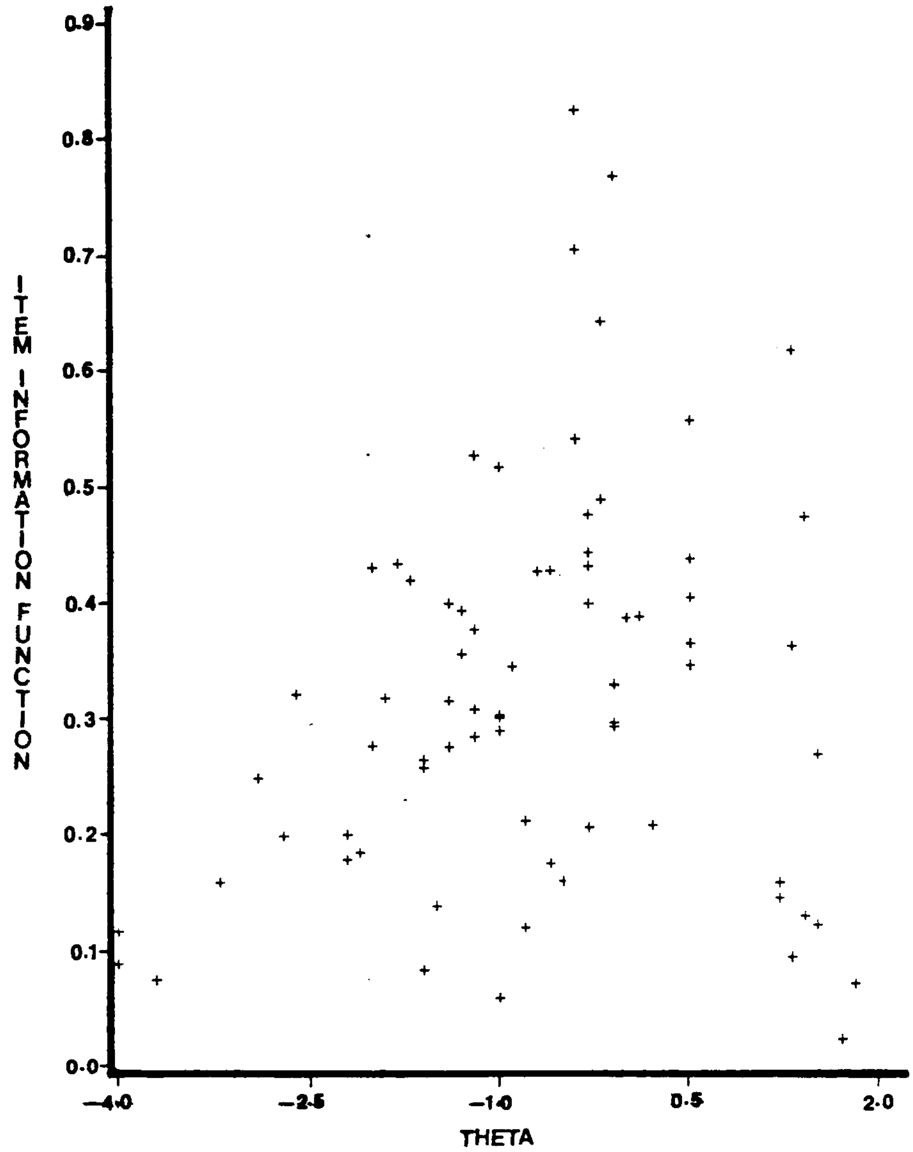


Figure IV.8. The highest value of the item information function and corresponding theta for each of 75 items

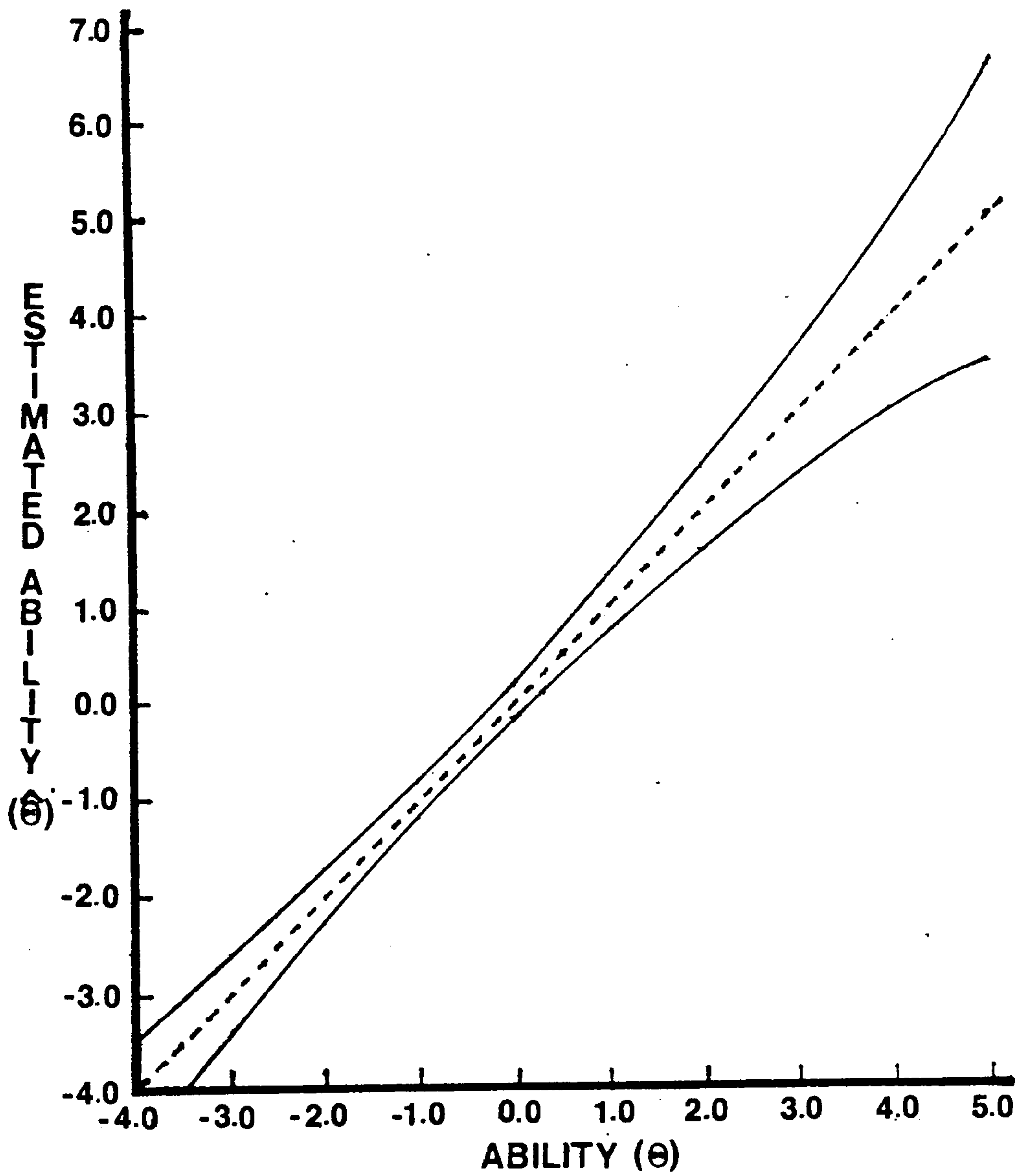


Figure IV.9. Relationship between theta and estimated theta

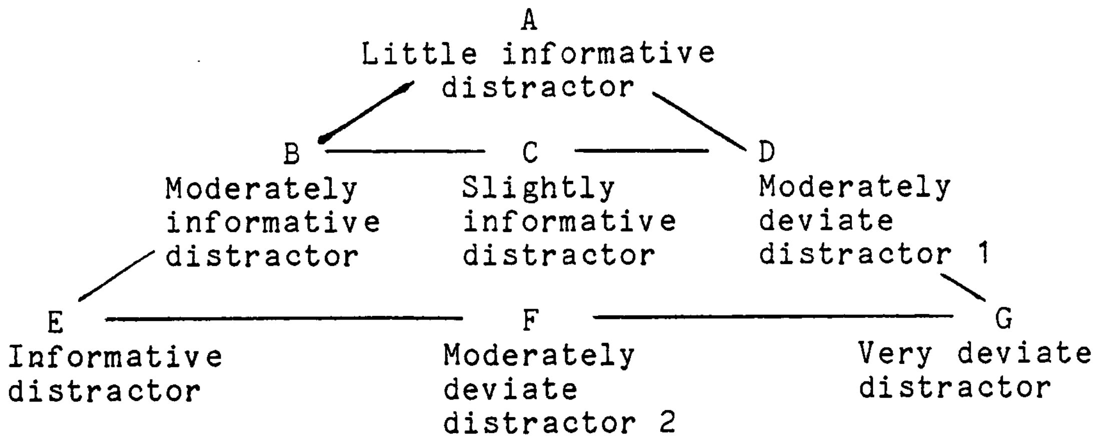


Figure IV.10. Seven groups of plausibiliy function

First of all, the plausibility functions of Categories A, E, and G are distinctively different from each other. The normal ogive model most closely fitted items in Group A, while these items provided the slightest information from the distractors. Group E is a group where the plausibility function of each item did not follow the item characteristic function very well, but one distractor's behavior was distinctively active. The plausibility functions of items in Group G were most deviate from the item characteristic functions. Behaviors of more than one distractor were erratic in this group. The rest of the items were placed among those distinct groups based on the relative closeness of their plausibility functions to those of the items in the three groups. Also, those items which the normal ogive model closely fitted--Groups A, B, C, and D--were placed at the top of the triangle. On the other hand, the bottom groups--Groups E, F, and G--are items which the normal ogive model did not satisfactorily fit.

Group B was placed between A and E, because one distractor's behavior started to be conspicuous. In Group D, more than one distractor were moderately active, so it was placed between A and G. The behavior of a single distractor of each item in Groups C and F appears to be still more active than the other distractors, but they also have started to be active. C and F were, therefore, considered to fall between B and D, and E and G, respectively.

A typical item from each of the seven groups was described in a graphic form. The items tending to fall into each group are listed below. Alternative codes, 1, 2, 3, and 4 are represented by the following symbols:

1: ●———● ;

2: △———△ ;

3: ○———○ ;

and, 4: ×———× ;

while solid lines represent item characteristic curves.

Group A: Little information from distractors group. None of the distractors were attractive to the subjects. Proportions correct were between about .7000 and .9600 ($.7000 < p < .9600$) in this category. A typical example, Item 1, is shown in Figure IV.11. Some of the items in this group had a very small range of the probability of getting the correct answer. This indicates that this item was too easy for the purpose of effectively distinguishing ability differences. The following items were categorized into this group: Items 1, 2, 3, 4, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 27, 28, 37, 38, 56, 57, 58, 59, and 60.

Group B: Moderately informative distractor group. One distractor of each item in this group is slightly more attractive than others to the subjects, especially low ability subjects. Figure IV.12 shows a typical example, Item 49. The plausibility function of the correct answer still followed the item characteristic curve well. The other items in this group were: Items 30, 39, 49, 50, 51, 52, 54, and 55.

Group C: Slightly informative distractor group. To some extent, all distractors had information about subjects with low abilities as seen in Figure IV.13 (Item 24). This group of items may give information about subjects similar to that obtained by Group B. However,

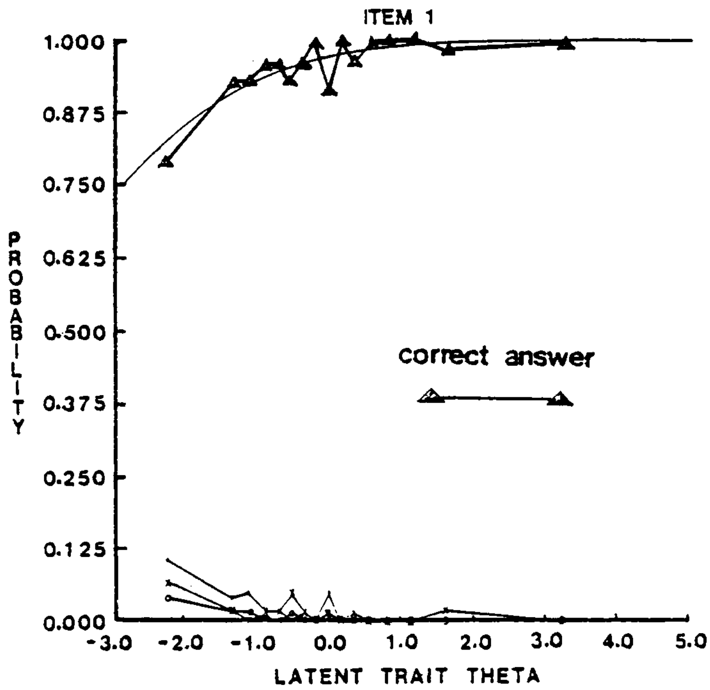


Figure IV.11. Item characteristic curve and plausibility function of Item 1 (Group A): $a_g = 0.43057$; $b_g = -4.40562$. All distractors provide little information.

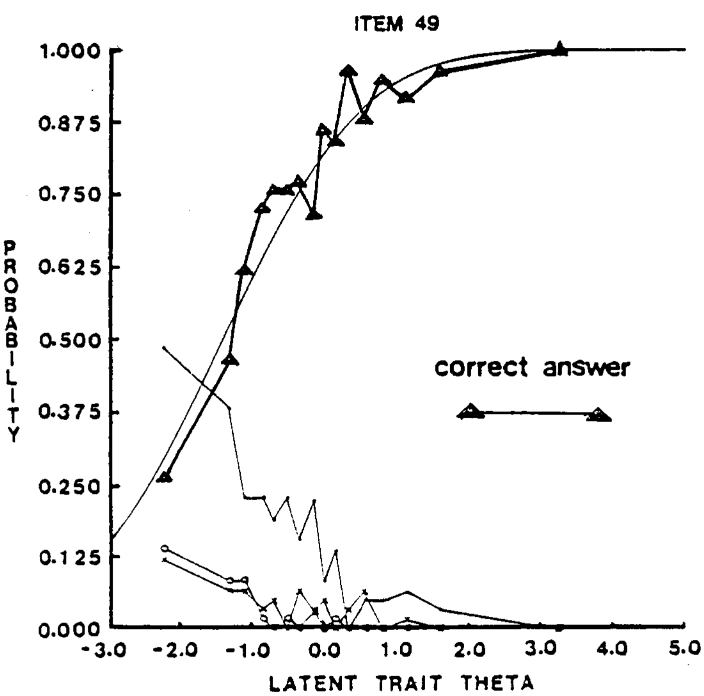


Figure IV.12. Item characteristic curve and plausibility function of Item 49 (Group B): a_g = 0.68707; b_g =-1.01375. One distractor is moderately informative.

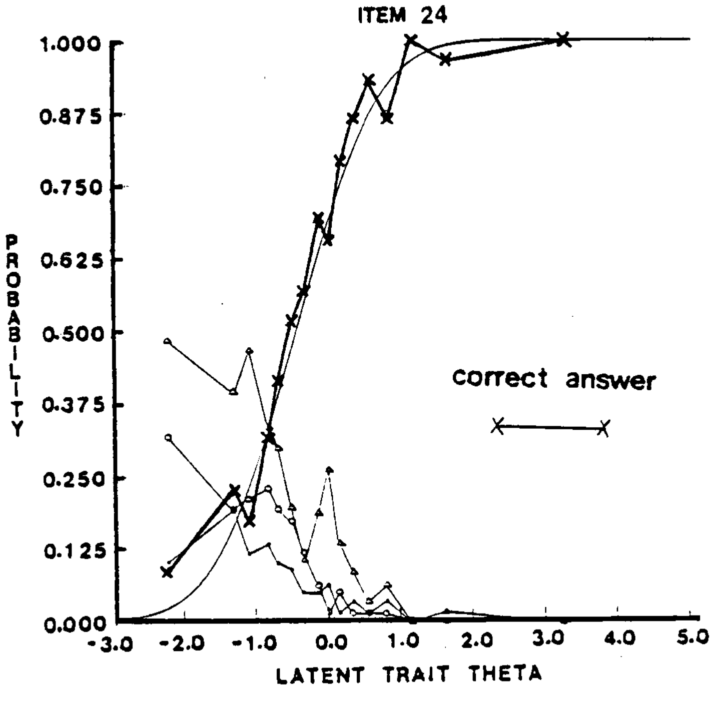


Figure IV.13. Item characteristic curve and plausibility function of Item 24 (Group C): a_g = 1.13798; b_g = -0.04257. One distractor is slightly informative.

the amount of information given by a single distractor was not as much. The other items included in this group were: Items 24, 34, and 40.

<u>Group D: Moderately deviate distractor group 1</u>. A typical example is shown in Figure IV.14 (Item 7). These distractors can provide some degree of information regarding subjects with low abilities, and to some extent, these distractors did not behave consistently. The range of the probability of getting the correct answer was very wide: In this example, it ranges from below 0.125 to near 1.000. The range of distinguishable theta is similar to those of items in Groups B and C. The following items were the other items belonging to this group: Items 5, 6, 8, 9, 10, 11, 12, 31, 32, 33, 35, 36, 53, 61, 62, 63, and 64.

<u>Group E: Informative distractor group</u>. Obviously, as seen in Figure IV.15 (Item 42), one distractor appears to be very attractive to subjects with below average ability. Also, within this low ability level, the function of the correct answer considerably deviates from the item characteristic function. Although this type of items could distinguish a wide range of abilities effectively, the item characteristic curves following the normal ogive model may give the false estimation regarding

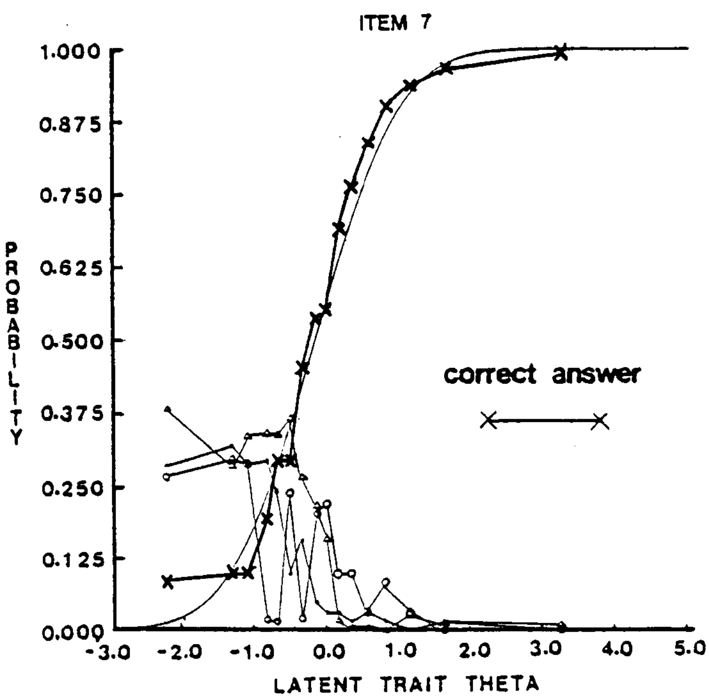


Figure IV.14. Item characteristic curve and plausibility function of Item 7 (Group D): a_g = 1.09826; b_g = -0.14745. All distractors are moderately deviate.

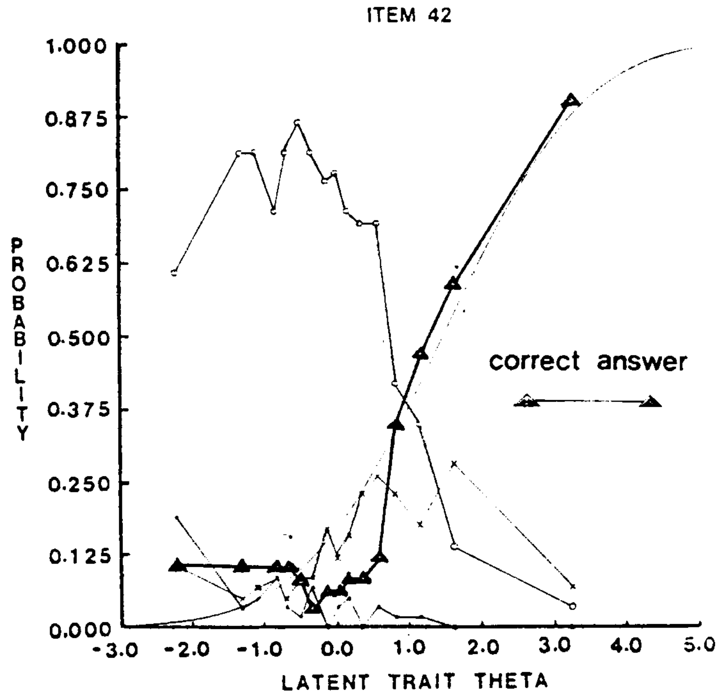


Figure IV.15. Item characteristic curve and plausibility function of Item 42 (Group E): $a_g = 0.98361$; $b_g = 1.29919$. One distractor is informative about subjects with below average ability.

the abilities of below average level. The other items in this group were: 41, 43, 44, 69, 70, 71, 72, 73, 74, 75, and 76.

Group F: Moderately deviate distractor group 2. As Figure IV.16 (Item 45) shows, the plot of the correct answer is only slightly distinctive. The behavior of one distractor still stands out to a certain extent, but the two other distractors also behave very actively. This may indicate the heavy involvement of guessing. The curve of the correct answer deviates from the item characteristic curve through the whole range of ability. The other items were: 46, 47, and 48.

Group G: Very deviate distractor group. As seen in Figure IV.17 (Item 66), all the distractors behave erratically throughout the whole range of ability levels. It seems that guessing is most heavily involved in this group. The plot of the correct answer deviates from the theoretical curve particularly at the higher ability levels. It may be, therefore, very likely that even the subjects at the highest ability level would miss the items in this category. The other items in this group were: 65, 67, and 68.

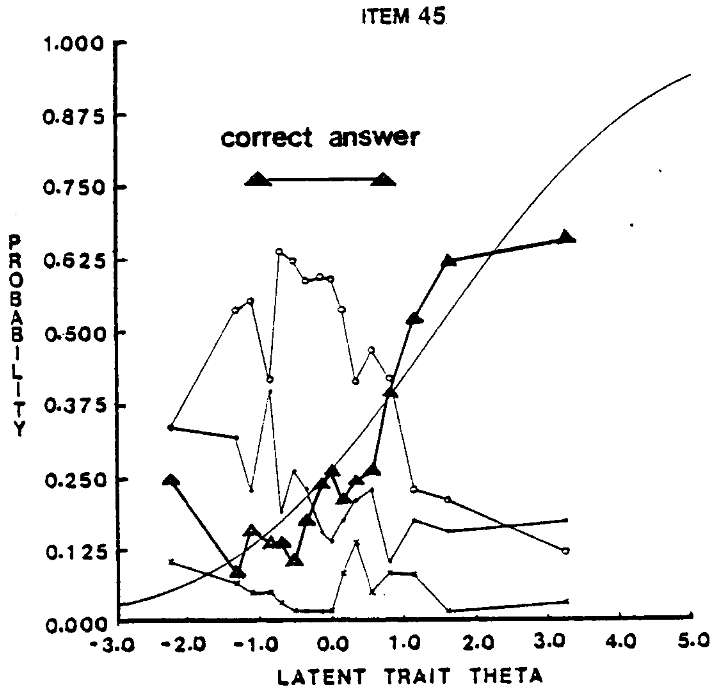


Figure IV.16. Item characteristic curve and plausibility function of Item 45 (Group F): $a_g = 0.47513$; $b_g = 1.20689$. All distractors are moderately deviate.

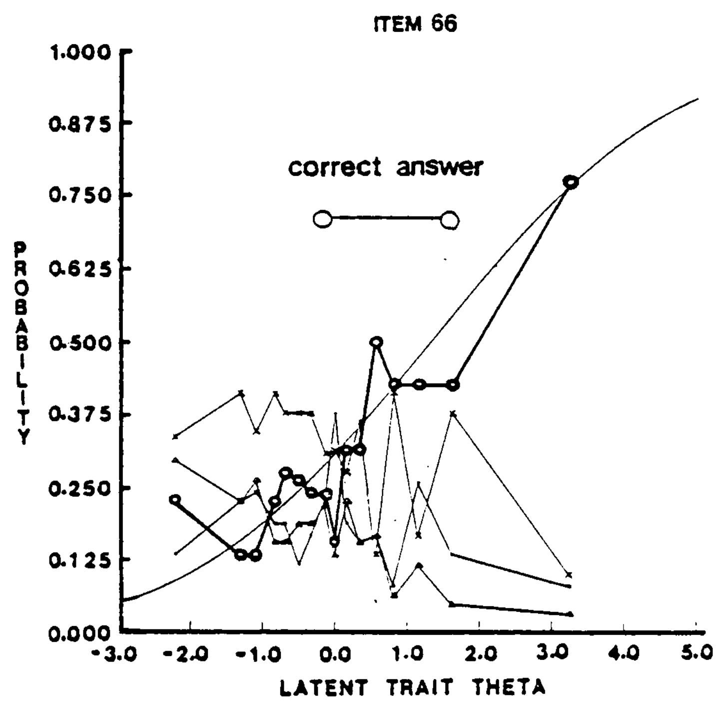


Figure IV.17. Item characteristic curve and plausibility function of Item 66 (Group G): $a_g = 0.38222$; $b_g = 1.36028$. The correct answer and all the distractors are deviate.

Comparison of Latent Trait Theory Methods and Conventional Methods

Frequency distributions of alternatives versus plausibility functions. Comparisons of frequency distributions of alternatives with plausibility functions and item characteristic functions were made to see whether there were any systematical relationships among them. The followings were some findings:

1. Category 1 of frequency distributions of alternatives had a tendency to fall into Group A or in the neighborhood of Group A or B.
2. Category 2 fell into Groups B, C, D, and E of the plausibility function and the item charcteristic function.

   Example 1: Item 55 and Item 75. Even though the distractors of Item 55 shown in Figure IV.18 seem to be as attractive as those of Item 75 shown in Figure IV.19, the distractors of Item 55 were apparently attractive over the wider range of the ability than those of Item 75. Item 55 belongs to Group B, while Item 75 is in Group E.

   Example 2: Item 6 and Item 39. In Item 6, more than one distractor seem to be attractive at the below average ability level (Figure IV.20).

a

b

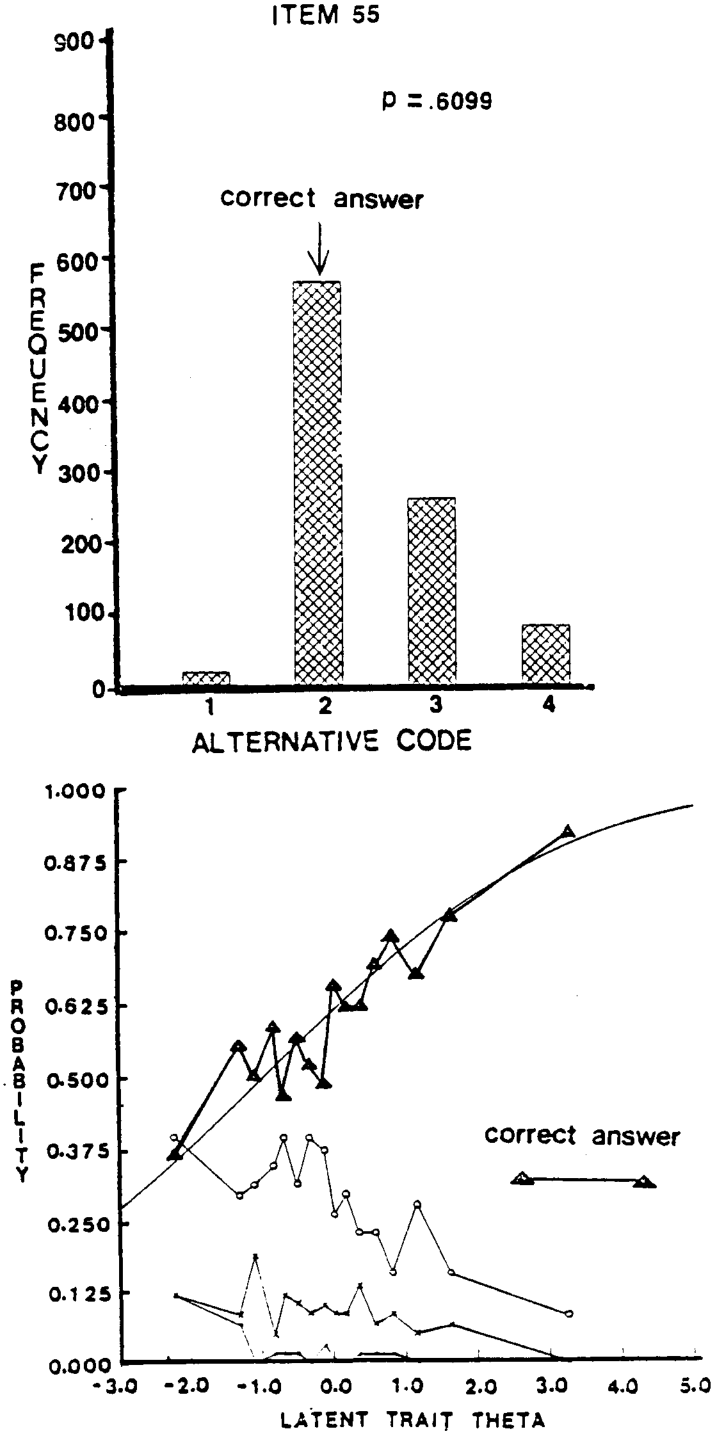


Figure IV.18. Frequency distribution of alternatives (a) and item characteristic curve and plausibility function (b)of Item 55: $a_g = 0.87609$; $b_g = -0.19008$

a

b

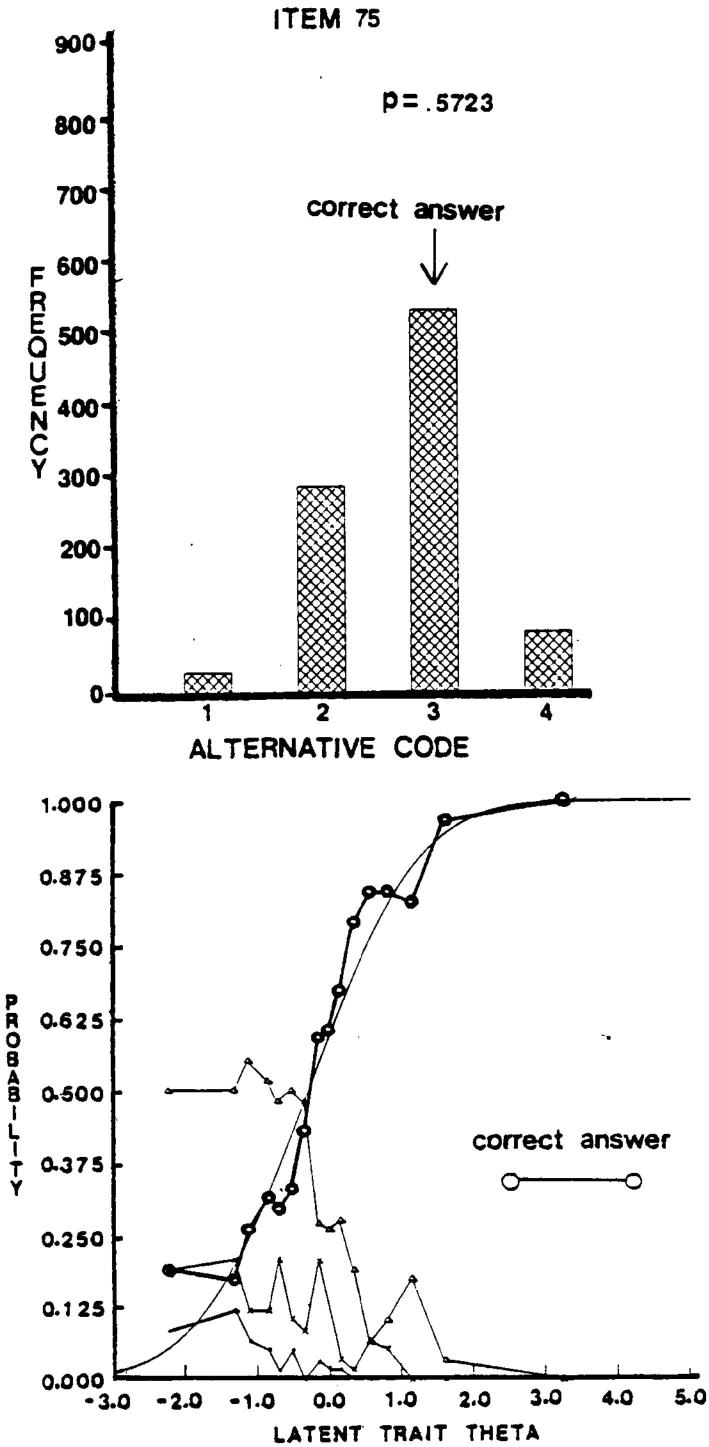


Figure IV.19. Frequency distribution of alternatives (a) and item characteristic curve and plausibility function (b) of Item 75: $a_g = 0.83423$; $b_g = -0.28466$

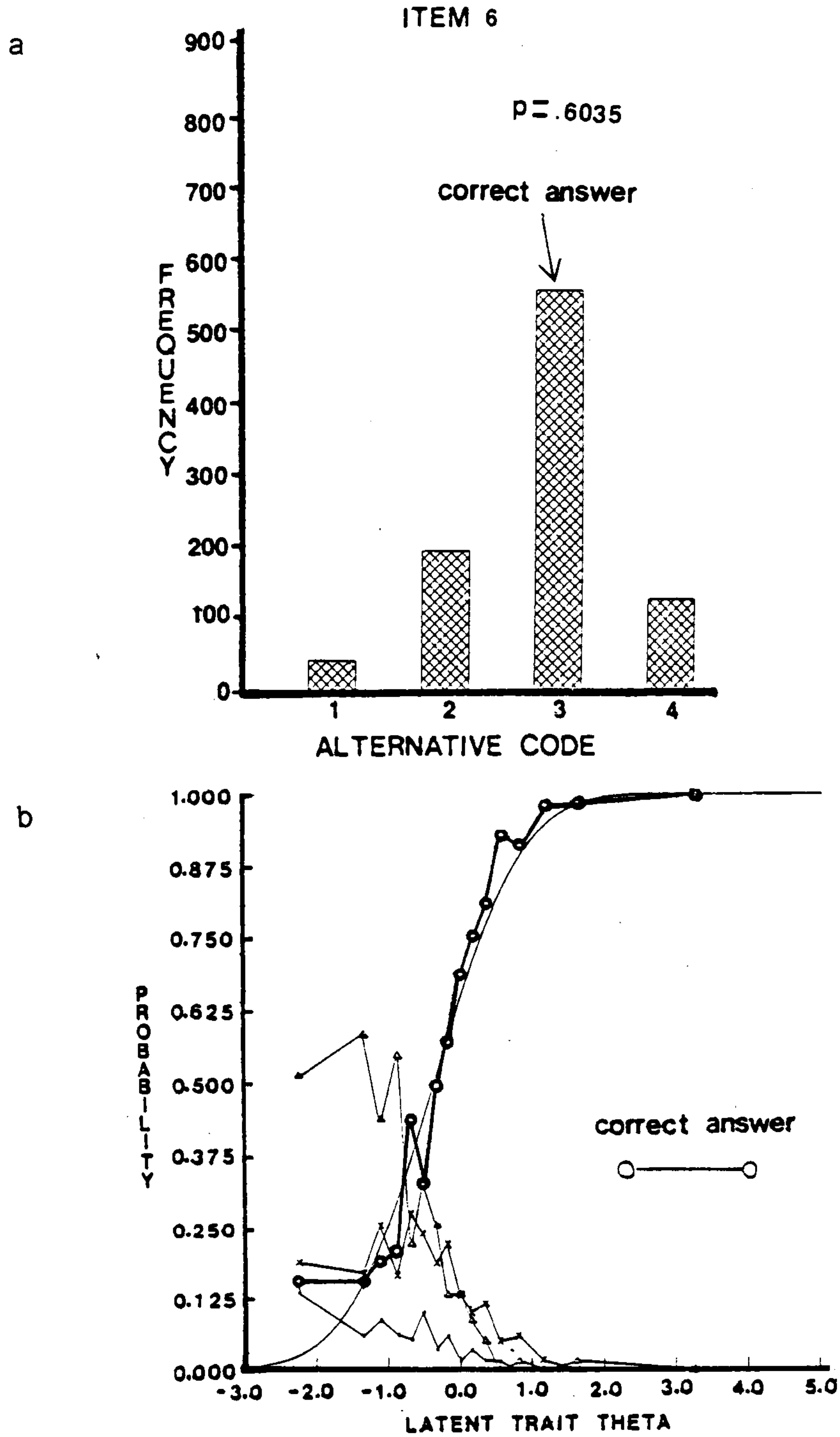


Figure IV.20. Frequency distribution of alternatives (a) and item characteristic curve and plausibility function (b) of Item 6: $a_g = 1.05169$; $b_g = -0.36186$

However, in Item 39, one distractor's behavior stands out through the whole range of ability (Figure IV.21). Item 6 belongs to Group B, while Item 39 is in Group D.

3. Like Category 2, Category 3 contains items that may be included in different types of the plausibility functions and the item characteristic functions (C, G, and E).

   Example: Item 34 and Item 65. As shown in Figure IV.22, Item 34 has one moderately attractive distractor for the below average students, while in Item 65, all distractors are apparently deviate as shown in Figure IV.23. Item 34 belongs to Group C and Item 65 appertains to Group G.

4. Category 4 apparently belongs to Group E.

In summary. Category 1 and 4 seem to fall into certain groups. However, Category 2 and 3 appear to be no systematical belongings of the plausibility function and the item characteristic function.

Distributions of raw test scores and MLE . Figure IV.24 shows the frequency distribution of raw scores to be skewed negatively, while the frequency distribution of MLE is slightly positively skewed (see Figure IV.25).

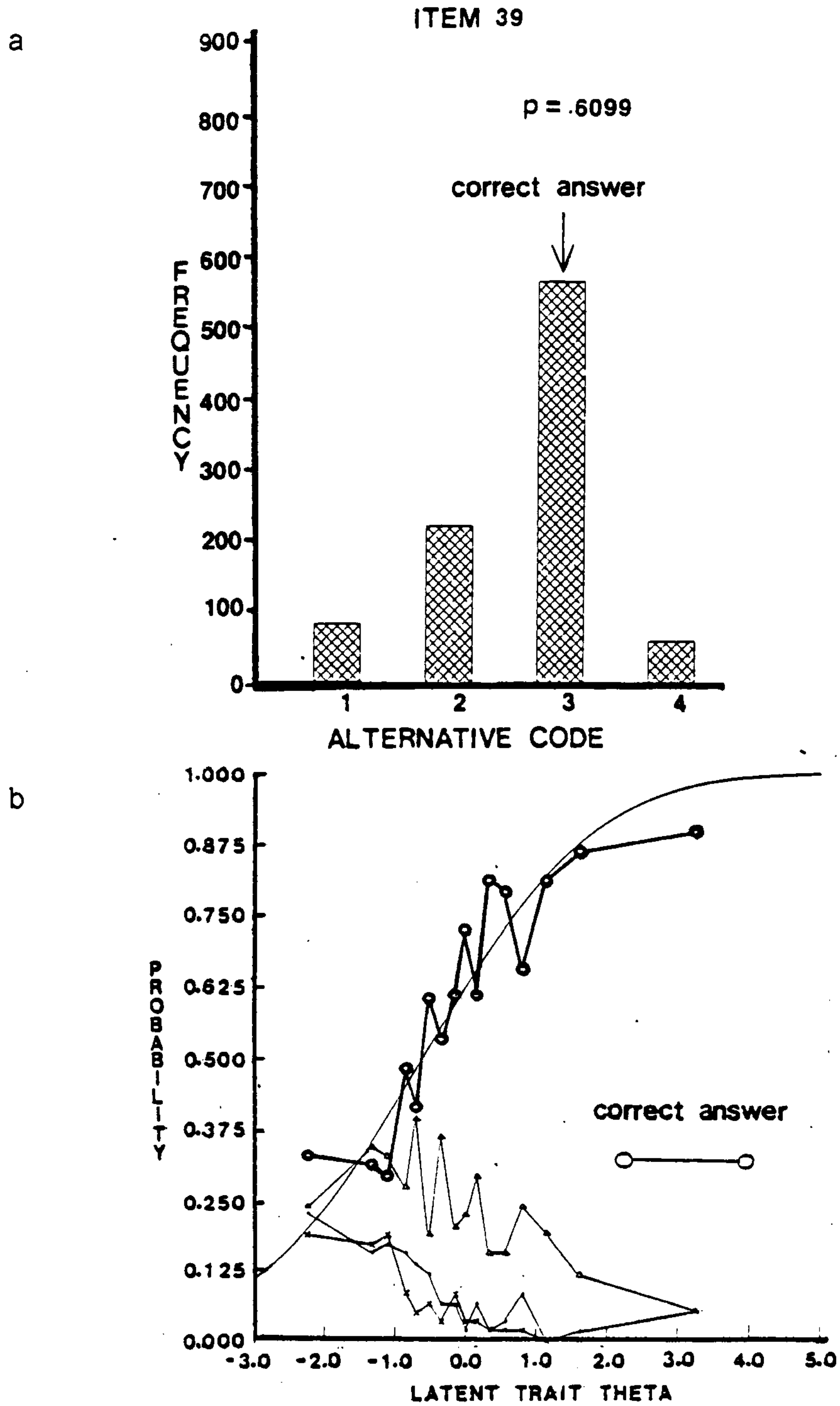
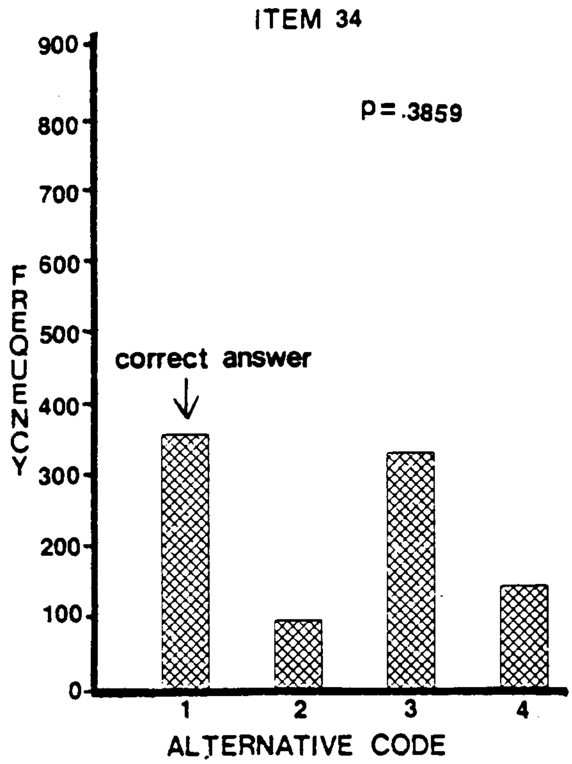


Figure IV.21. Frequency distribution of alternatives (a) and item characteristic curve and plausibility function (b) of Item 39: $a_g = 0.52327$; $b_g = -0.60168$

a



b

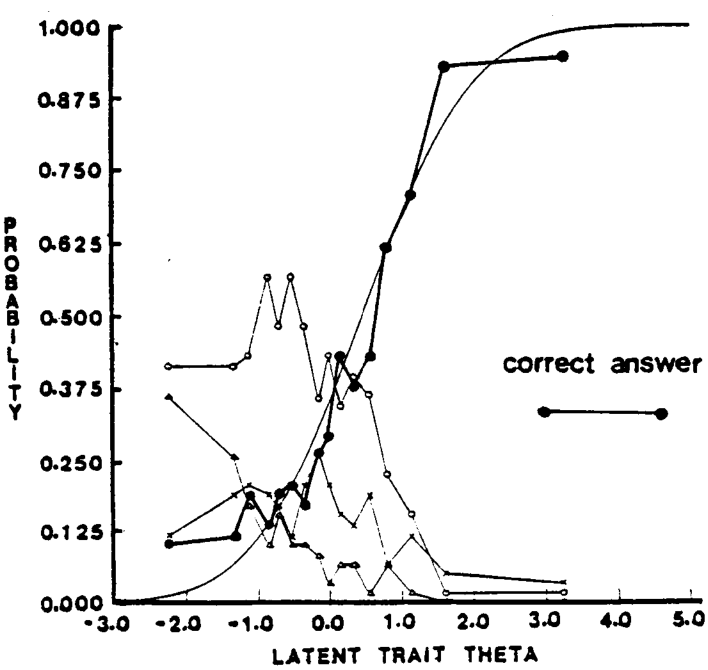


Figure IV.22. Frequency distribution of alternatives (a) and item characteristic curve and plausibility function (b) of Item 34: $a_g = 0.82895$; $b_g = 0.45464$

a

b

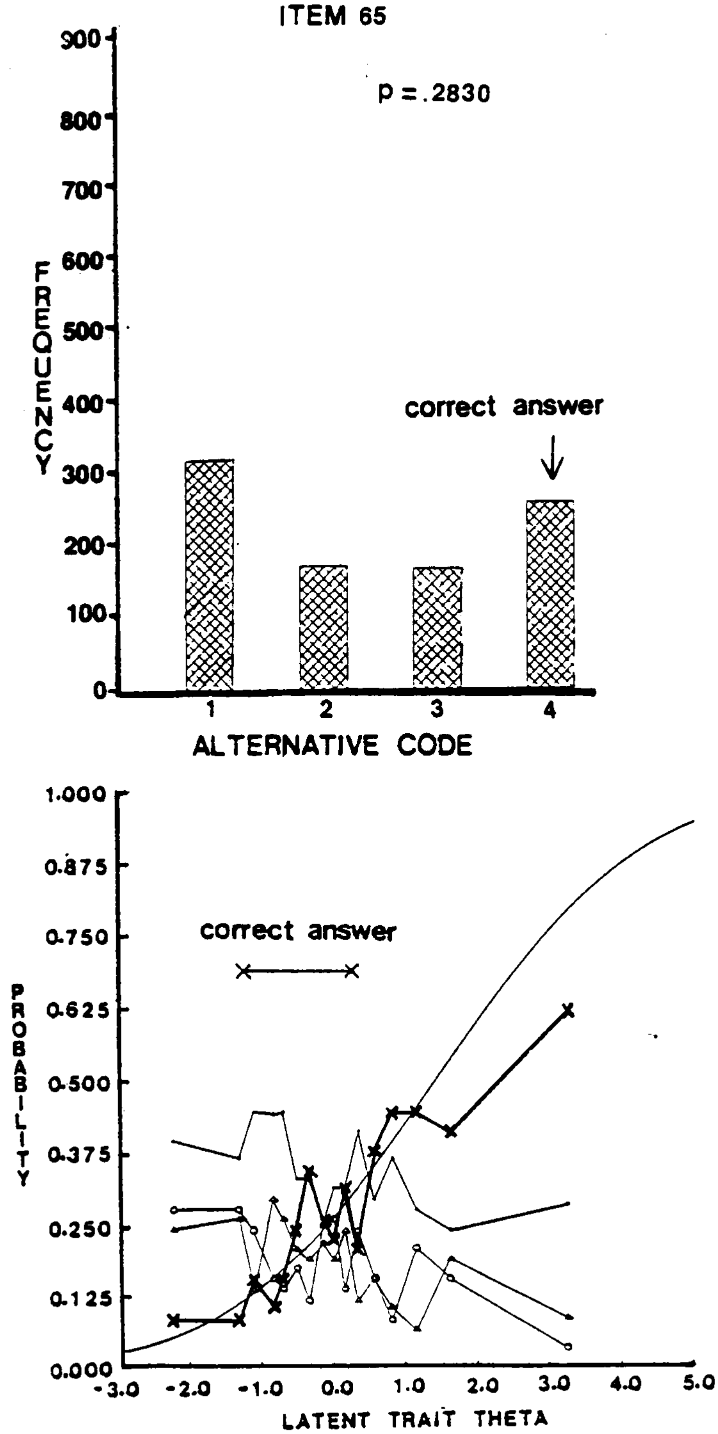


Figure IV.23. Frequency distribution of alternatives (a) and item characteristic curve and plausibility function (b) of Item 65: $a_g = 0.38222$; $b_g = 1.30628$

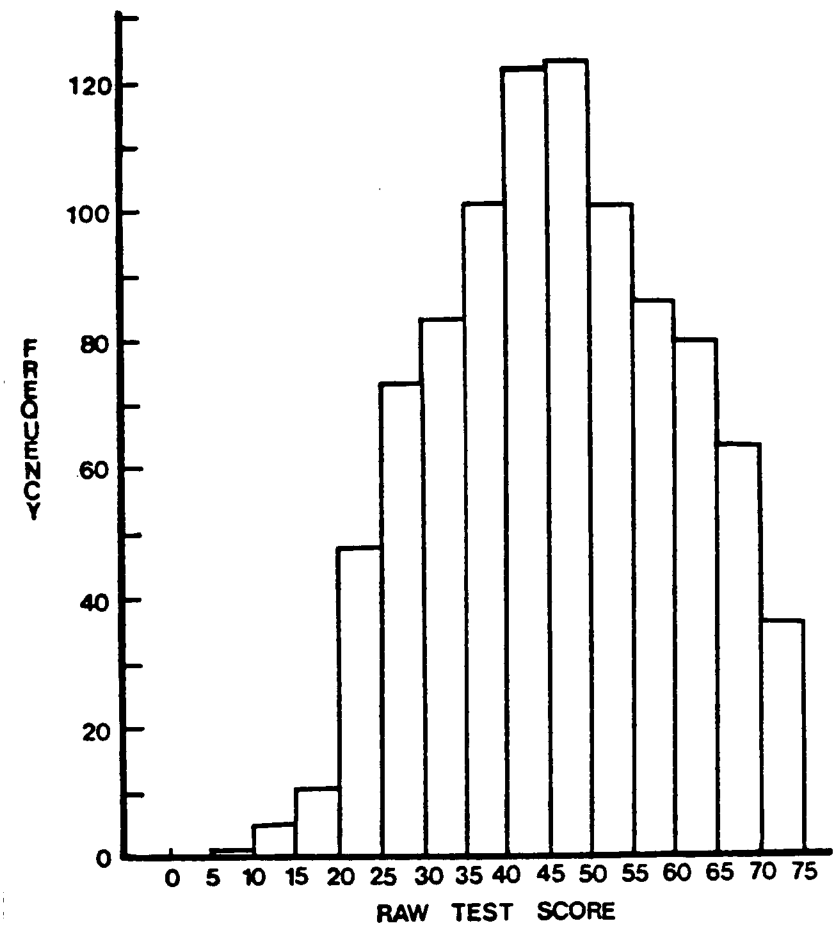


Figure IV.24. Frequency distribution of raw scores (conventional)

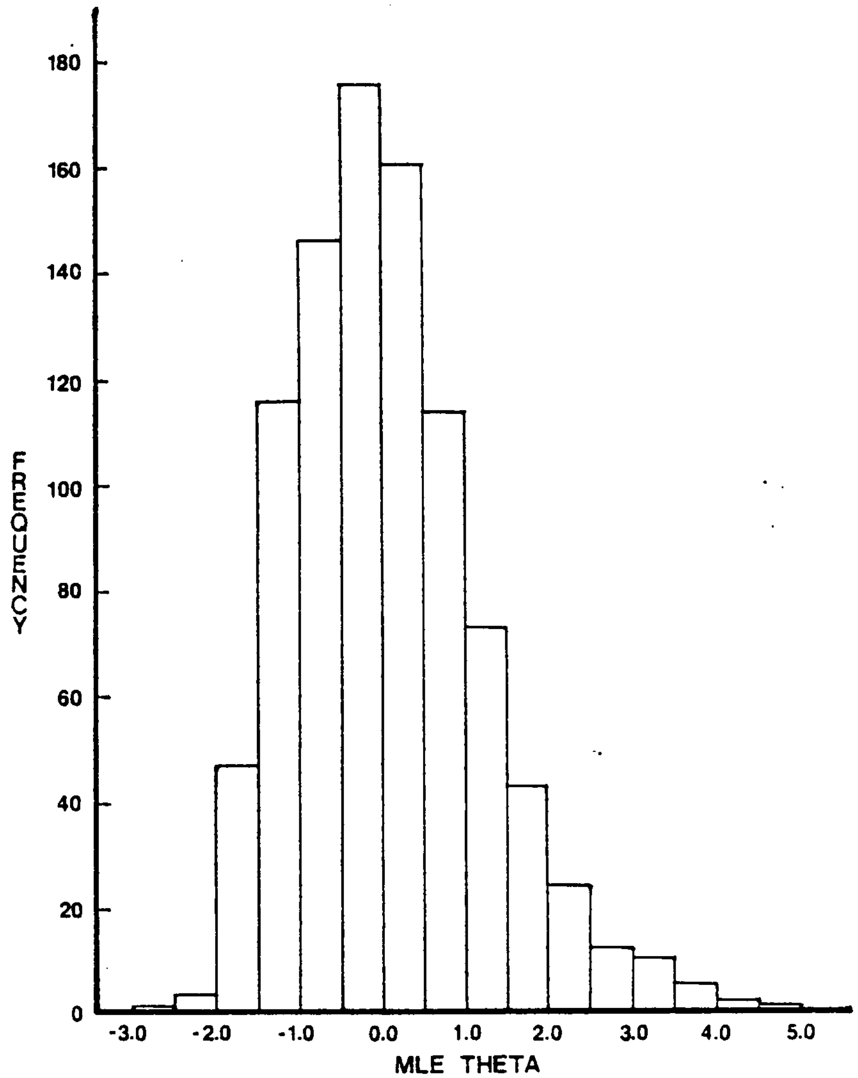


Figure IV.25. Frequency distribution of MLE θ (Latent trait theory)

Relationship between raw test score and MLE θ (Figure IV.26) indicates that the higher the raw score, the wider the range of MLE θ, the more discrepancies one finds between the raw score and MLE θ.

Relationships between difficulty index and discrimination index (conventional), and item difficulty and discrimination parameters (latent trait theory). For the purpose of item selection, it may be helpful to find the relationship between item difficulty and item discrimination (in both conventional and latent trait methodologies) in order to locate specific items with good discrimination power at the selected difficulty levels (Hopkins, 1981, p. 272; Thorndike, 1971, pp. 143-144). When the point biserial correlation coefficient was employed as a discrimination index, the discrimination power was maximized at the difficulty level of 0.5 (Figure IV.27). When the biserial correlation coefficient was employed as discrimination index, it was hard to see any specific relationship between difficulty and discrimination indexes as shown in Figure IV.28. The relationship between the two item parameters showed that the values of the discrimination parameters were at the maximum when its difficulty level was about between -0.5 to 0.1 as shown in Figure IV.29.

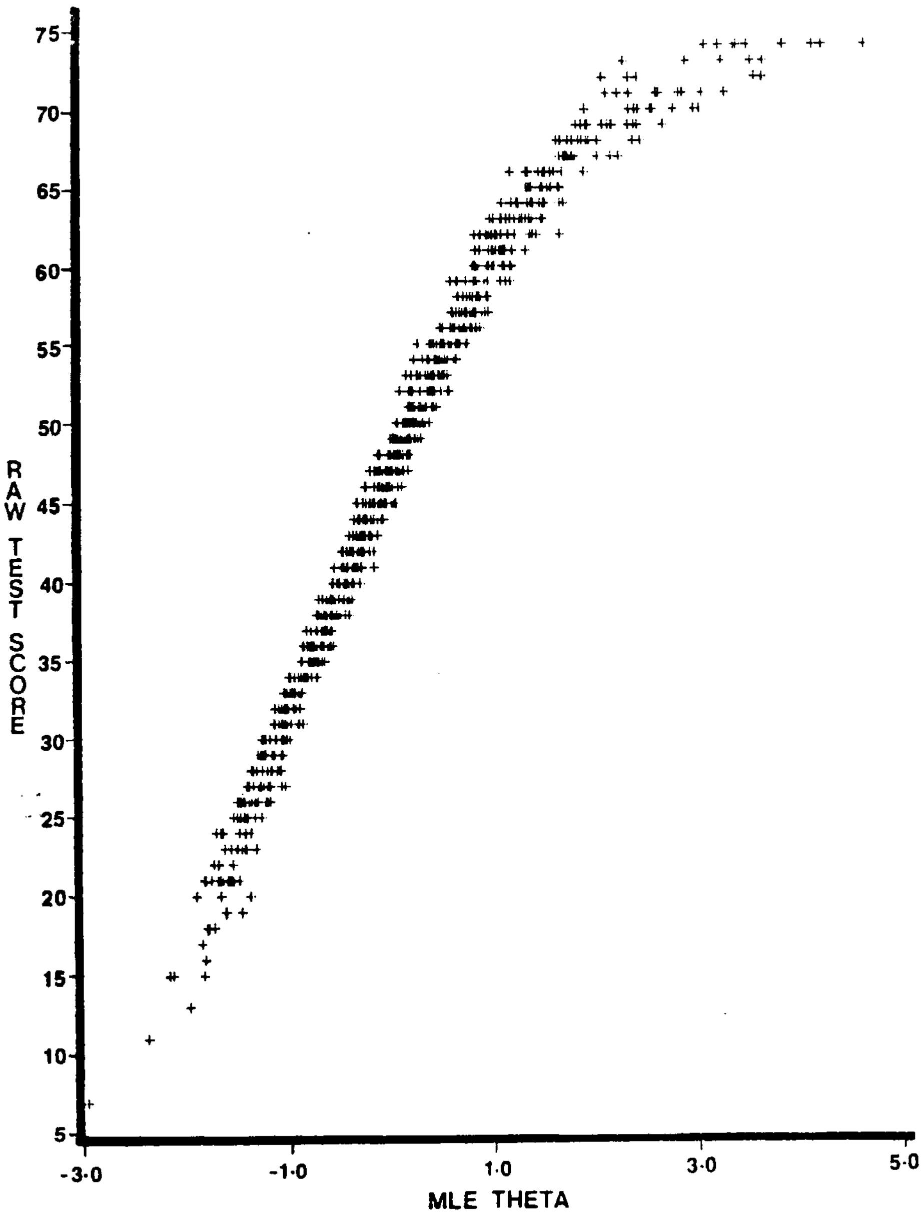


Figure IV.26. Relationship between raw test score and MLE θ.

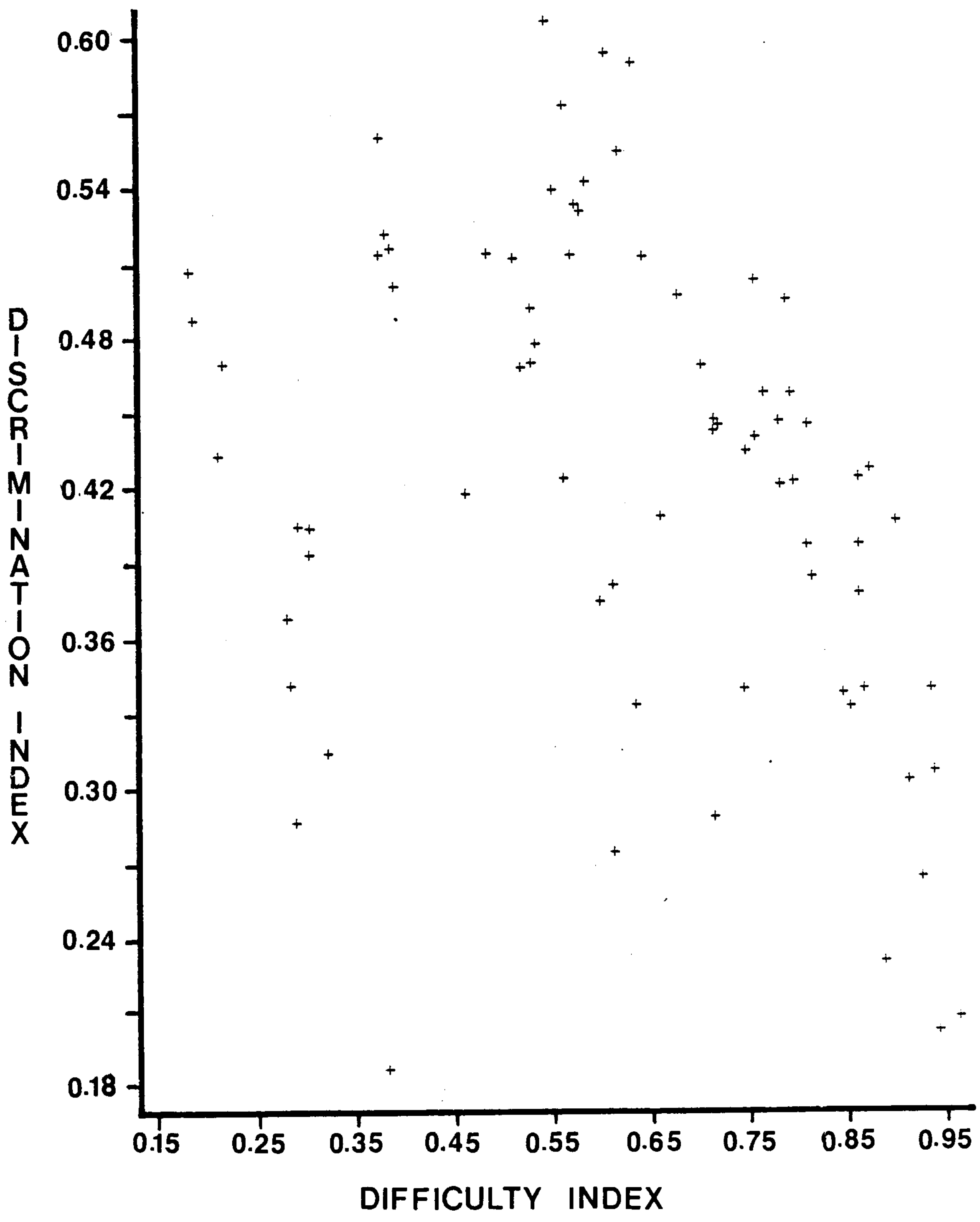


Figure IV.27. Relationship between difficulty index and discrimination index (point biserial, convenational)

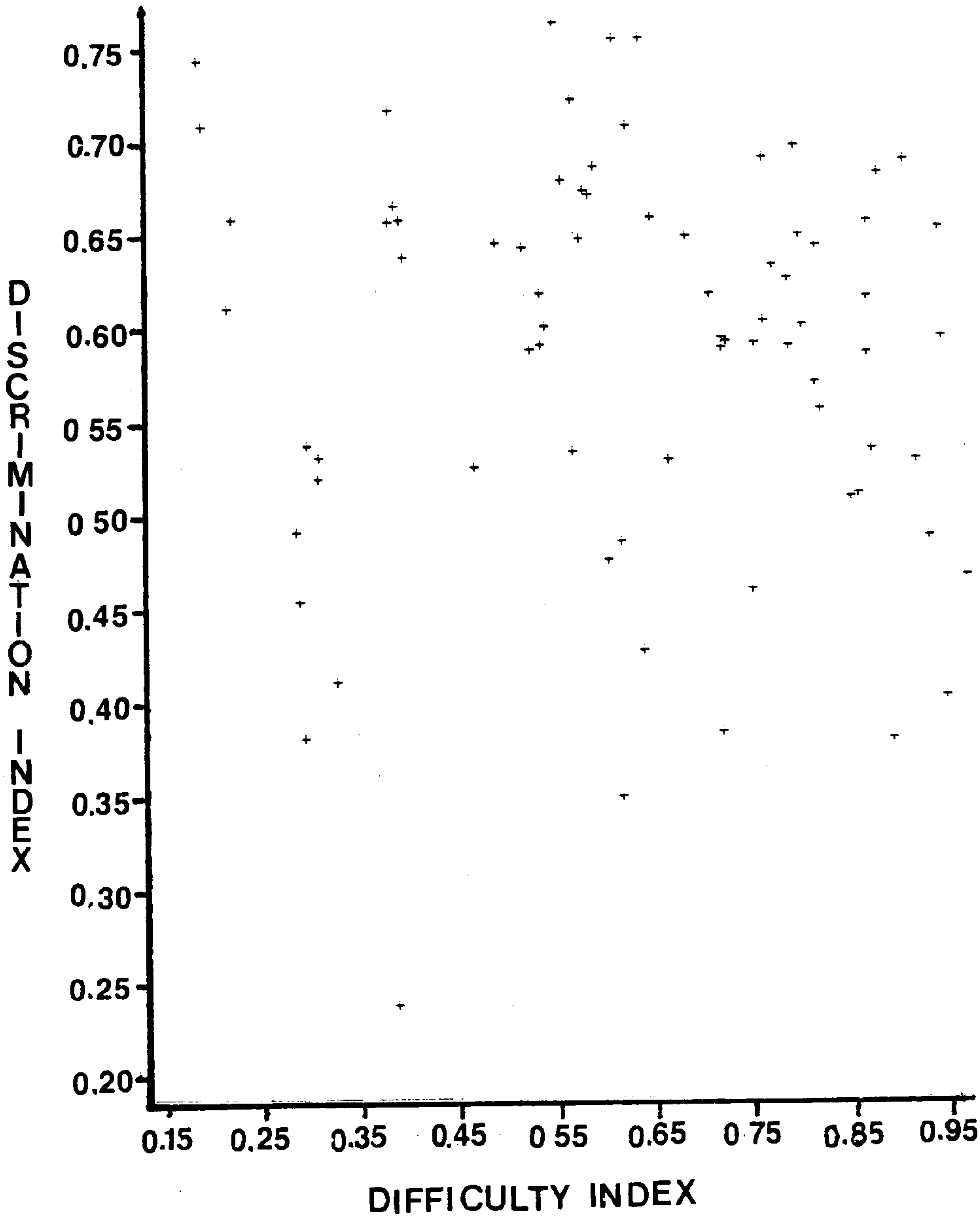


Figure IV.28. Relationship between difficulty index and discrimination index (biserial, conventional)

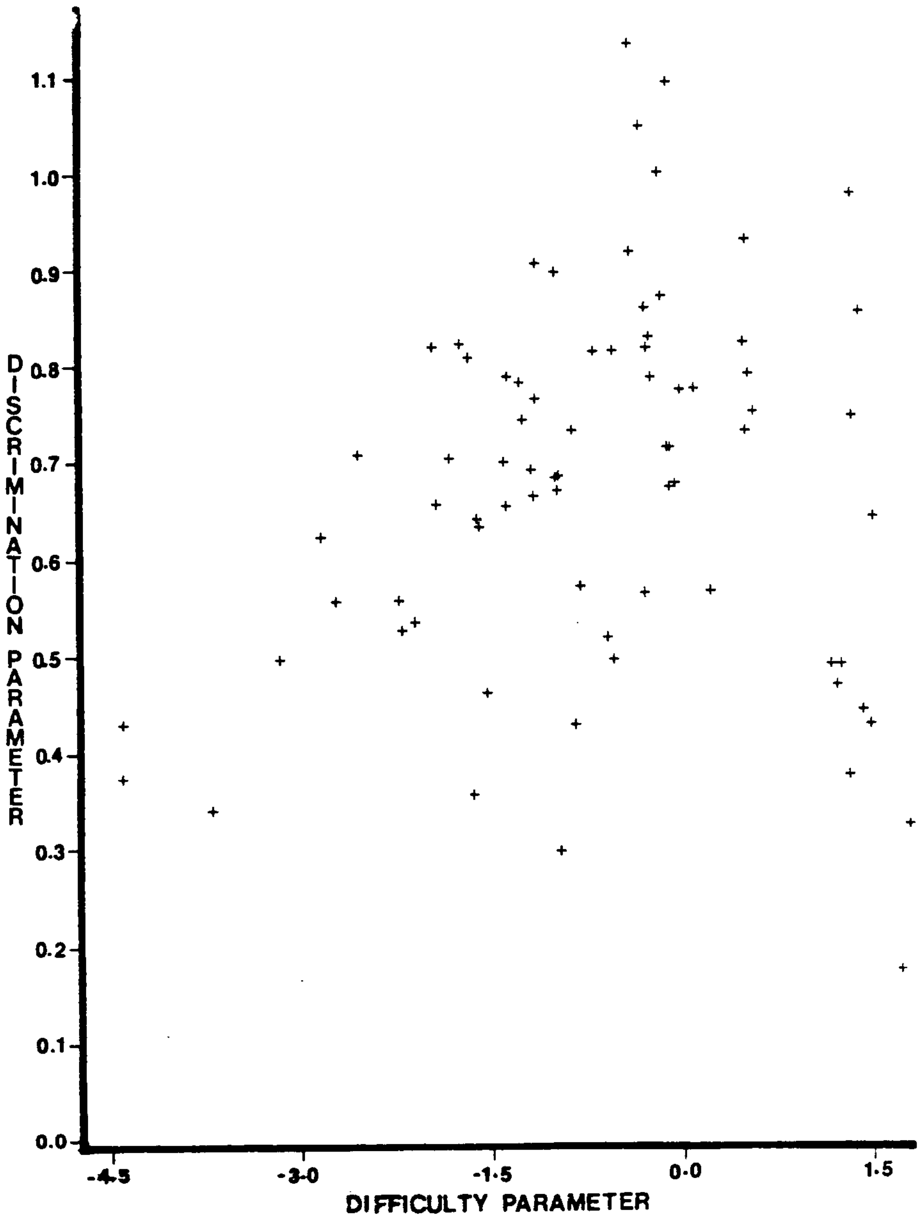


Figure IV.29. Item difficulty vs item discrimination parameters of 75 items

CHAPTER V

CONCLUSION

Some findings of this study are as follows:

1. For very simple arithmetic test items, only the correct answers were informative about subjects' abilities as shown in Figure IV.11 (p. 66). This is what would be predicted.
2. When there was an effective attractive distractor in the set of alternatives the normal ogive model did not fit the item as shown in Figure IV.15 (p. 71). The plausibility function plot of the correct answer (△—△) goes along well with the item characteriustic curve (solid line) from about the latent trait theta, 0.9 to 3.0. It, however, dropped sharply from the curve and formed a U-shape between around the latent trait theta of -0.9 and 0.9. From the latent trait theta -0.9 to -2.0, it is almost a straight line which departs from the item characteristic curve.
3. When items had a proportion correct of less than 0.39, they contained attractive distractors or some terms, the concepts of which might not have been understood by certain examinees. In this study, the normal ogive model did not fit the item.

4. The 76 items, when considered as a test, accurately measured the middle range of student ability, that is, students who scored from 45 to 90 percent correct. This indication is seen in Figure IV.6 through Figure IV.9 (pp. 58-59, 61-62). The highest value of the test information function lay around the theta point of -0.4. Therefore, the standard error of estimate was also the smallest around this point as shown in Figure IV.7. Items with higher values of the information function were also found around this point as shown in Figure IV.8. Moreover, Figure IV.9 indicates the smallest difference between theta (θ) and estimated theta ($\hat{\theta}$) around this point. If the purpose of the test is to pass or fail students on the basis of a pre-determined cut score, it would be important to build the test so that it could most accurately measure ability at that score. Although this was a field test study, and sixth- and seventh-grade students were used, it can be used illustratively. From Figure IV.26, this ability level is represented by a raw score of approximately 45, which is a percent-correct

score of 60. A cut score of about 60 would be ideal because that score would discriminate maximally among those scoring above and below that point.

5. Most students' abilities lay between -1.0 and 1.5 of MLE theta (see Figure IV.25), where the test information function was maximized (see Figure IV.6, p. 58). Within this range, the higher the MLE theta goes, the more discrepancy between raw test score and MLE theta exists. This indicates that small differences among high scores can be detected in detail using MLE theta rather than raw scores. This indication may be due to the ceiling effect because the test was not for the students with high abilities. The field test items were expected to measure a minimum competency. Therefore, students with a wide range of abilities would get high raw scores. If the test had been built to discriminate among abilities, the distribution of raw scores and MLE perhaps would be more similar.
6. The comparison of the frequency distribution of raw test scores (Figure IV.24) to MLE theta (Figure IV.25) also indicated that MLE theta showed small differences for the higher estimated

abilities (between estimated abilities 1.5 and 5.0) than the raw test scores.

7. Factor analyses indicated that there were more than one dimension in many items in this set. In this study, however, all items were treated as unidimensional items.

8. When the categorization of the frequency distribution of alternatives was used to find out some systematic relationship to the different groups of the plausibility function and the item characteristic function, the following findings were obtained: (a) Category 1 and 4 fell into Group A and E. respectively; but (b) Category 2 and 3 fell into several different groups, even though the configurations of the frequency distributions of alternatives were similar within each group. If the findings of this study regarding the relationship between the frequency distribution of alternatives and the plausibility functions were confirmed by using different content areas, such as language arts, science, and so on, and different population groups, the frequency distribution could be used to preselect items which should be further analyzed using plausibility functions.

9. The relationship between item discrimination and difficulty parameters in Figure IV.29 indicates that the discrimination parameter is the highest at the difficulty parameter of around 0.3. Also, the relationship between point biserial discrimination index and p (proportion correct) in Figure IV.27 provided similar information: It was maximized at around 0.5 of p (proportion correct). However, the relationship between biserial correlation discrimination index and p in Figure IV.28 did not show any particular relations.

10. If the test constructor uses the normal ogive model to determine whether the items are good or bad, items belonging to Group A through Group D may be recommended because the normal ogive model seems to fit those items. However, if the test constructor uses some method to estimate the student's ability using information from the distractor as well as information from the correct answer, items belonging to Group E may also be good. If the technique of estimating the student's ability from the distractor is beyond the test constructor's situation, he/she may want to revise specific attractive distractors so that one distractor may not be particularly

attractive, and then the normal ogive model may fit the revised items in Group E.

11. Items belonging to Group F and G may be deleted from the set of items for the following reasons: (a) Low discrimination parameters of these items indicate discriminating the wide range of abilities. Since the purpose of a minimum competency test is to discriminate students around the cutting score, for example, 60% of correct answers, items belonging to these groups may not serve the purpose of the test administration effectively; (b) Neither did the normal ogive model fit them, nor did the items seem to have any apparent attractive distractors.

12. It was said that when all distractors were frequently chosen, this set of distractors is acceptable (Thorndike, 1971, p. 138). This is the case of Category 3 of the frequency distribution in this study. The results showed items in this cateory belong to different groups of plausibility functions and item characteristic functions in this category. Some of these items were not considered to be good items.

Educators and researchers need to be aware of the implications and refinements of educational testing made possible by latent trait theory. The test constructor needs to be aware of the need for consistency of discrimination parameters and difficulty parameters year after year, as in state proficiency tests. Some of those proficiency tests are used as one indicator of eligibility to graduate from high school. If item parameters differ greatly from year to year; students who passed one form of the test one year might not have passed another form of the test in another year, due to varying difficulty parameters unless some equating method is employed. If the constructor knows how different types of items will produce different parameter values, he/she can make a proper selection of items more easily. Innes (Personal communication, 1984) said:

> Traditional methods can accomplish the yearly production of equivalent tests; however, latent trait methods seem to provide more capacity to fine tune a test through the selection of items with certain parameter values. Furthermore, the effect of distractors can be made more visible by latent trait methodology, so that the production of various kinds of tests can be enhanced.

Further investigations may be recommended regarding the following points:

1. The dimensionality of this test may be able to be further analyzed so that the first eigenvalues may be higher than they were in this study.
2. An elaboration of the type of analysis introduced in this study could involve the effects of

instructional or cultural variance on item parameters. Harnisch (1983) mentioned that the effect of large school-to-school variation on different objectives and on the different distractors need to be considered.

3. It may be interesting to perform distractor analysis using raw test scores instead of MLE θ and then to compare those two distractor analyses.
4. Test items of different content areas, or of the same content area with different difficulty levels may be studied in the same way as in this study to confirm the results of this study.

The ultimate goal of the present study is to establish some kind of tools to distinguish good items which will estimate the examinee's ability accurately. However, this is a preliminary study toward that goal. Many studies from different angles as mentioned above need to be done before making the definite conclusions.

REFERENCES

REFERENCES

Anastasi. A. (1976). Psychological testing (4th Ed.). New York: MacMillian.

Bejar, Issac I. (1983). Introduction to item response models and their assumption. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 1-19).

Benson. J., Croker, L. and Ware, W. (1982). A comparison of three types of item analysis in test development using classical and latent trait methods. Educational and Psychological Research, 2, 199-212.

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. 397-472). Reading, Mass: Addison-Wesley.

Carter, J. E. (1980). A test of the adequacy of the Rasch model in the detection of unidimensionality in achievement test data. Unpublished master's thesis. University of Illinois at Chicago Circle.

Glass. U. G. (1966). Note on rank biserial correlation. Educational and Psychological Measurement, 26. 623-631.

Hambleton, R. K. and Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Harman. H. H. (1967). Modern factor analysis. Chicago: The University of Chicago Press.

Harnisch. D. L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Helwig, J. T. and Council, K. A. (Eds.). (1979). SAS user's guide (1979 ed.). Cary, NC: SAS Institute Inc.

Hopkins. K. D. and Stanley. J. C. (1981). Educational and psychological measurement and evaluation. Englewood Cliffs. NJ: Prentice-Hall.

Hulin. C. L., Drasgow. F., and Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Lazarsfeld, P. F. and Henry. N. W. (1968). Latent structure analysis. Reading, MA: Addison-Wesley.

Lindsey. W. W. and Douglas, S. K. (1981). SAS/GRAPH user's guide (1981 ed.). Cary, NC: SAS Institute Inc.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7.

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. (1980). Application of item response theory to practical testing problems. Hillsdale. NJ: Erlbaum.

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

Nie. N. H., Hull, C. H., Jenkins, J.G., Steinbrenner, K., and Bent, D. H. (1975). Statistical package for the social sciences (2nd ed.). New YOrk: McGraw-Hill.

Rasch. G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, No. 17.

Samejima, F. (1972) A general model for free-response data. Psychometric Monograph, No. 18.

Samejima, F. (1973). Homogeneous case of the continuous response level. Psychometrika, 38, 203-219.

Samejima, F. (1977a). Effects of individual optimization in setting the boundaries of dichotomous items on accuracy of estimation. Applied Psychological Measurement, 1, 77-94.

Samejima, F. (1977b) A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

Samejima, F. (1977c). Weakly parallel tests in latent trait theory with some criticisms on classical test theory. Psychometrika, 42, 193-198.

Samejima, F. (1979). Constant information model: A new promising item characteristic function. Knoxville. TN: University of Tennessee, Department of Psychology.

Samejima, F. (1980). MLE theta: FSPC-80-7, FSCT-80-30 [Computer program]. Knoxville, TN: Author.

Samejima, F. (1980). Tau transformation: FSCT-80-26, FSPL-78-63, FSPL-79-21, FSCT-80-24 [Computer program]. Knoxville, TN: Author.

Samejima, F. (1981). Final Report: Efficient methods of estimating the operating characteristics of item reponse categories and challenge to a new model for the multiple-choice item. Knoxville, TN: Univeristy of Tennessee, Department of Psychology.

Samejima, F. (1983). Operating characteristics in normal ogive model with and without random guessing: FSPC-83-1 [Computer program]. Knoxville, TN: Author.

Samejima, F. and MacCarter, C. (1981). Item parameter computation from proportion correct and correlation between item variable and theta on normal ogive model on dichotomous response level: FSCM-81-6 [Computer program]. Knoxville, TN: Primary author.

Samejima, F. and MacCarter, C. (1982). Polycholic correlation coefficients computed from response patterns or contingency tables: FSCM-82-8 [Computer program]. Knoxville, TN: Primary author.

The State Department of Education. (1981). Manual of directions: Field test. Nashville, TN.

The State Department of Education. (1982). Objectives for the Tennessee Proficiency Test. Nashville, TN.

The State Department of Education. (1982). You and the proficiency test. Nashville, TN.

Swaminathan, H. and Gifford, J. A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-191.

Thorndike. R. L. (1971). Educational measurement (2nd Ed.). Washington, D. C.: American Council of Education.

Yen. C. C. (1980). Cal-comp graph drawing of points, lines and/or curves: FSPC-80-18 [Computer program]. Knoxville, TN.

APPENDICES

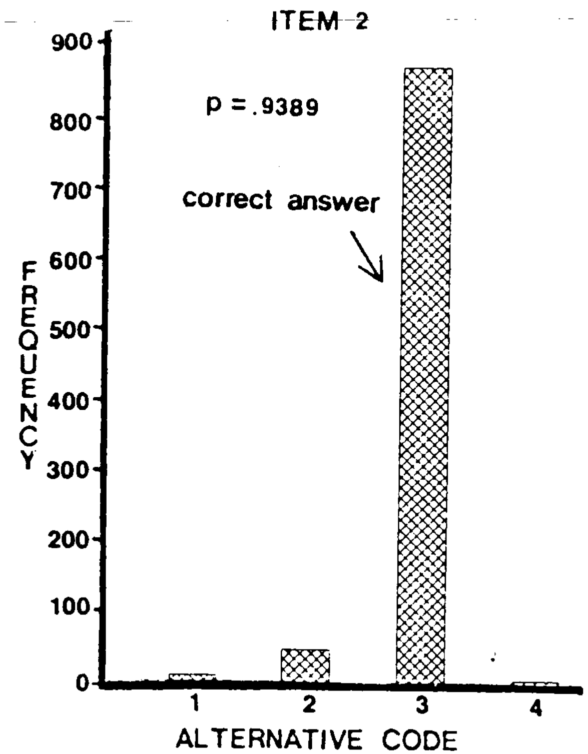
APPENDIX A

Response Patterns of Students Who Responded
Fewer than 87% of Items

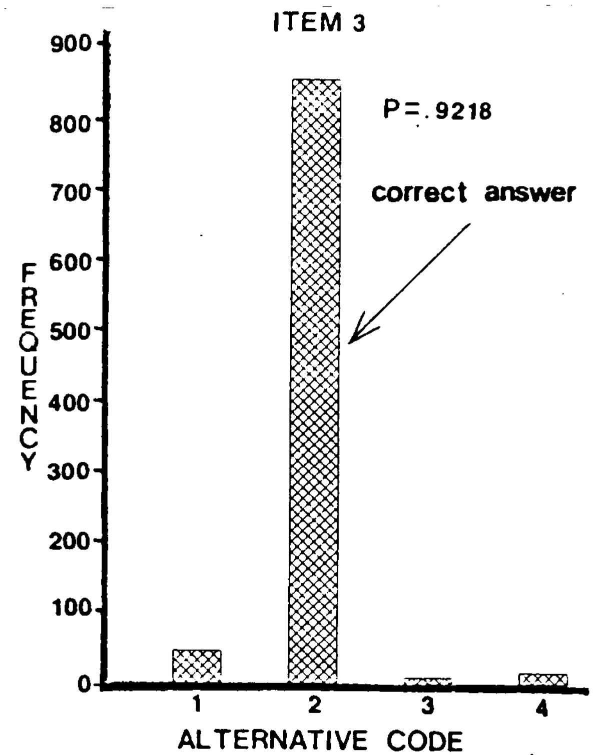
```
             1         2         3         4         5         6         7
    12345678901234567890123456789012345678901234567890123456789012345678901234567890123456
001 KLKLJLMKKKLMMMJMLLMMKLJMKJKMMMJLMMMKK
002 KLKKJKLKM                                              J                   M
003 KLKLJLMKKLLMMMJMLLMMKLJJKJKMKMJL
004 KLKLJKKLKJLKMMJMLMMMMLJMKJKMMMJLLLJJKKJJ
005 KLKLMLLJMJLMMMJMLLMMKLJMKJKMMMJLJJKJKKLKLLKJLLLJKJMKML               JJMKK
006 KLKLJLMKKKKMMMJMLLMMKLJMKJKMMMJLLLLJLKLKKMKJL
007 KLKLJLMKKKLMMMJMLLMMKLJMKJKMMMJL    KKLM
008 KLKLMLKLMJLL
009 KLKLKKLJJLMKMMJMLLM
```

```
             1         2         3         4         5         6         7
    1234567890123456789012345678901234567890123456789012345678901234567890123456
010 KLKMJKMLLKKMMMJMLLKMLLJLMKLKLKLKMKLKJKJL    KMJLKKMKKKLJMKKKMMLJ          JL
011 KLKLJLMKKKLMMMJMLLMMKLJ
012 KLKLKKKMJKJKMLJMLLMMJJL KJKMJLKKMMML   MLLKJLKKKJJKJLLKMMMJKKJKLKM
013 KLKLJLMJKKLMMMJMLLMMKJJLLKJLMMJLKLMKKKLKLLKJLLLJKKMKLLLJMK
014 KLKLMKLMLJLJMMJMLLMMLJKKKKJJMMLKMMJM
015 KLKLJKMKMMKJMJJMLLMMKLJM
016 KLKLJKMKJJKLMMJMLLMMKLJMKJKMLLJLLJKJKKLKLLKJJKKKKKMKMLKJMKKKMMJL
017   JJJKJ      J
```

APPENDIX B

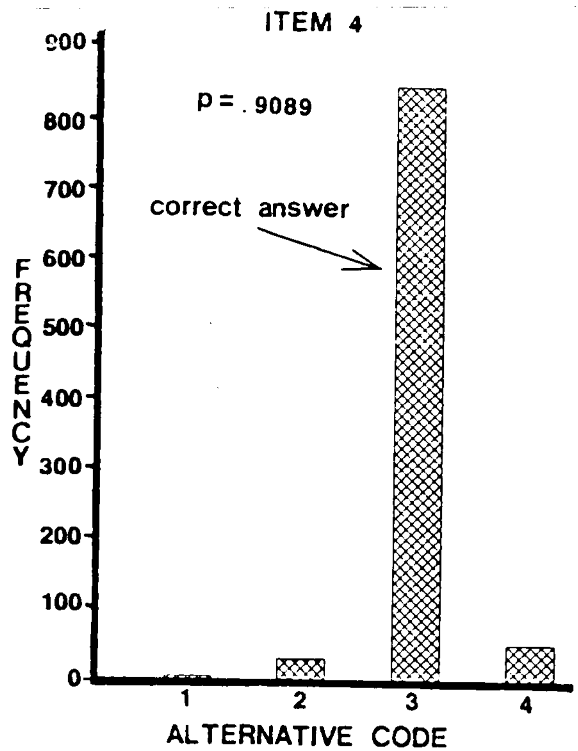


(1) Item 2

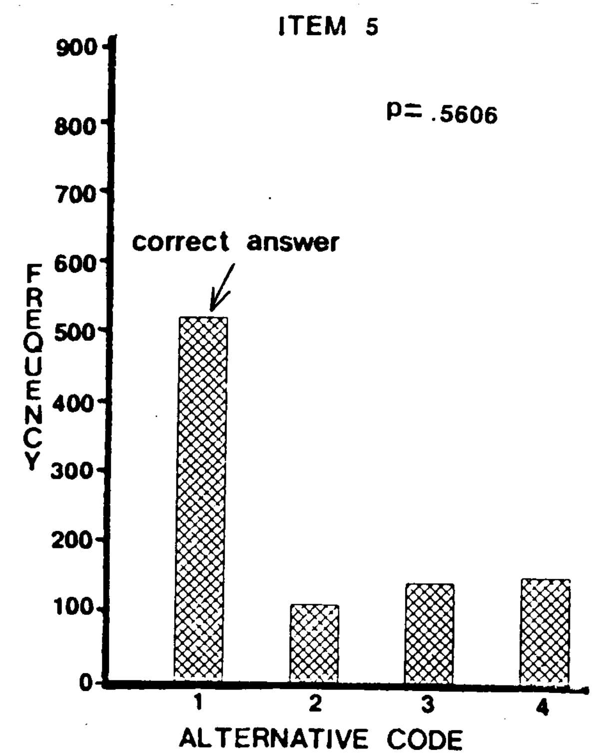


(2) Item 3

Figure B.1. Frequency distributions of alternatives of test items.

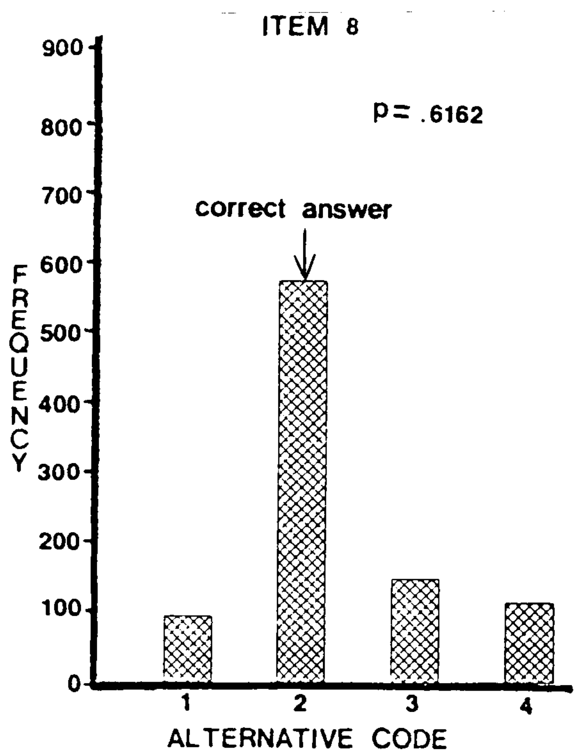


(3) Item 4

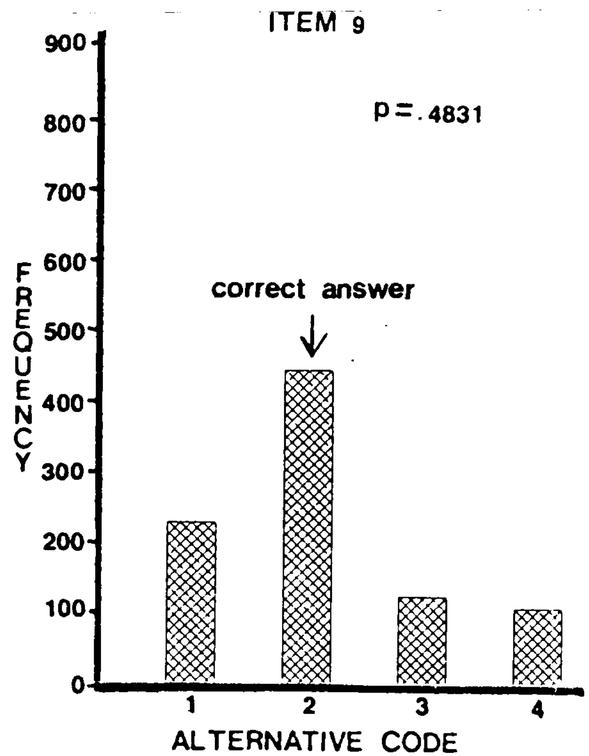


(4) Item 5

Figure B.1. (Continued)

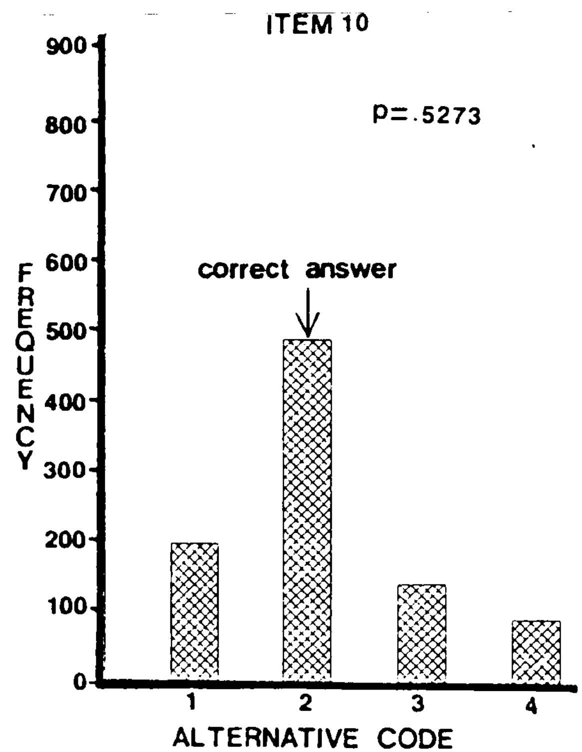


(5) Item 8

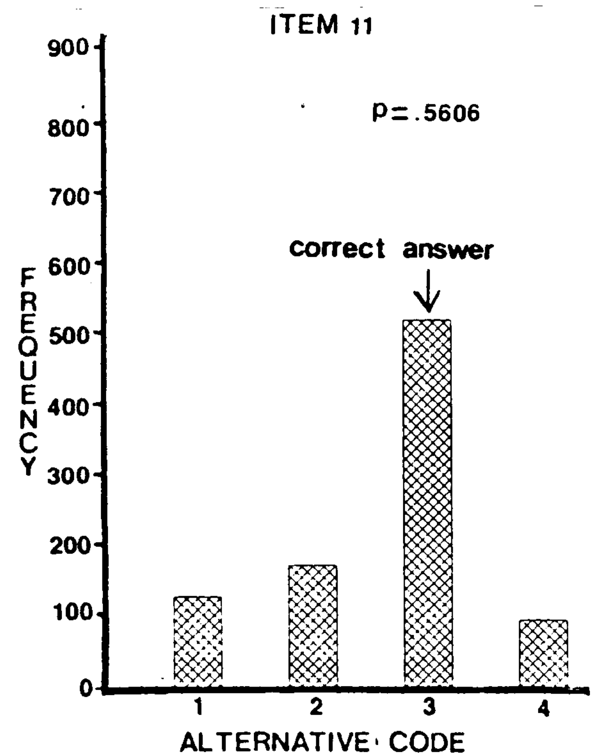


(6) Item 9

Figure B.1. (Continued)

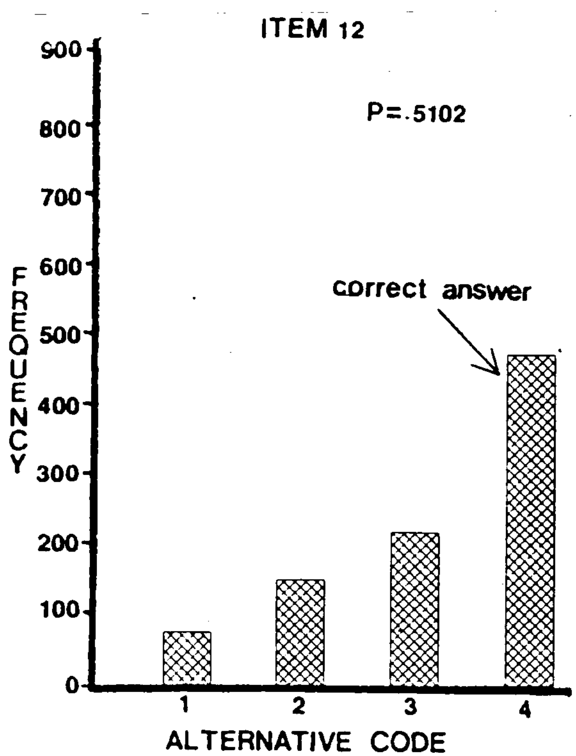


(7) Item 10

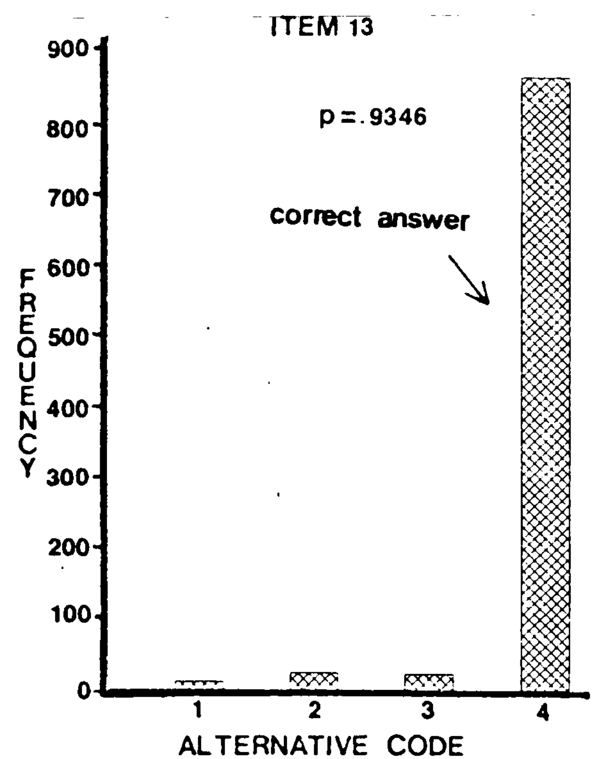


(8) Item 11

Figure B.1. (Continued)

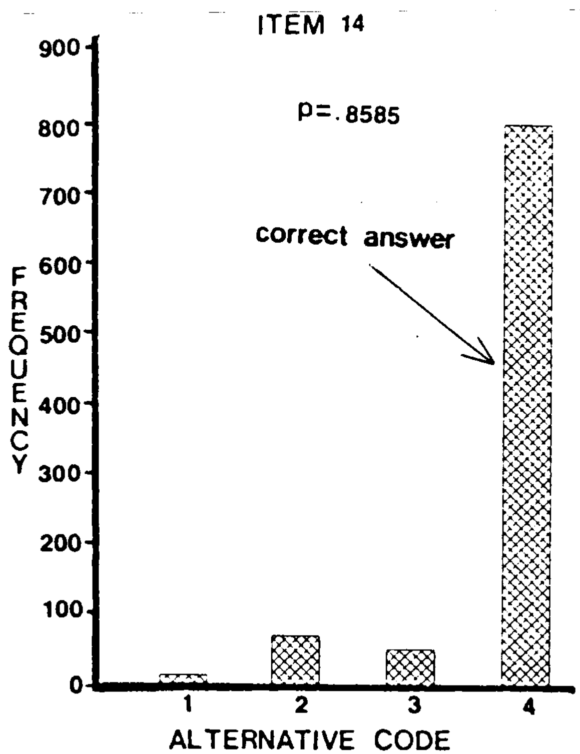


(9) Item 12

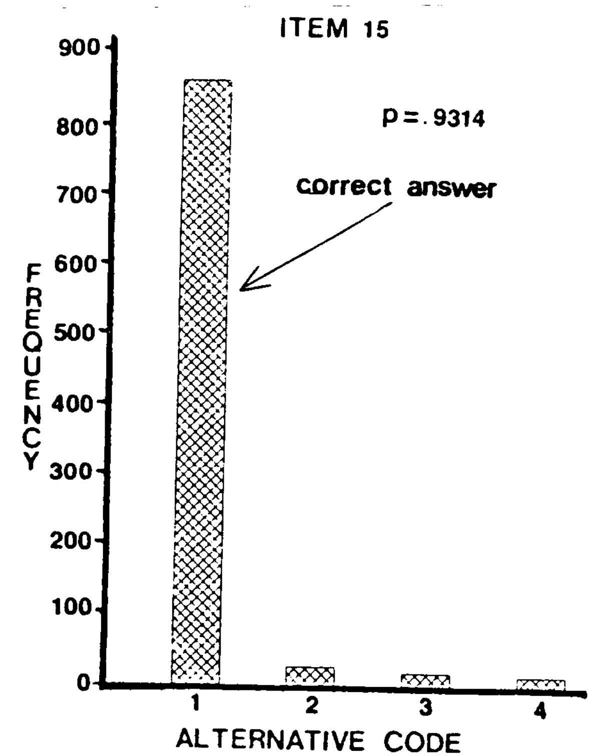


(10) Item 13

Figure B.1. (Continued)

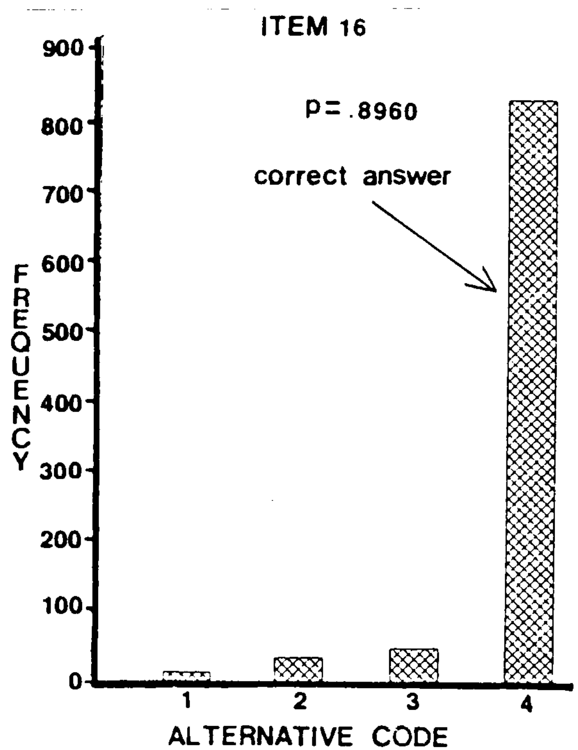


(11) Item 14

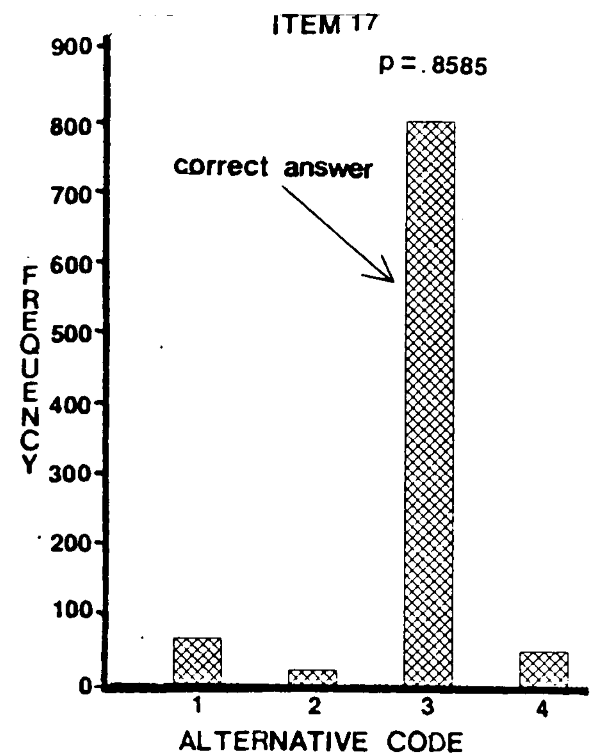


(12) Item 15

Figure B.1. (Continued)

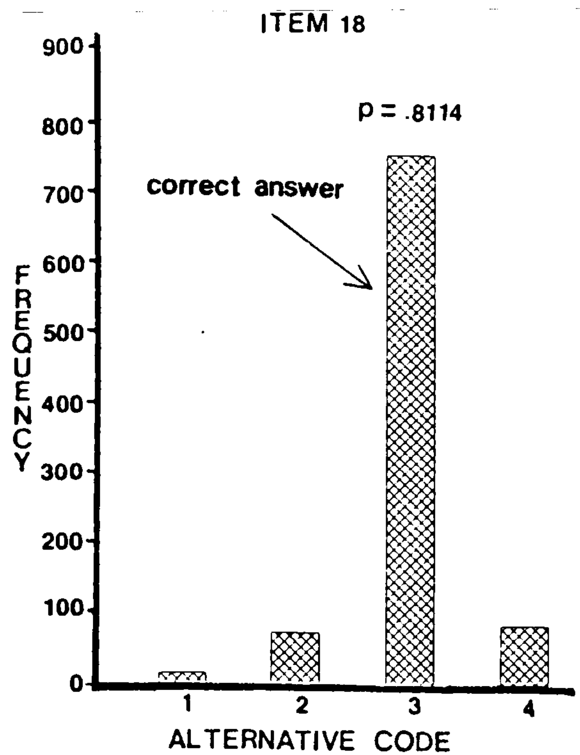


(13) Item 16

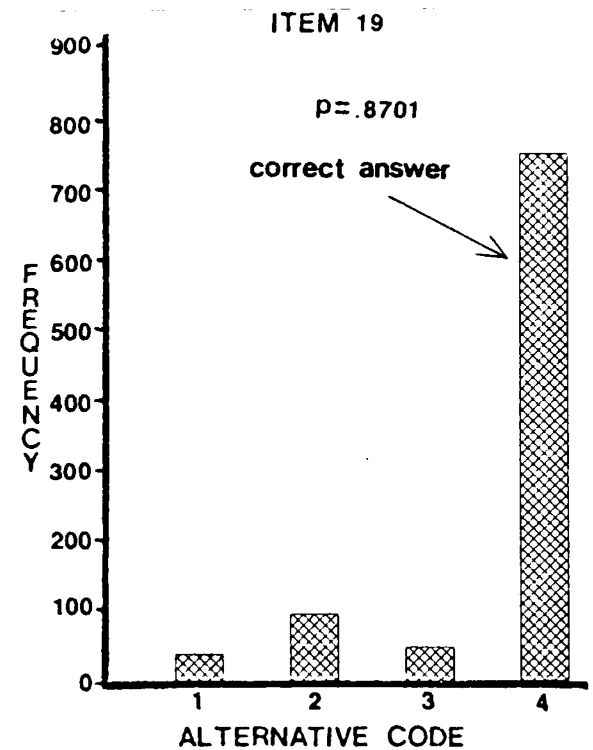


(14) Item 17

Figure B.1. (Continued)

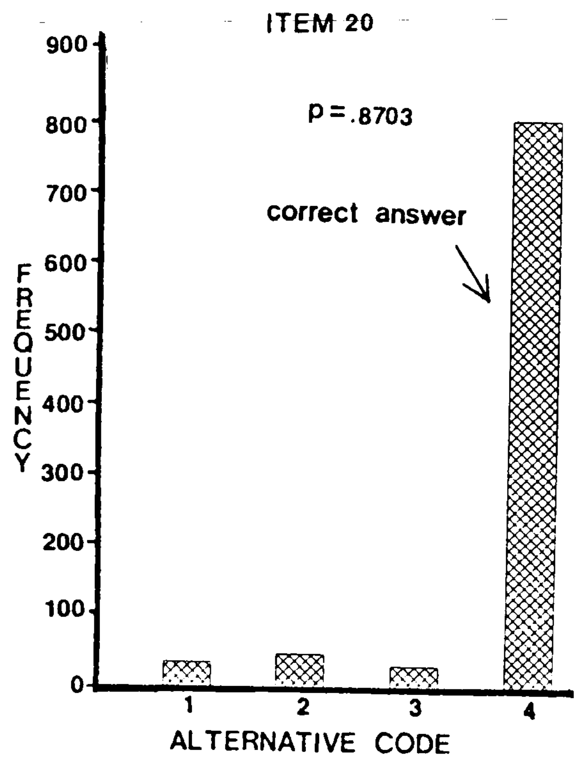


(15) Item 18

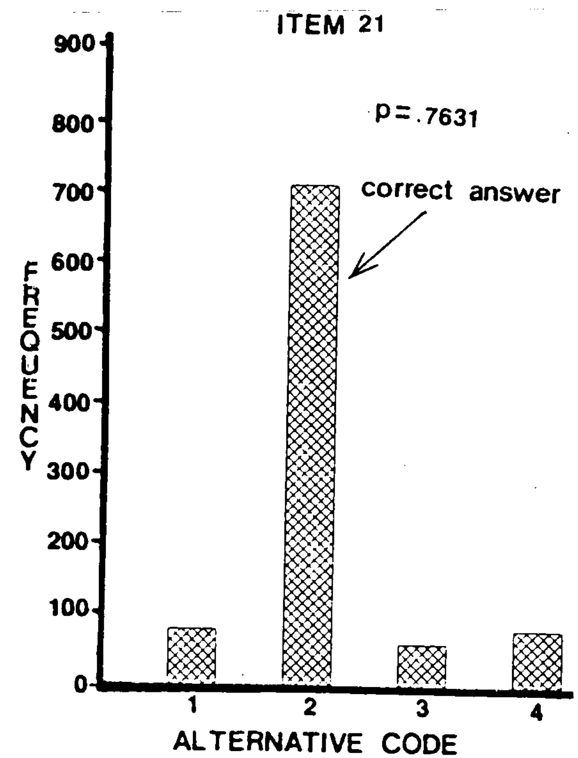


(16) Item 19

Figure B.1. (Continued)

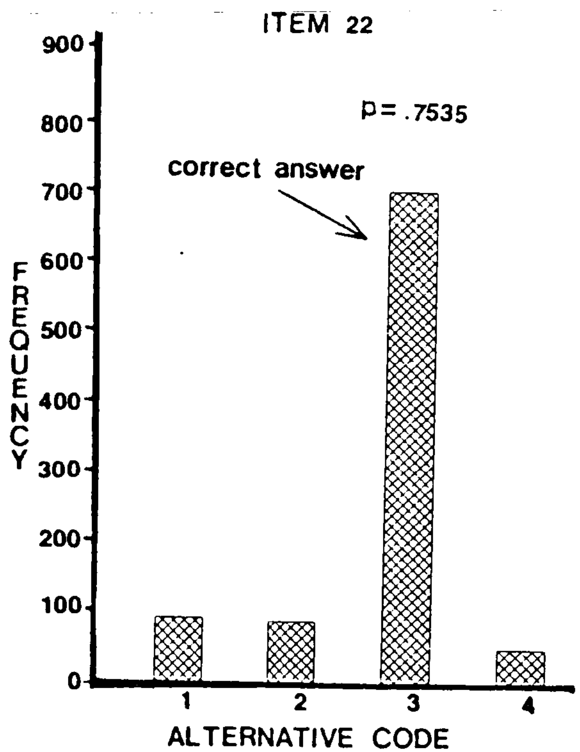


(17) Item 20

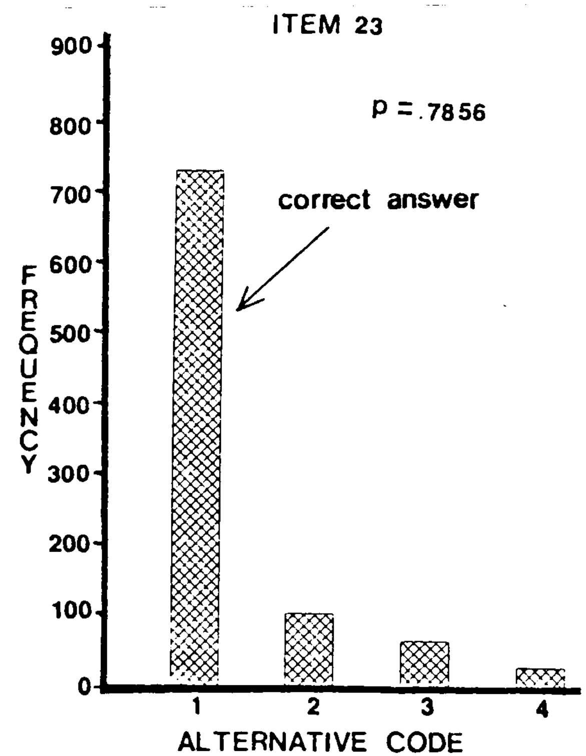


(18) Item 21

Figure B.1. (Continued)

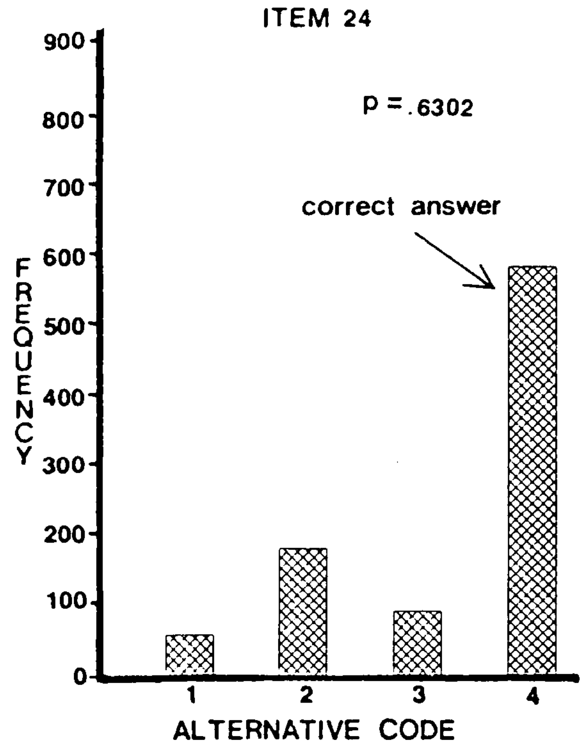


(19) Item 22

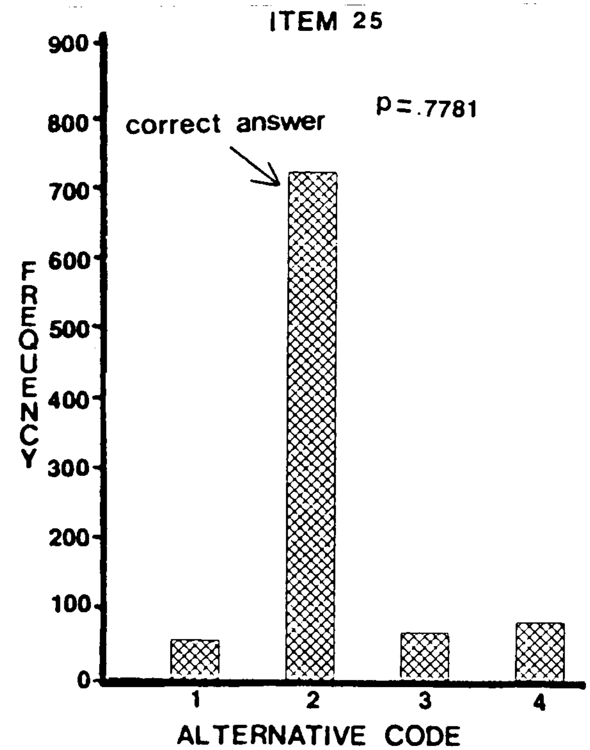


(20) Item 23

Figure B.1. (Continued)

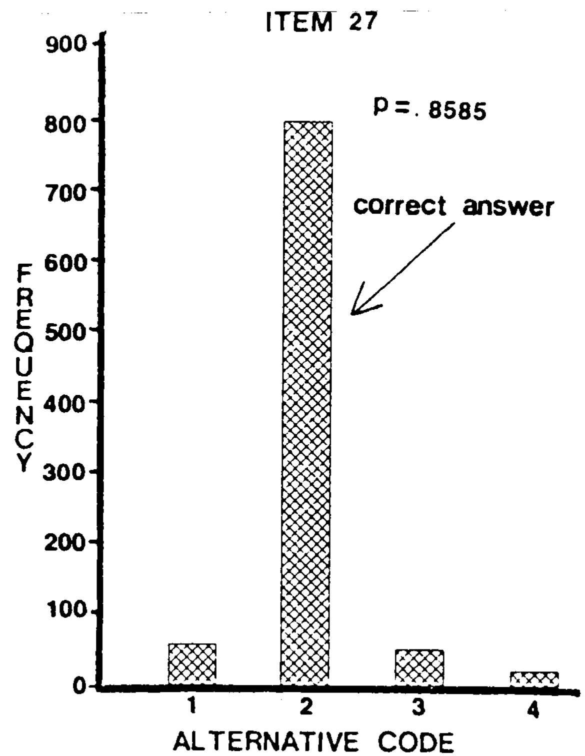


(21) Item 24

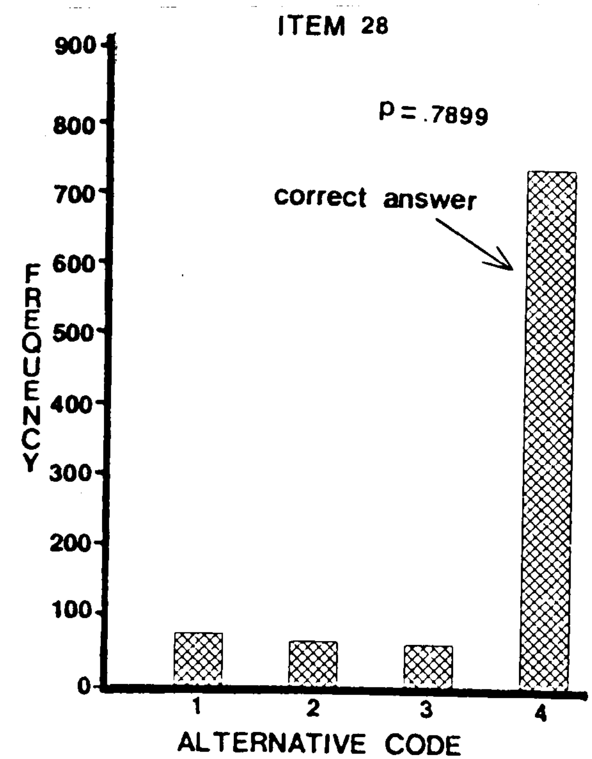


(22) Item 25

Figure B.1. (Continued)

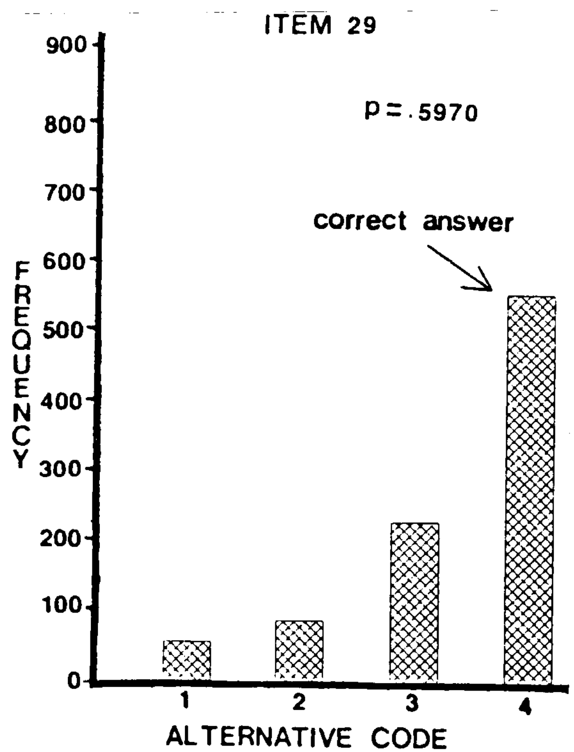


(23) Item 27

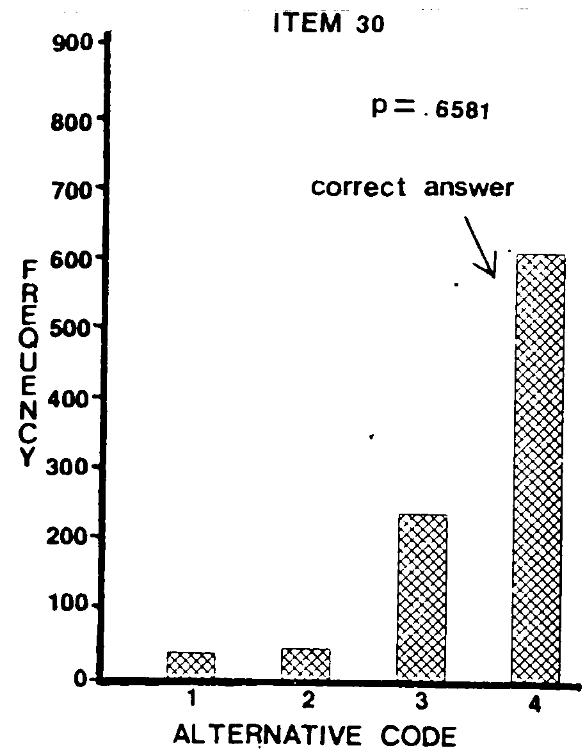


(24) Item 28

Figure B.1. (Continued)

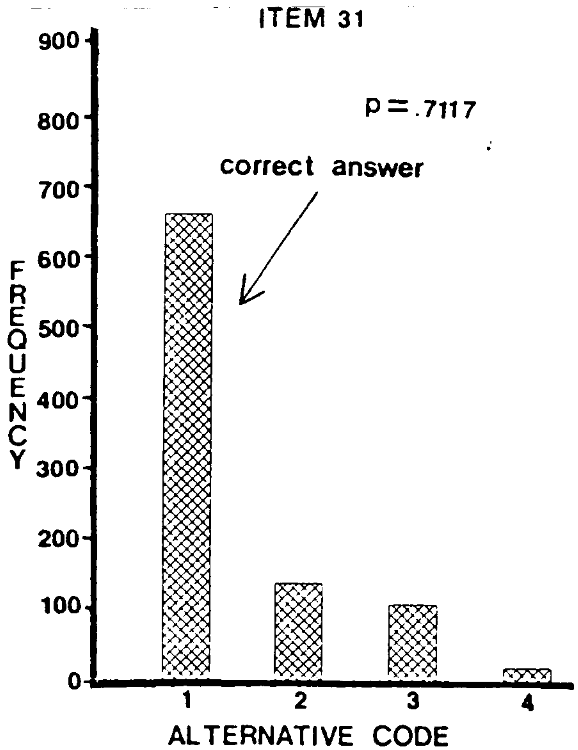


(25) Item 29

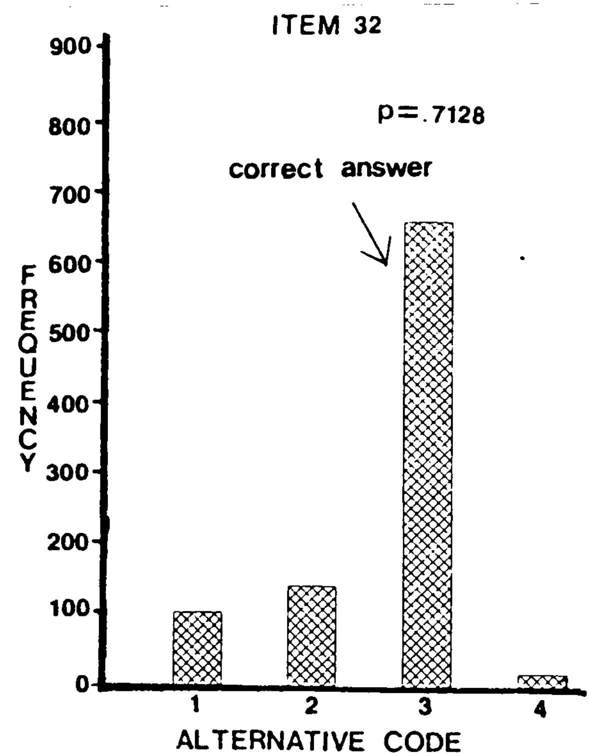


(26) Item 30

Figure B.1. (Continued)

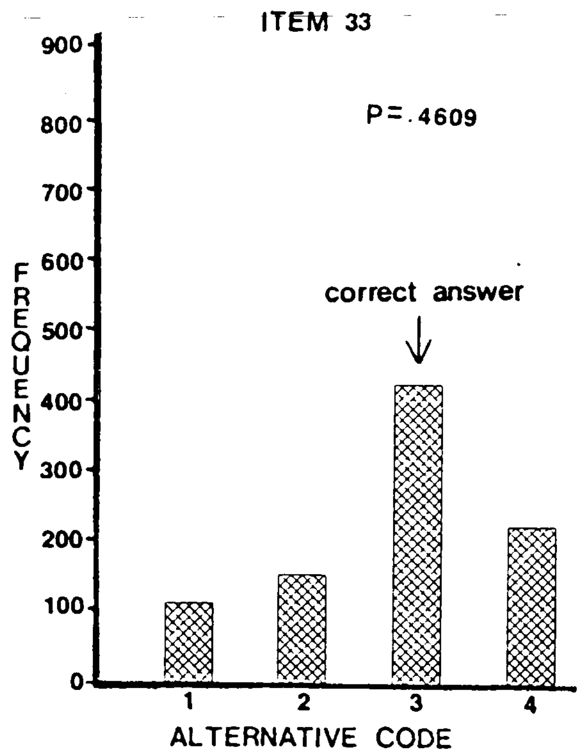


(27) Item 31

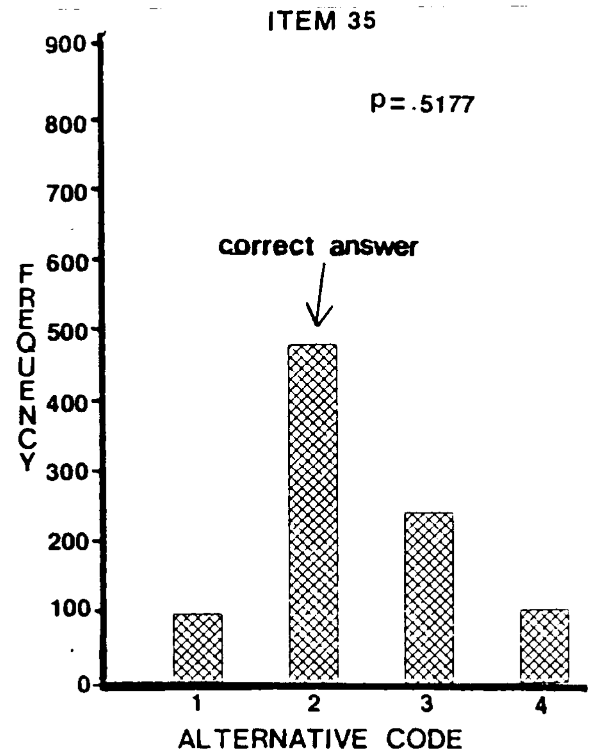


(28) Item 32

Figure B.1. (Continued)

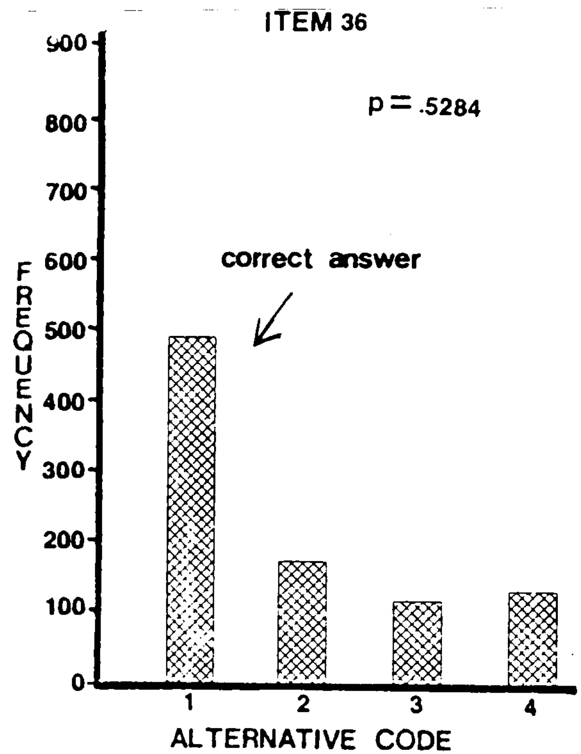


(29) Item 33

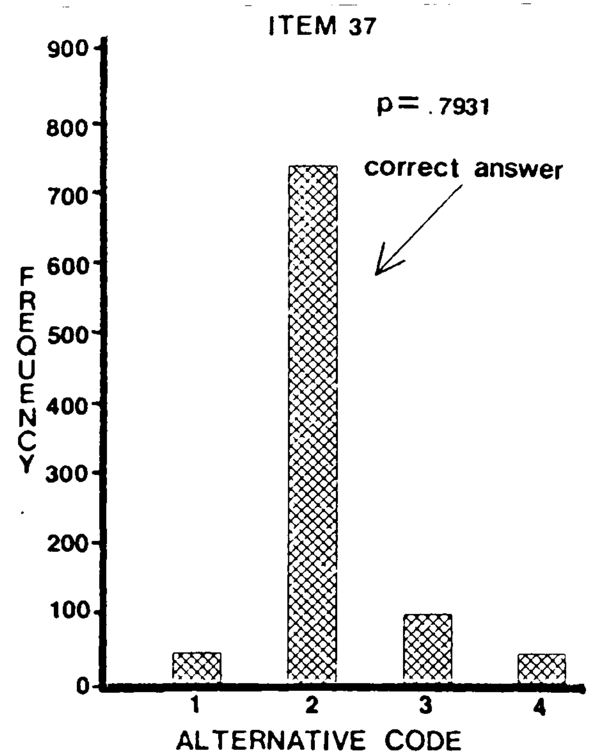


(30) Item 35

Figure B.1. (Continued)

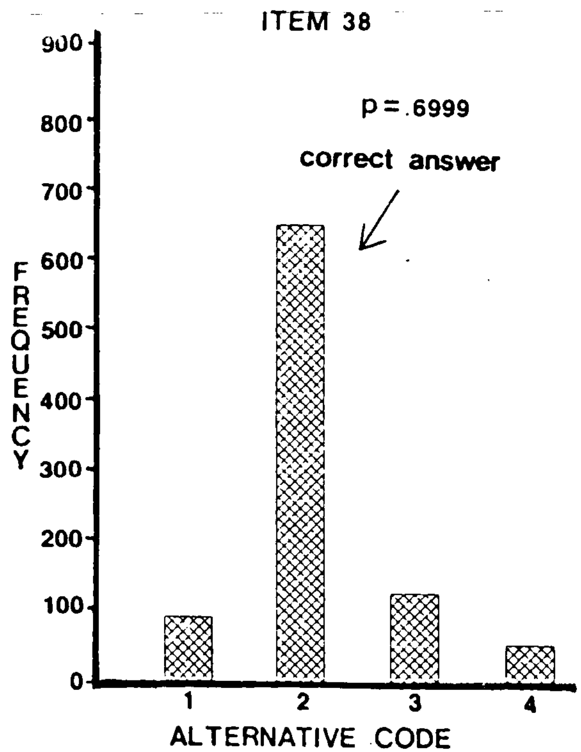


(31) Item 36

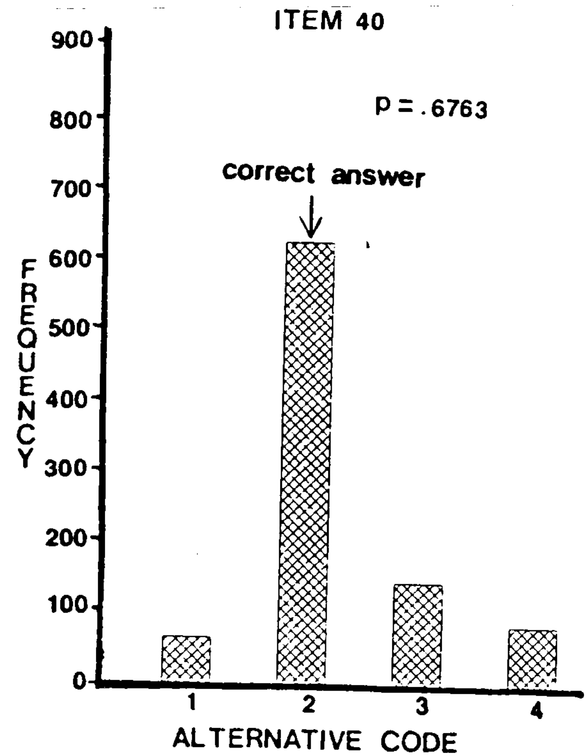


(32) Item 37

Figure B.1. (Continued)

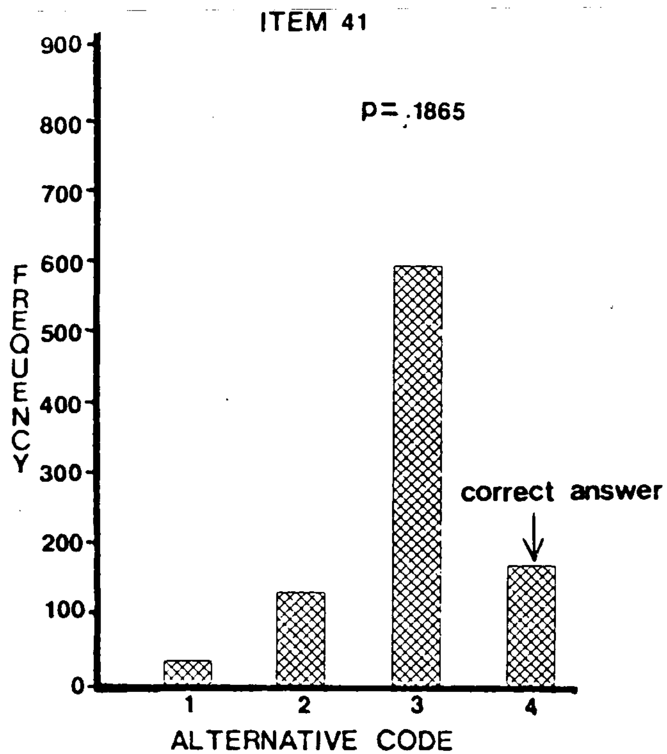


(33) Item 38

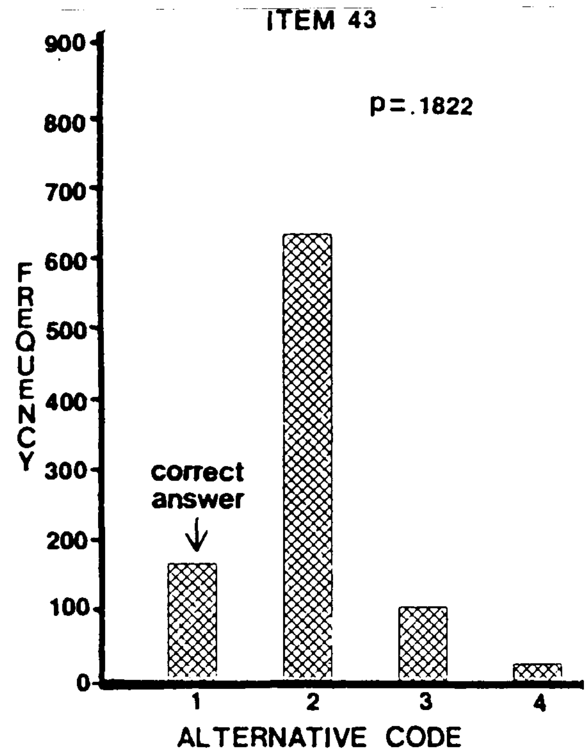


(34) Item 40

Figure B.1. (Continued)

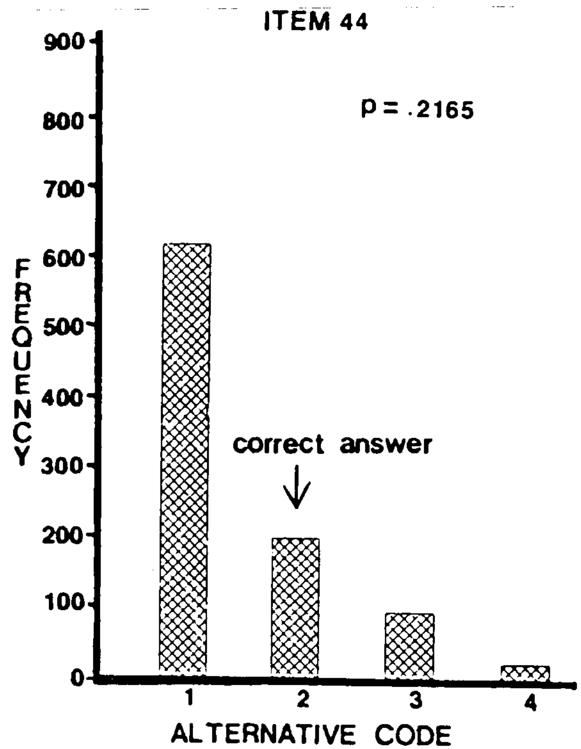


(35) IOtem 41

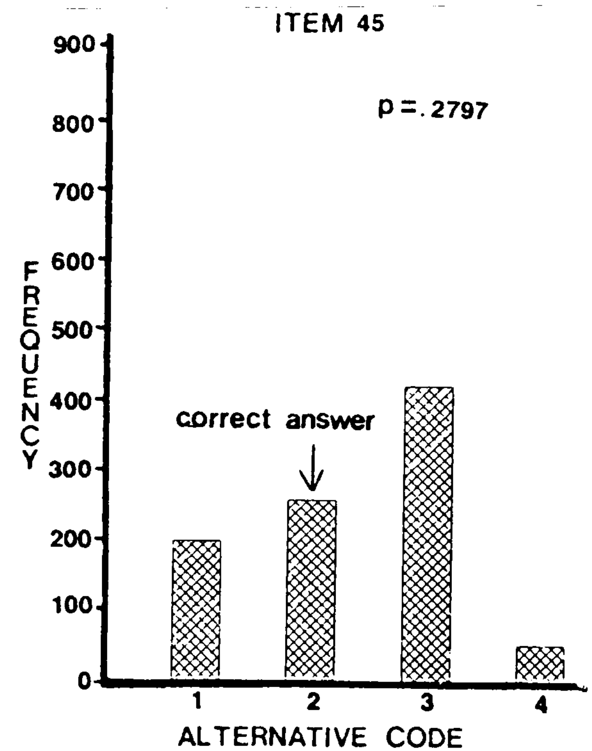


(36) Item 43

Figure B.1. (Continued)

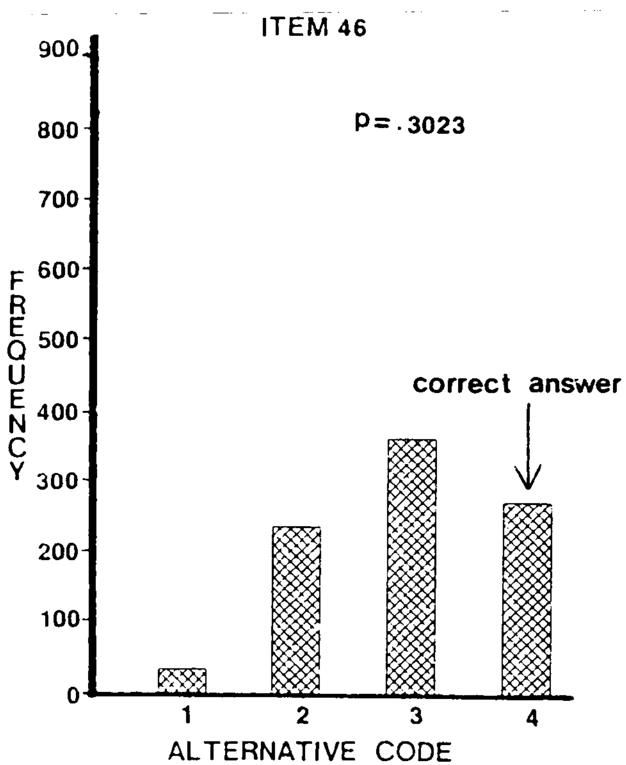


(37) Item 44

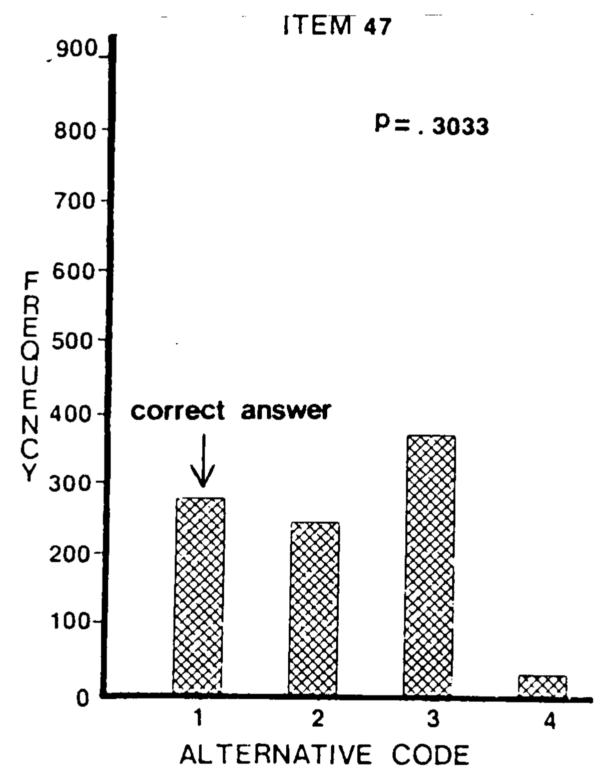


(38) Item 45

Figure B.1. (Continued)

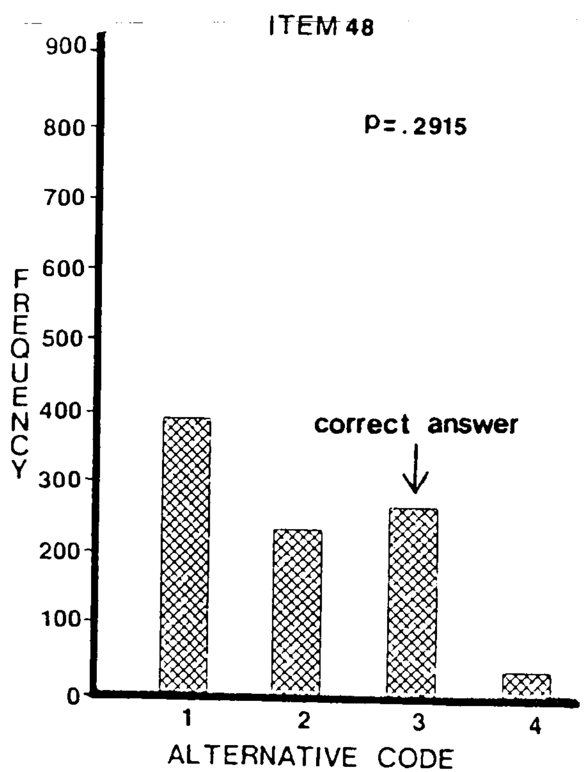


(39) Item 46

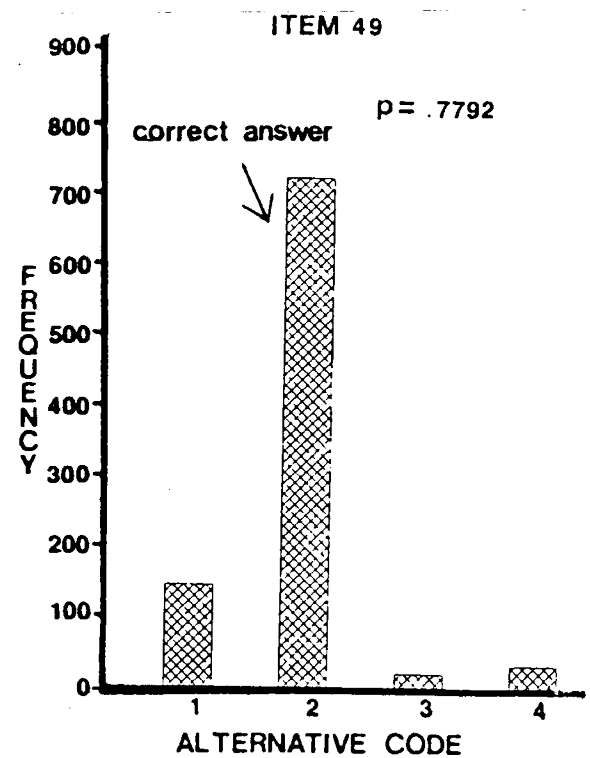


(40) Item 47

Figure B.1. (Continued)

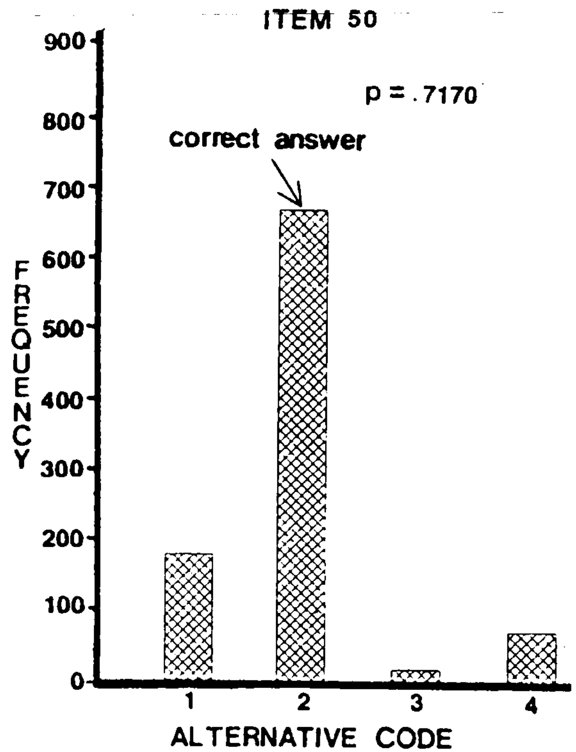


(41) Item 48

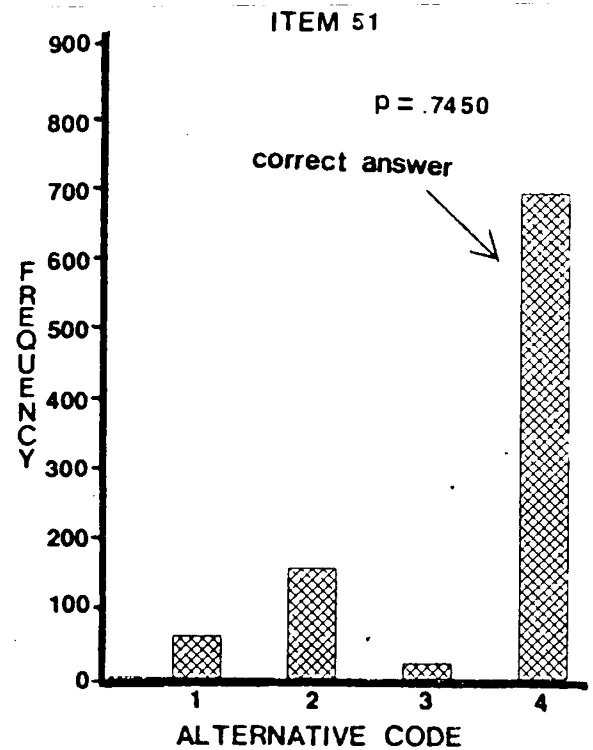


(42) Item 49

Figure B.1. (Continued)

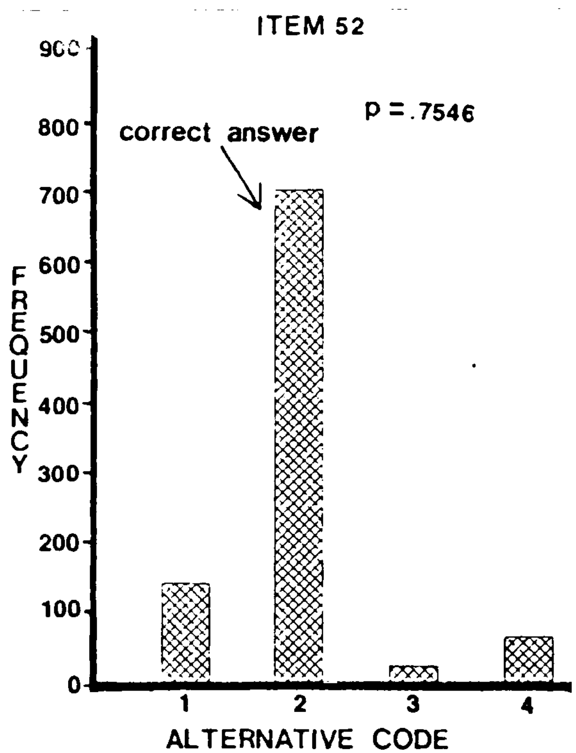


(43) Item 50

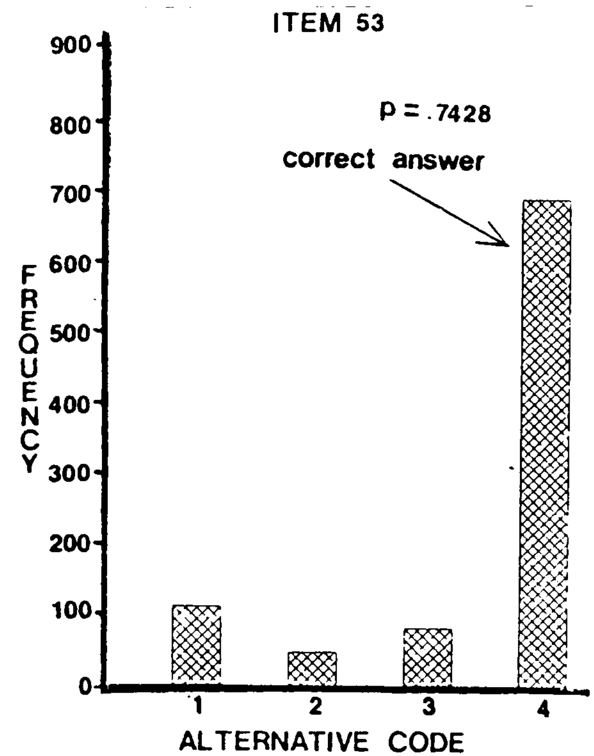


(44) Item 51

Figure B.1. (Continued)

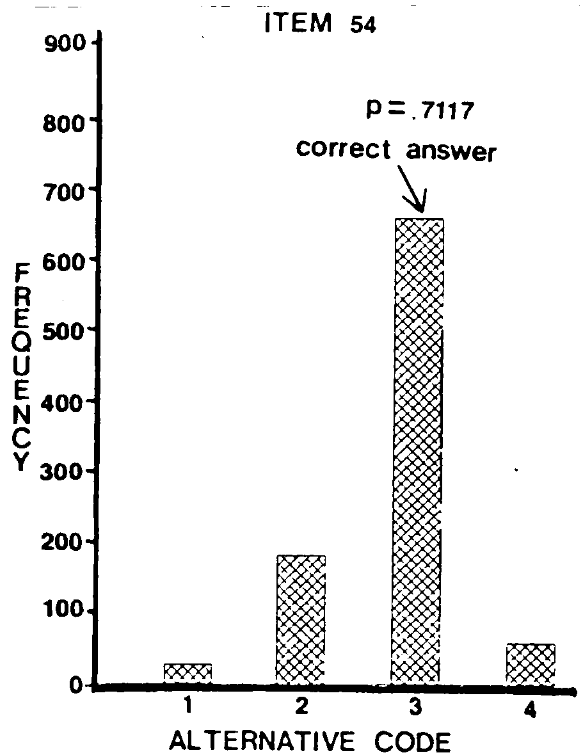


(45) Item 52

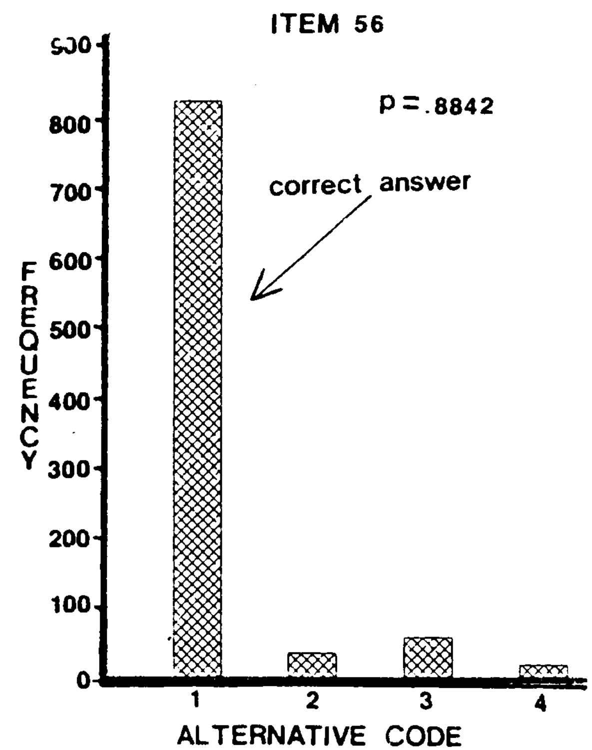


(46) Item 53

Figure B.1. (Continued)

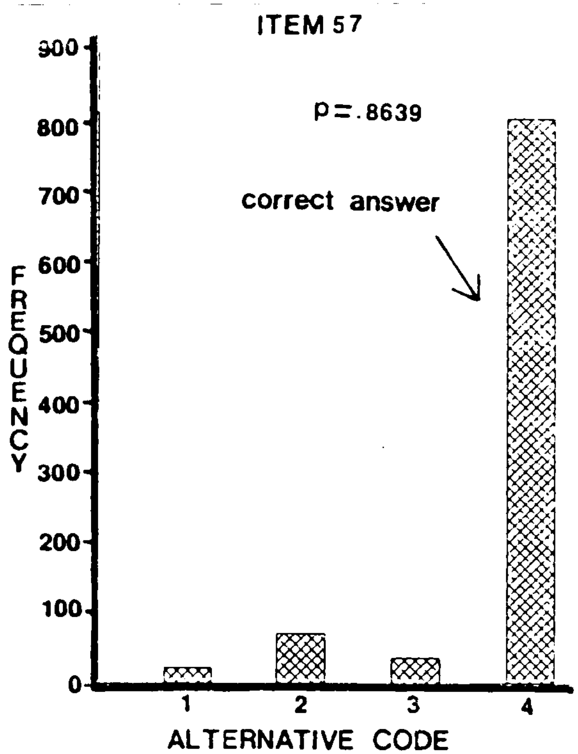


(47) Item 54

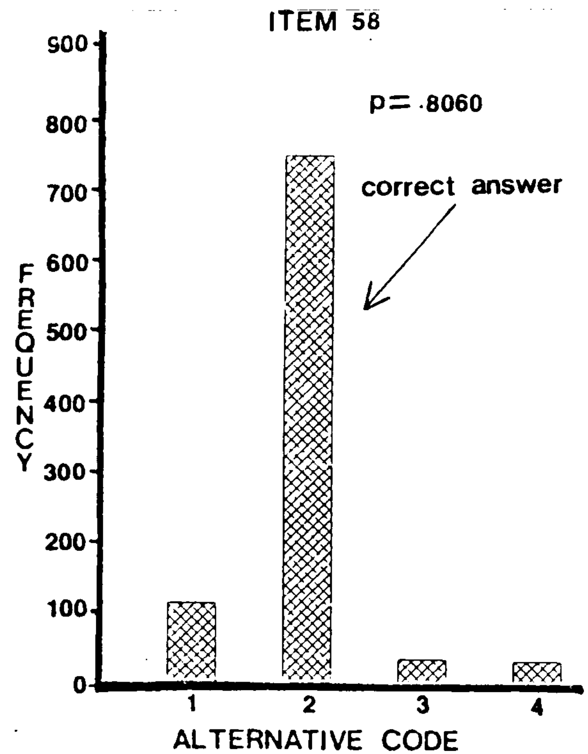


(48) Item 56

Figure B.1. (Continued)

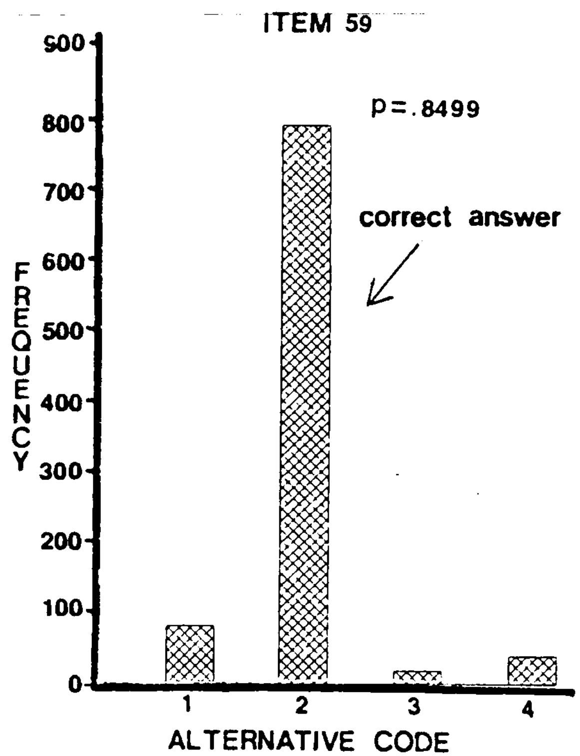


(49) Item 57

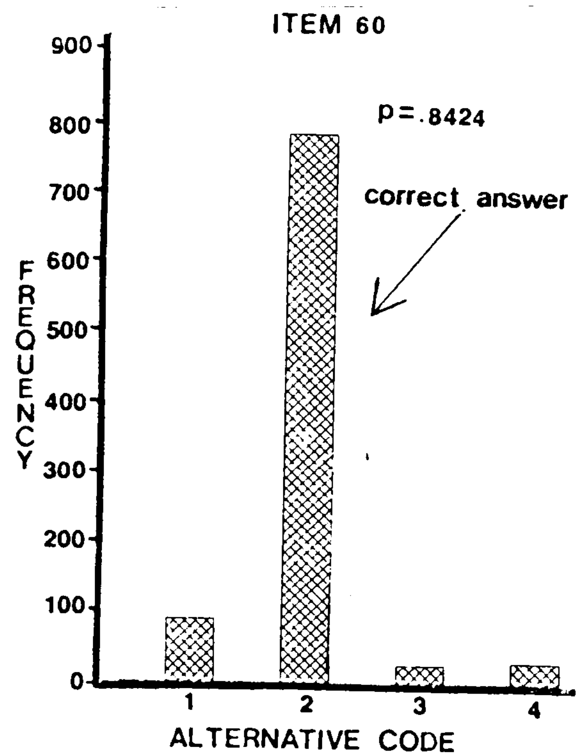


(50) Item 58

Figure B.1. (Continued)

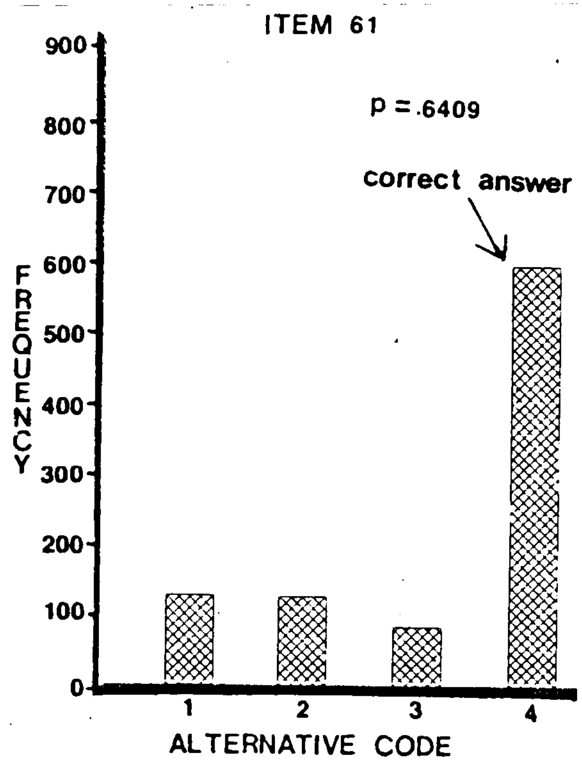


(51) Item 59

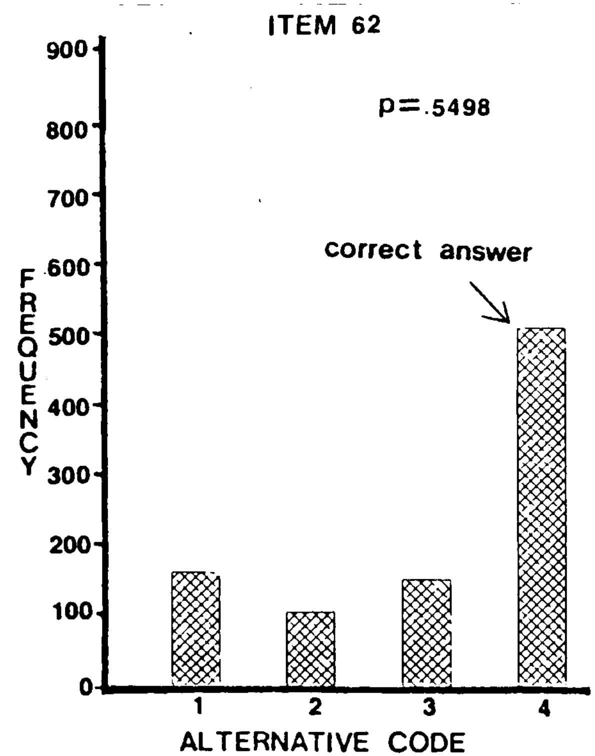


(52) Item 60

Figure B.1. (Continued)

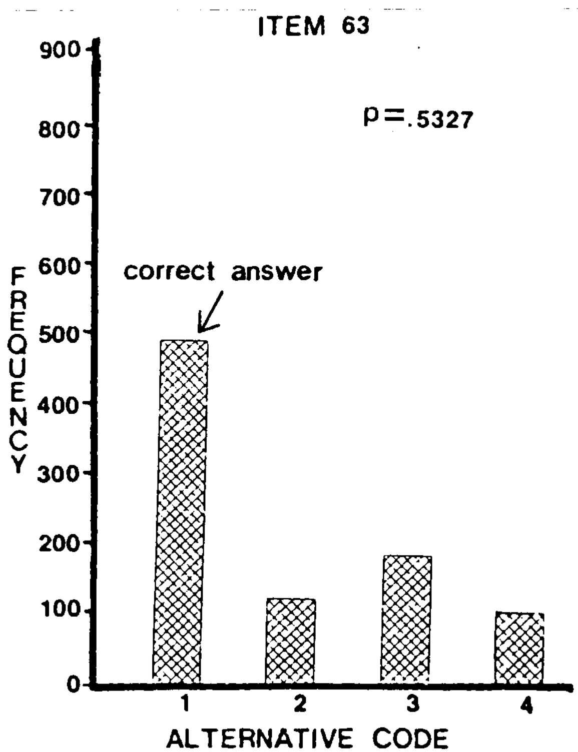


(53) Item 61

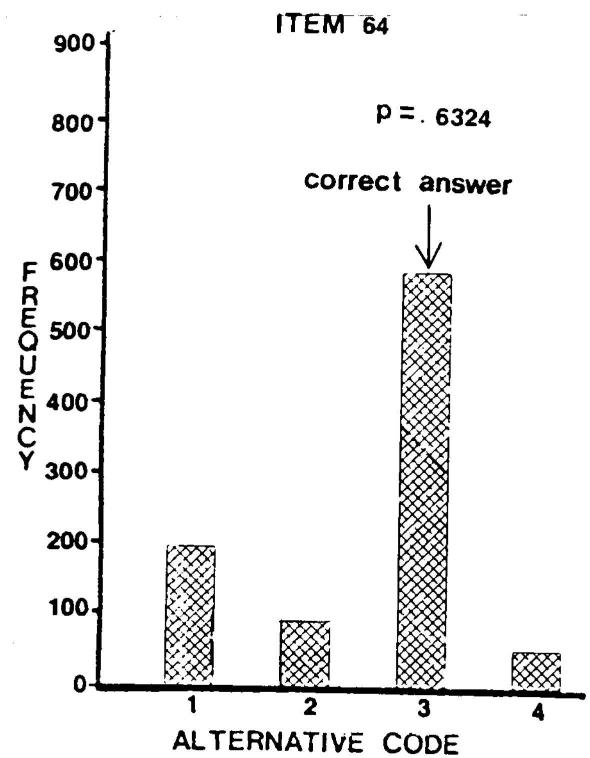


(54) Item 62

Figure B.1. (Continued)

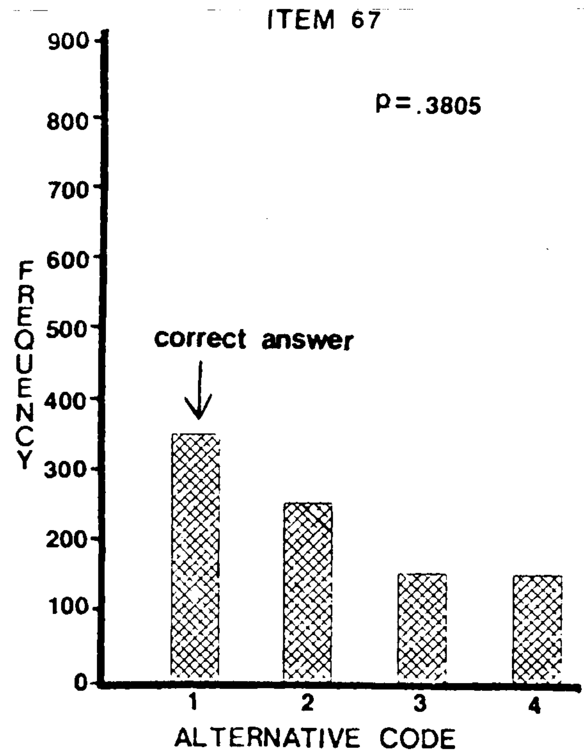


(55) Item 63

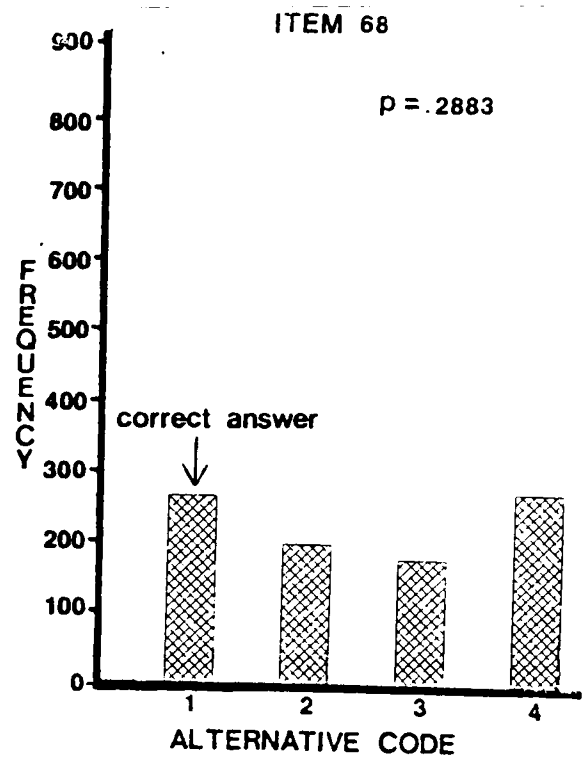


(56) Item 64

Figure B.1. (Continued)

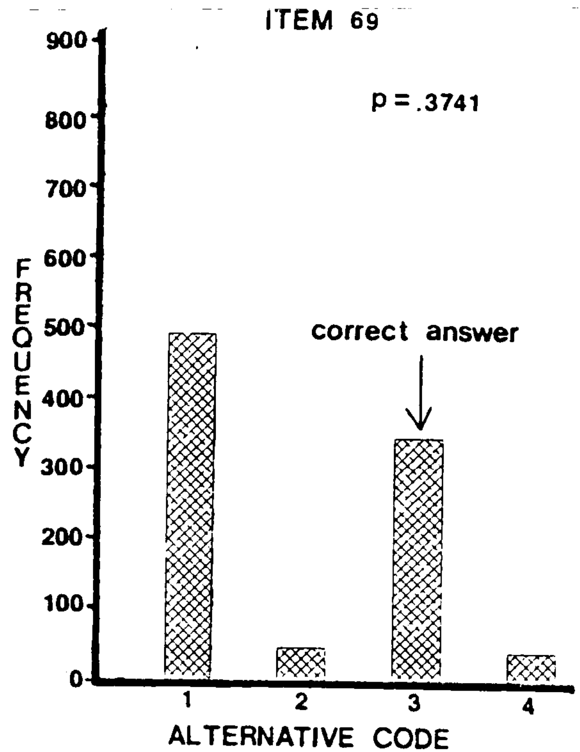


(57) Item 67

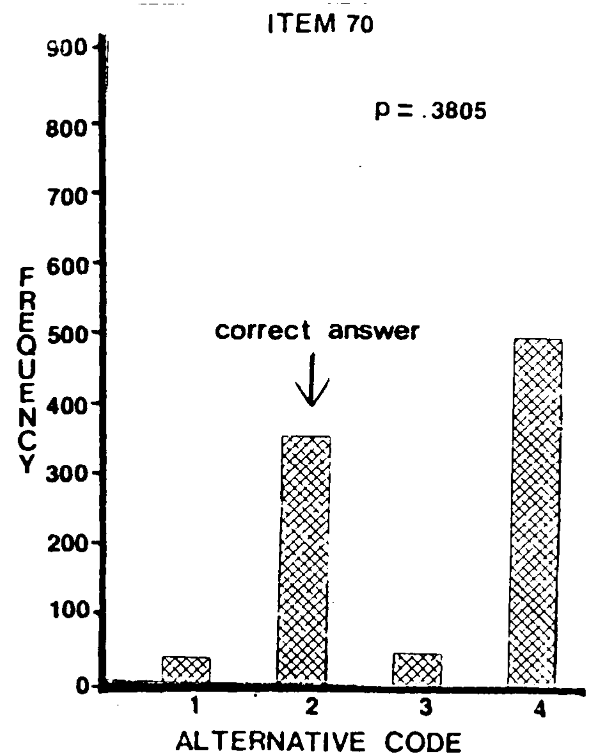


(58) Item 68

Figure B.1. (Continued)

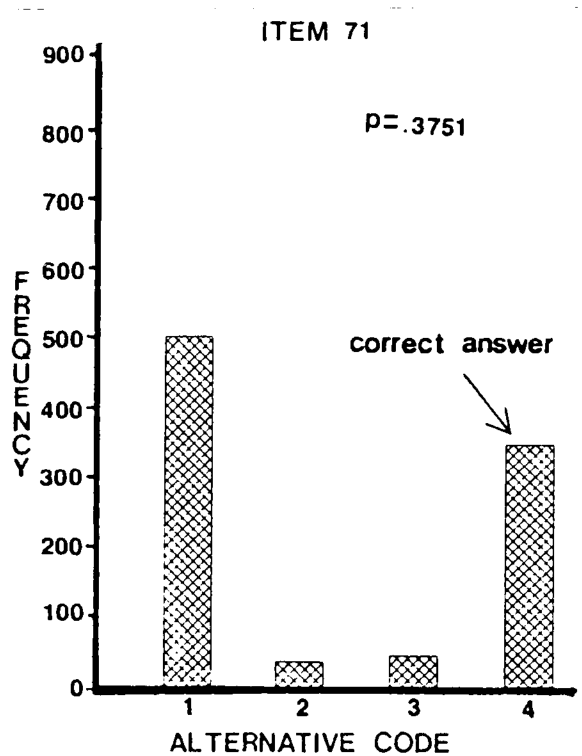


(59) Item 69

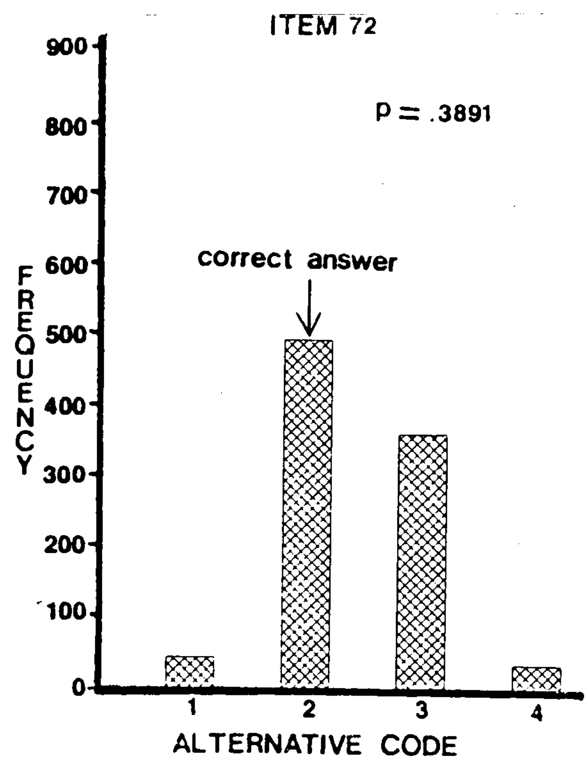


(60) Item 70

Figure B.1. (Continued)

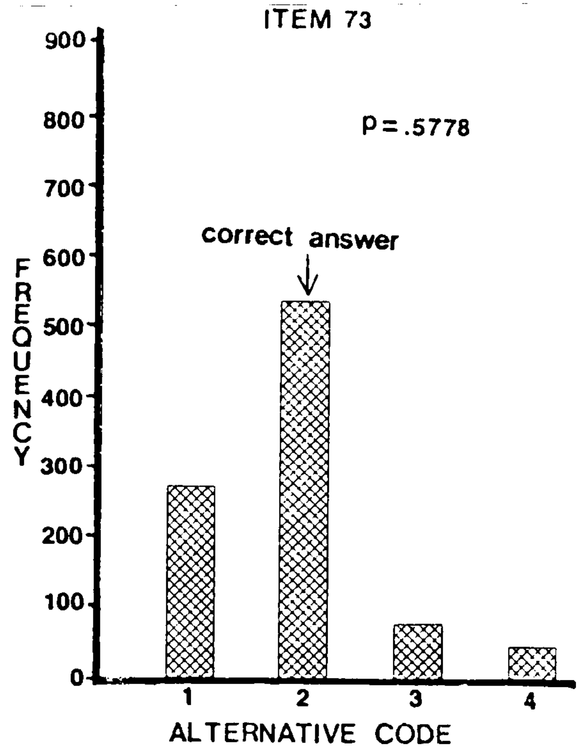


(61) Item 71

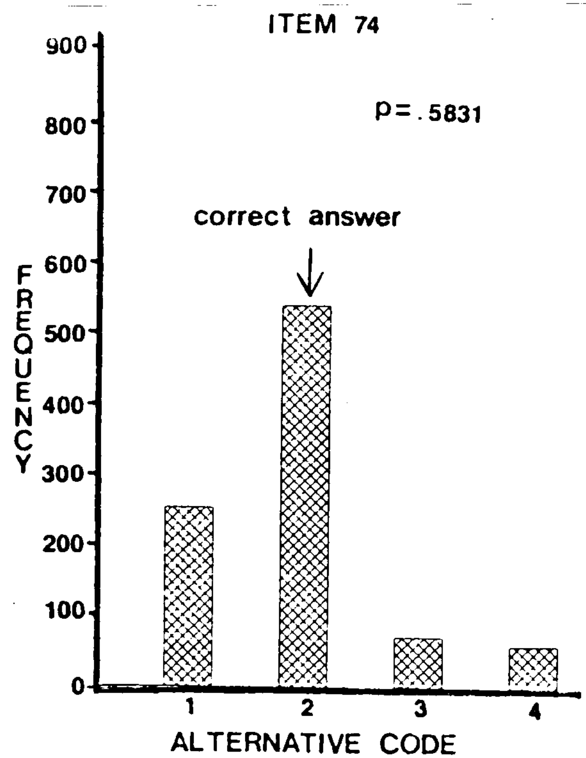


(62) Item 72

Figure B.1. (Continued)

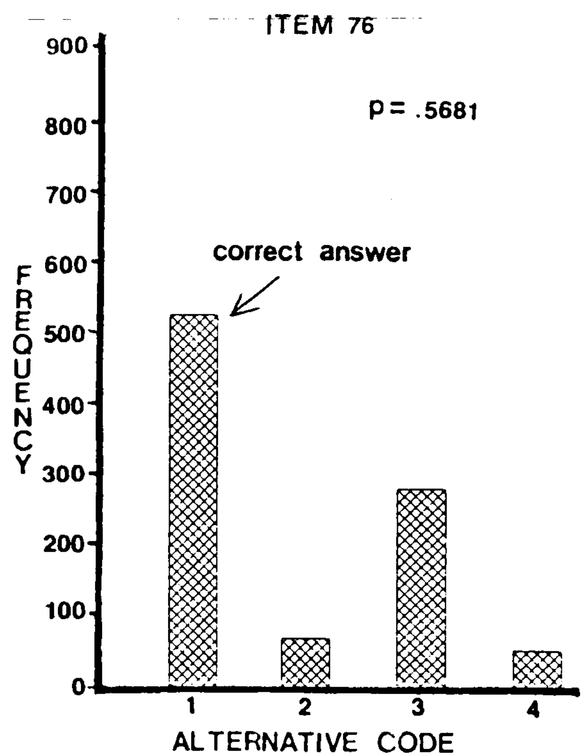


(63) Item 73



(64) Item 74

Figure B.1. (Continued)



(65) Item 76

Figure B.1. (Continued)

APPENDIX C

Table C.1. Factor loadings of 75 items

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
X1	0.39547	0.36799	-0.06049	0.10634
X2	0.35069	0.38233	-0.08204	-0.01292
X3	0.44619	0.34284	-0.07863	-0.07835
X4	0.48736	0.29076	0.07552	-0.07966
X5	0.70859	-0.06031	-0.16958	-0.04155
X6	0.72469	-0.10926	-0.11930	-0.03434
X7	0.73941	-0.06526	-0.19359	-0.01931
X8	0.67781	-0.07027	-0.15674	-0.04868
X9	0.61523	-0.24309	-0.26215	-0.09813
X10	0.58399	-0.23401	-0.19739	-0.11505
X11	0.49425	-0.24674	-0.24932	-0.09973
X12	0.61472	-0.16570	-0.20612	-0.12450
X13	0.52968	0.49241	0.09756	-0.06657
X14	0.63003	0.30704	0.02895	-0.02762
X15	0.57895	0.58717	0.05488	-0.03697
X16	0.63487	0.47663	0.06840	-0.06611
X17	0.55011	0.45759	0.03657	-0.11900
X18	0.54154	0.30247	0.10619	-0.08402
X19	0.62077	0.30627	0.00952	-0.11275
X20	0.63643	0.36386	0.03648	-0.07664
X21	0.60962	0.11557	-0.11948	-0.16554
X22	0.66919	0.15227	-0.08118	-0.14984
X23	0.67277	0.19225	-0.09375	-0.16438
X24	0.75118	0.13269	-0.07893	-0.03789
X25	0.59853	0.18558	0.29015	-0.08079
X27	0.57709	0.30171	0.22314	-0.08894
X28	0.61784	0.16190	0.27801	-0.03835
X29	0.44758	0.01998	0.22996	0.03918
X30	0.49887	0.02061	0.24235	0.06020
X31	0.55879	0.16003	0.35596	0.07858
X32	0.56763	0.11764	0.29668	0.14588
X33	0.49543	-0.18685	-0.06070	0.19491
X34	0.63819	-0.23488	-0.04649	0.30573
X35	0.56331	-0.23776	-0.10220	0.29945
X36	0.56132	-0.18984	-0.02526	0.23397
X37	0.57542	0.27463	-0.17068	-0.06976
X38	0.59297	0.10485	-0.07672	-0.02991
X39	0.46363	0.08798	-0.02031	-0.03632
X40	0.63354	0.10763	-0.11869	-0.06626
X41	0.65276	-0.50250	-0.18405	0.31515
X42	0.54432	-0.54267	-0.17802	0.37006
X43	0.70124	-0.45040	-0.13396	0.31201

Table C.1. (Continued)

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
X44	0.60177	-0.54411	-0.14973	0.28485
X45	0.39874	-0.52196	0.53414	0.05309
X46	0.42915	-0.49086	0.58394	0.02870
X47	0.44504	-0.47577	0.58683	0.01922
X48	0.44464	-0.50992	0.57025	-0.00036
X49	0.54937	0.29189	-0.27787	0.42535
X50	0.56629	0.28690	-0.28427	0.44818
X51	0.55530	0.33178	-0.32283	0.44496
X52	0.57090	0.35676	-0.33885	0.40478
X53	0.42204	0.25482	0.27291	0.24111
X54	0.33885	0.14058	0.20155	0.19414
X55	0.28926	-0.07754	0.20526	-0.00877
X56	0.32358	0.40224	0.24115	0.10383
X57	0.48832	0.27121	0.34265	0.00714
X58	0.53654	0.07320	0.21009	-0.00749
X59	0.46729	0.04472	0.27358	0.07954
X60	0.47338	0.19495	0.27237	0.08423
X61	0.63370	-0.07158	0.21889	0.00238
X62	0.65897	-0.12847	0.05289	0.02198
X63	0.58382	-0.06164	0.11304	-0.00187
X64	0.39775	0.00606	-0.08819	0.04014
X65	0.40968	-0.15667	0.09022	0.17087
X66	0.35703	-0.24802	-0.02428	0.11698
X67	0.17694	-0.01695	0.01253	0.01779
X68	0.31379	-0.24423	0.09307	0.16607
X69	0.60333	-0.44413	-0.20684	-0.21645
X70	0.62293	-0.38036	-0.17878	-0.21028
X71	0.68288	-0.33953	-0.13857	-0.21917
X72	0.59305	-0.41743	-0.15000	-0.26566
X73	0.63544	-0.18743	-0.11973	-0.42071
X74	0.65392	-0.07519	-0.08811	-0.45925
X75	0.64059	-0.14277	-0.13096	-0.46043
X76	0.62075	-0.13423	-0.08022	-0.47544

Table C.2. Factor loadings of 47 items

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
X3	0.46734	-0.28977	-0.17537	0.17689
X4	0.51423	-0.31610	-0.06828	0.21079
X5	0.71852	0.22688	-0.13573	0.27324
X6	0.74231	0.27332	-0.10172	0.31494
X7	0.76856	0.26660	-0.16218	0.30758
X8	0.69242	0.24533	-0.13329	0.31540
X9	0.55129	0.23348	-0.11452	0.09757
X14	0.64096	-0.24996	-0.05974	0.04049
X16	0.68575	-0.46392	-0.06987	0.01682
X17	0.59726	-0.41752	-0.19569	0.04171
X18	0.58180	-0.33752	-0.15720	-0.04171
X19	0.65444	-0.29539	-0.22410	-0.00978
X20	0.69476	-0.33382	-0.21094	0.03665
X21	0.62555	-0.07450	-0.14737	-0.09307
X22	0.69969	-0.09483	-0.26370	-0.16232
X23	0.69347	-0.16980	-0.24757	-0.05131
X24	0.76466	-0.07224	-0.20429	-0.04747
X25	0.64295	-0.20490	0.31578	-0.38030
X27	0.61924	-0.30026	0.29787	-0.34222
X28	0.65207	-0.17715	0.32841	-0.41021
X32	0.52641	-0.06448	0.10480	-0.10407
X33	0.50875	0.31443	-0.02027	-0.09047
X34	0.62296	0.41972	0.09793	-0.04491
X35	0.54612	0.48338	0.10409	-0.01747
X36	0.56157	0.37132	0.11535	-0.18747
X37	0.59561	-0.12343	-0.13263	-0.04163
X38	0.62676	-0.00623	-0.17662	-0.05150
X39	0.49991	-0.04969	-0.12435	0.02121
X40	0.64516	0.00579	-0.19909	-0.15288
X43	0.62667	0.44525	0.02671	0.05585
X45	0.32195	0.30610	0.15747	-0.01748
X50	0.53531	-0.04058	-0.06560	-0.03886

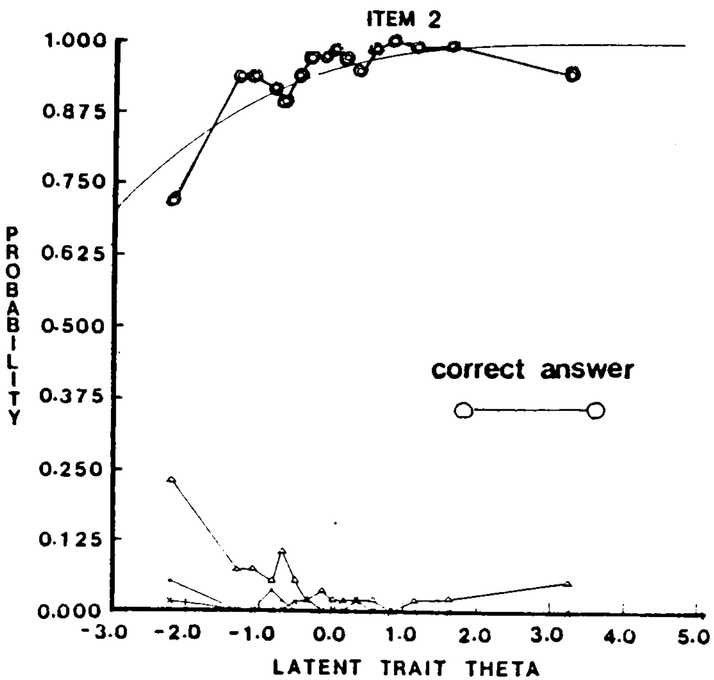
Table C.2. (Continued)

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
X54	0.34400	-0.06174	0.36264	0.41202
X55	0.27751	0.05243	0.31600	0.35903
X56	0.34969	-0.40000	0.48884	0.55374
X57	0.51234	-0.24498	0.40780	0.07454
X58	0.54464	-0.06041	0.32330	-0.18450
X60	0.46815	-0.14680	0.41127	0.07926
X61	0.62834	0.11430	0.20144	-0.09807
X62	0.64458	0.21245	0.19938	-0.04222
X63	0.58788	0.07852	0.07027	-0.04592
X64	0.38854	0.07986	0.16747	0.02568
X65	0.40500	0.24061	0.07885	-0.21619
X66	0.33363	0.30467	-0.02842	-0.07358
X67	0.17280	-0.04249	-0.0C714	0.09829
X69	0.51785	0.38780	-0.05926	-0.06393
X73	0.56369	0.13284	-0.05598	-0.02654

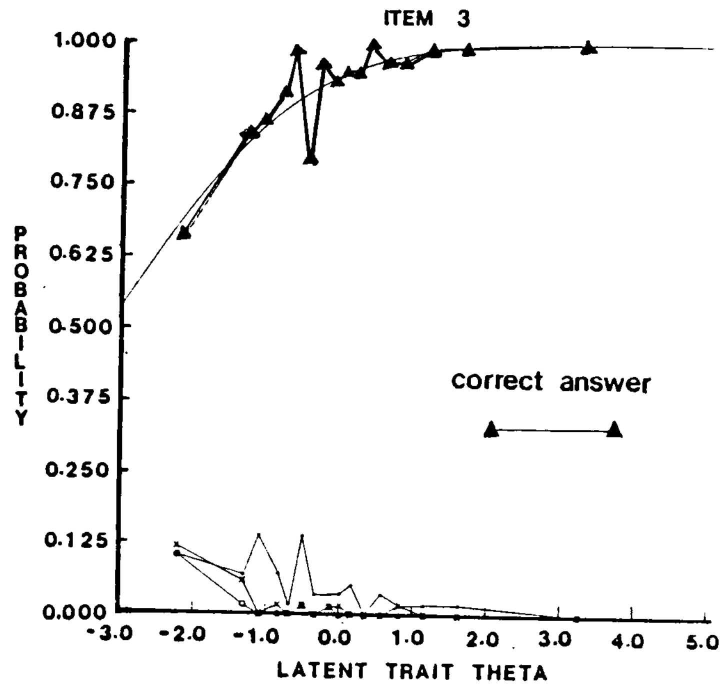
Table C.3. Factor loadings of 20 items

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
X3	0.40692	-0.32187	-0.07939	-0.05929
X5	0.68284	0.09092	-0.05557	-0.06035
X9	0.62038	0.21357	-0.20810	-0.03176
X14	0.63545	-0.28134	0.04626	-0.03480
X19	0.61188	-0.24548	-0.07698	0.02677
X21	0.60108	-0.15748	-0.18240	0.13443
X25	0.56877	-0.20156	0.24813	0.12035
X32	0.49873	-0.13390	0.20066	-0.00811
X33	0.48899	0.10222	-0.03022	0.00072
X37	0.60314	-0.27072	-0.19510	-0.02161
X40	0.63629	-0.14231	-0.09305	-0.02270
X43	0.73115	0.56458	0.06223	-0.37039
X45	0.34899	0.35449	0.29668	0.08517
X50	0.55158	-0.16236	-0.07010	-0.21201
X55	0.26922	0.12751	0.18616	0.21526
X57	0.48546	-0.25441	0.41455	-0.05214
X64	0.39914	0.05294	0.09554	-0.04791
X66	0.31129	0.25014	-0.04901	-0.02623
X69	0.58788	0.35718	-0.15836	0.30119
X73	0.61846	0.12703	-0.02324	0.23980

APPENDIX D

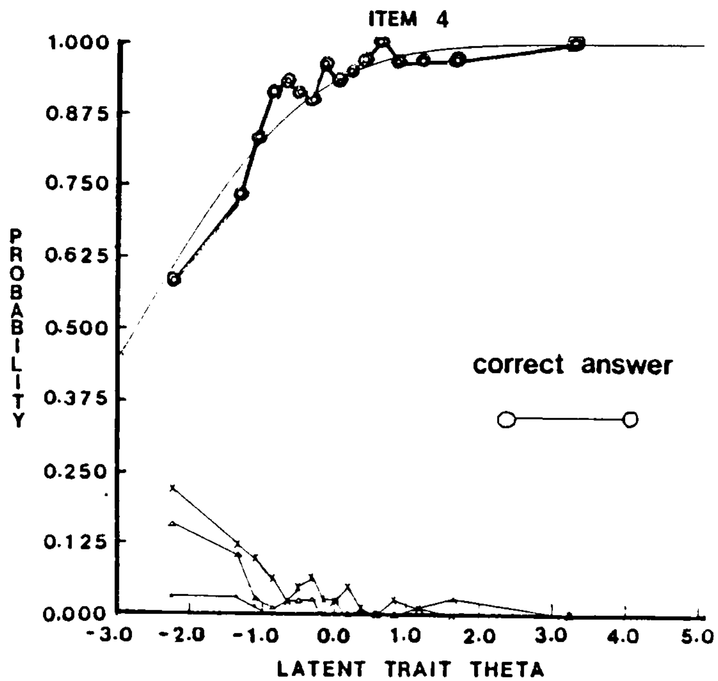


(1) Item 2: $a_g = 0.37447$
$b_g = -4.40749$

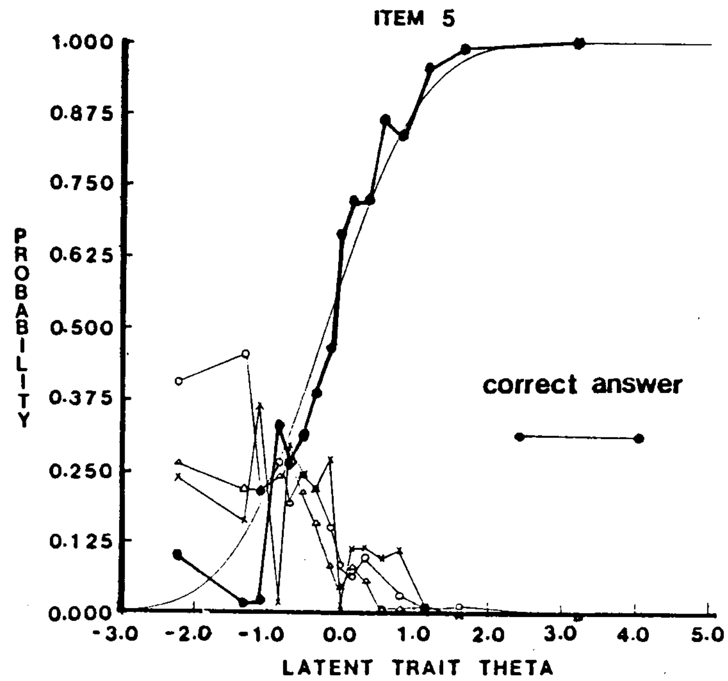


(2) Item 3: $a_g = 0.49857$
$b_g = -3.17577$

Figure D.1. Item characteristic curves and plausibility functions of test items.

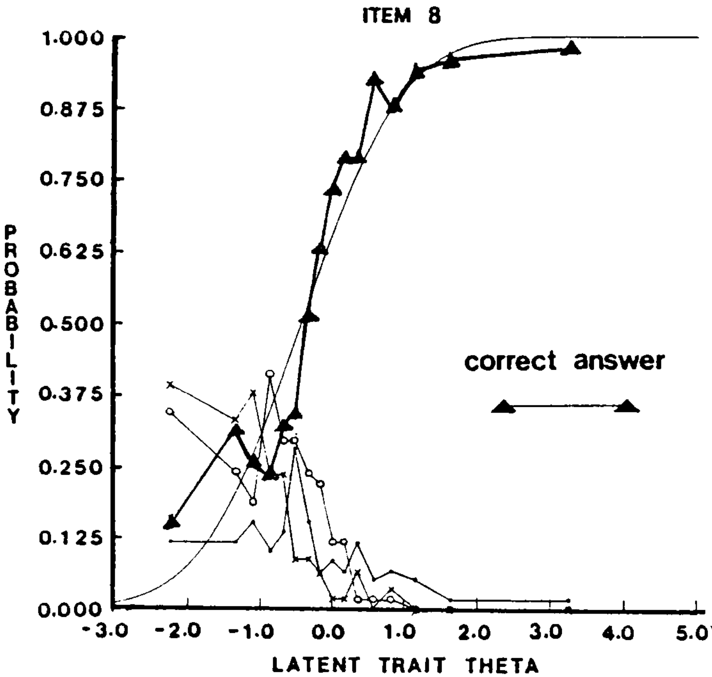


(3) Item 4: $a_g = 0.55813$
$b_g = -2.73717$

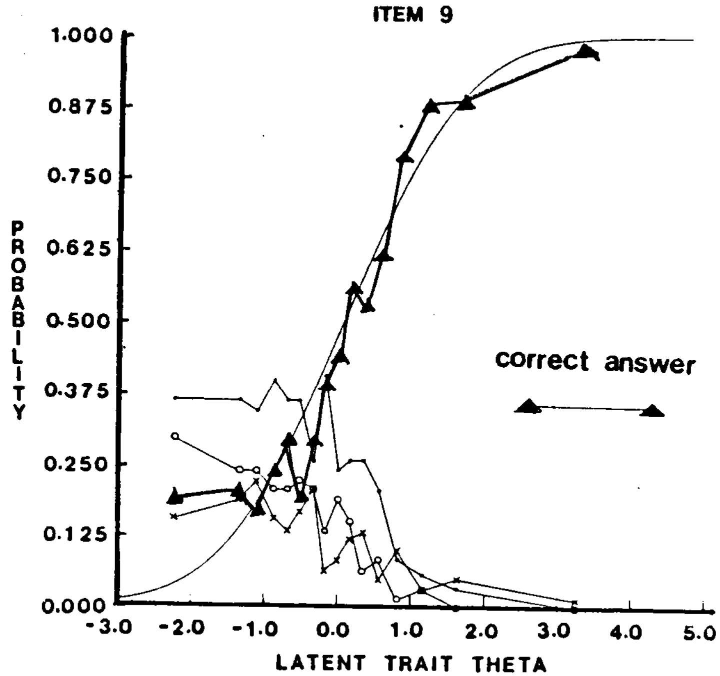


(4) Item 5: $a_g = 1.00421$
$b_g = -0.21505$

Figure D.1. (Continued)

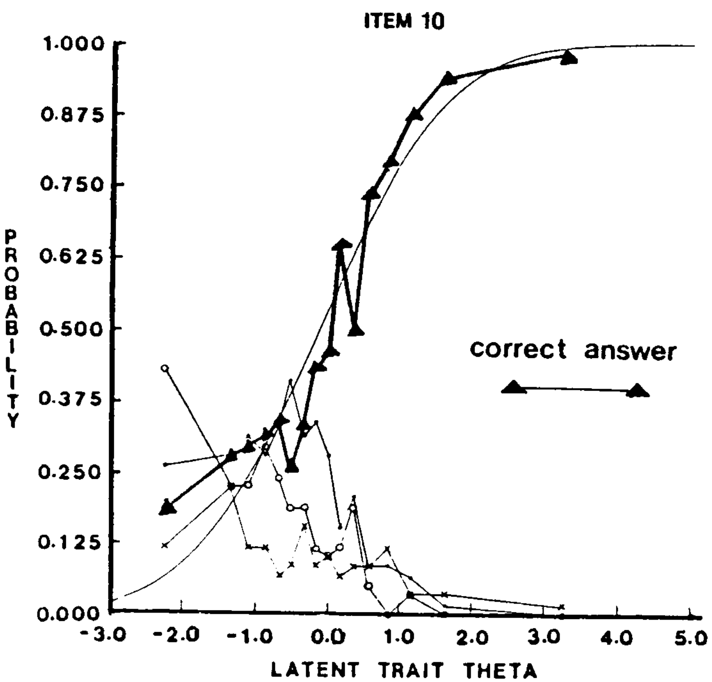


(5) Item 8: $a_g = 0.92189$
$b_g = -0.43634$

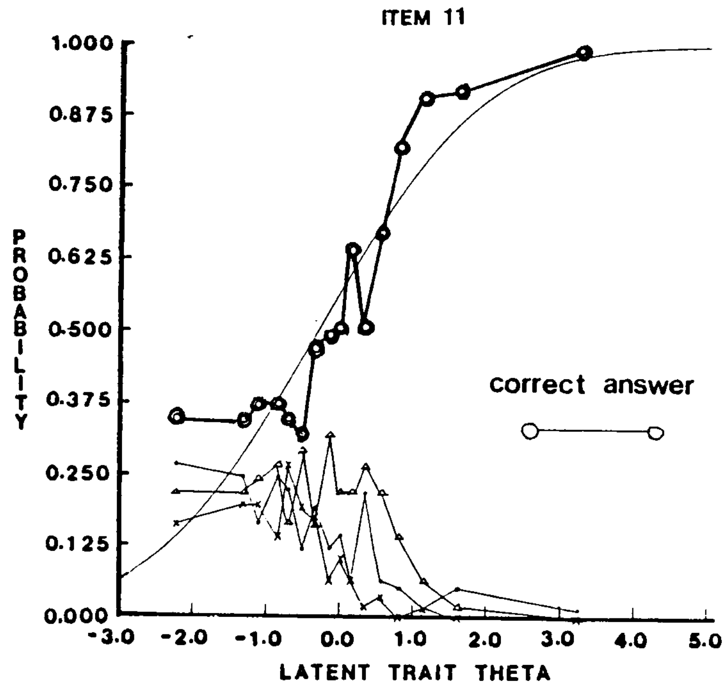


(6) Item 9: $a_g = 0.78040$
$b_g = 0.06771$

Figure D.1. (Continued)

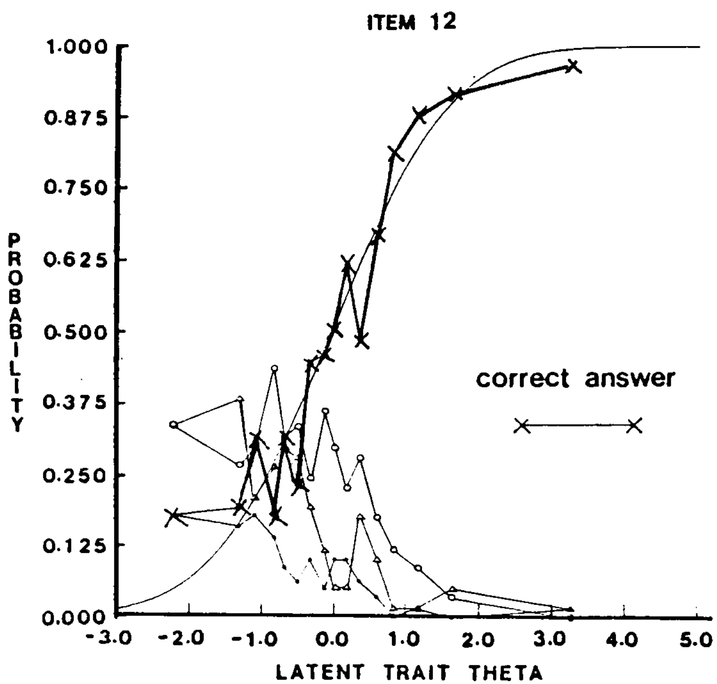


(7) Item 10: $a_g = 0.71941$
$b_g = -0.11740$

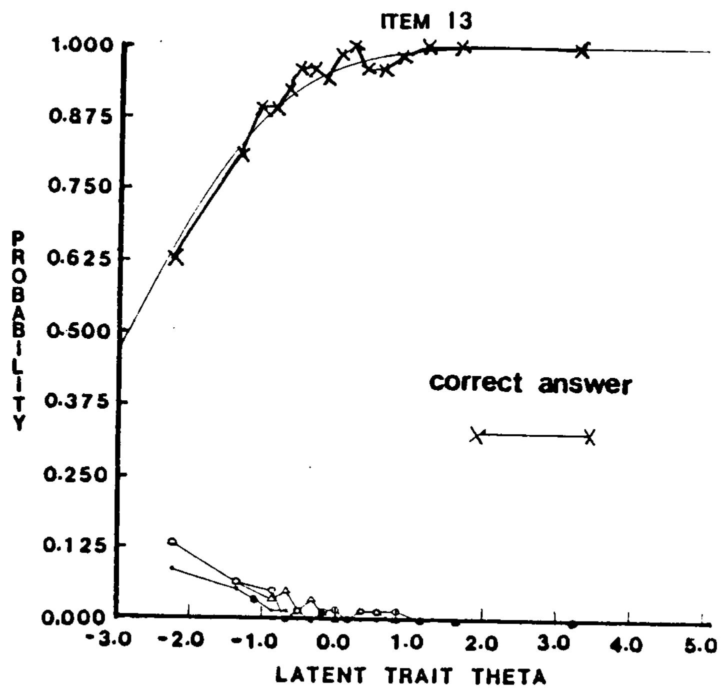


(8) Item 11: $a_g = 0.56855$
$b_g = -0.30831$

Figure D.1. (Continued)

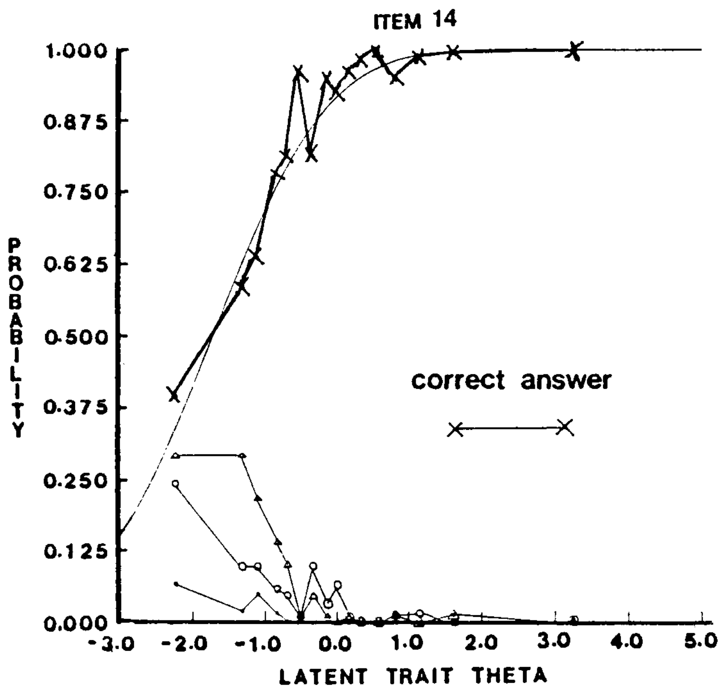


(9) Item 12: $a_g = 0.77936$
$b_g = -0.04152$

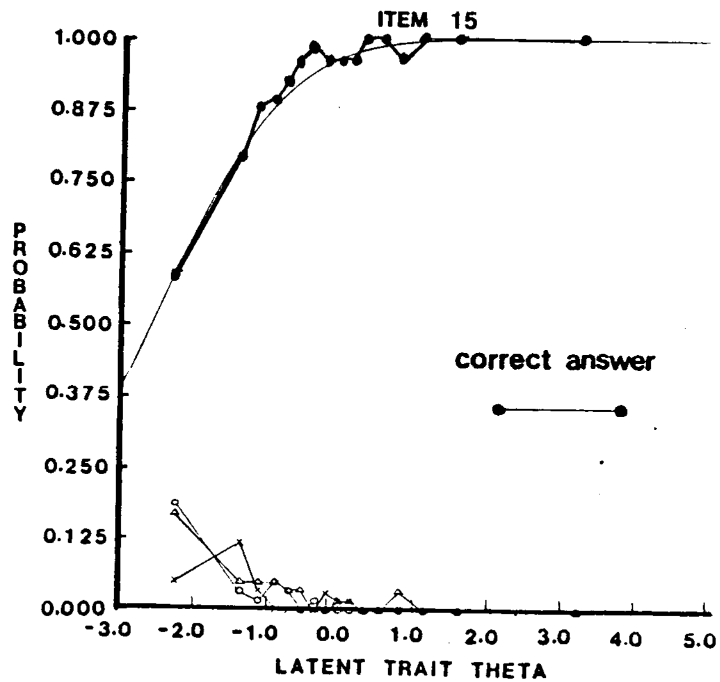


(10) Item 13: $a_g = 0.62448$
$b_g = -2.85287$

Figure D.1. (Continued)

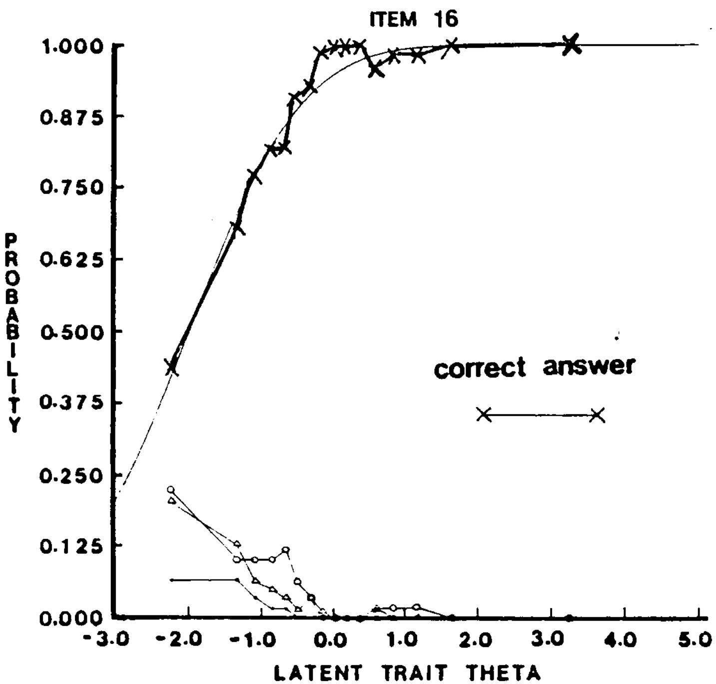


(11) Item 14: $a_g = 0.81130$
$b_g = -1.70420$

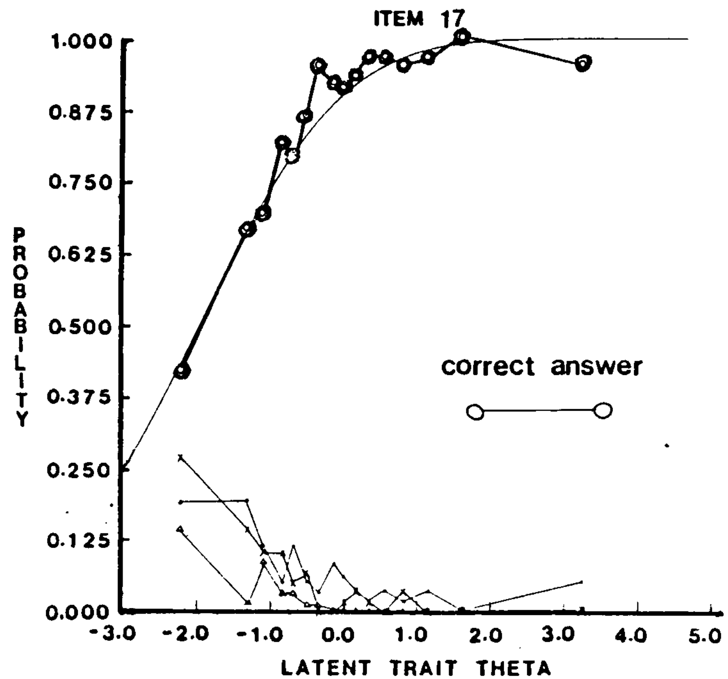


(12) Item 15: $a_g = 0.71005$
$b_g = -2.56729$

Figure D.1. (Continued)

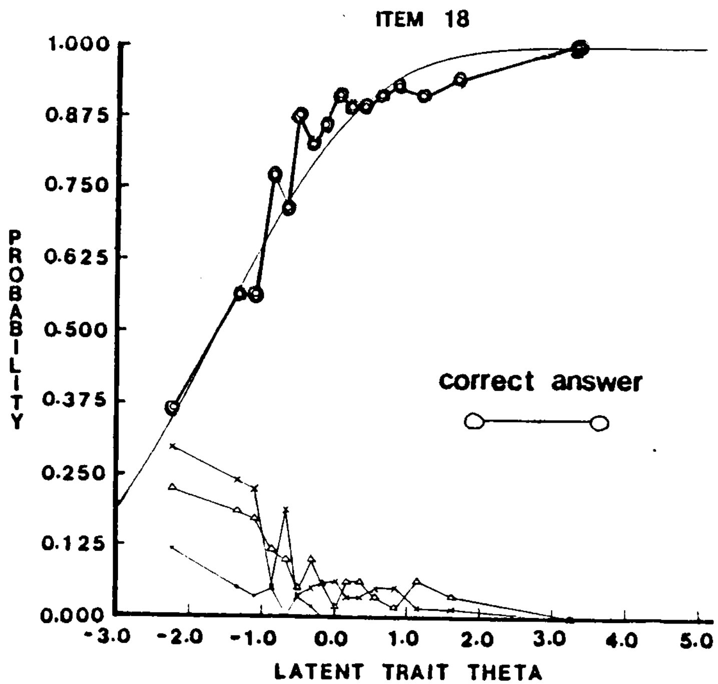


(13) Item 16: $a_g = 0.82171$
$b_g = -1.98351$

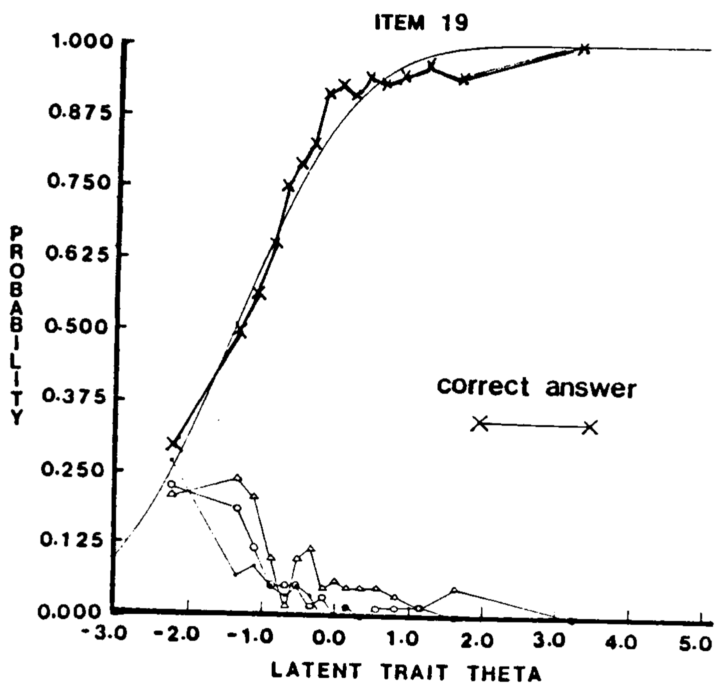


(14) Item 17: $a_g = 0.65874$
$b_g = -1.95179$

Figure D.1. (Continued)

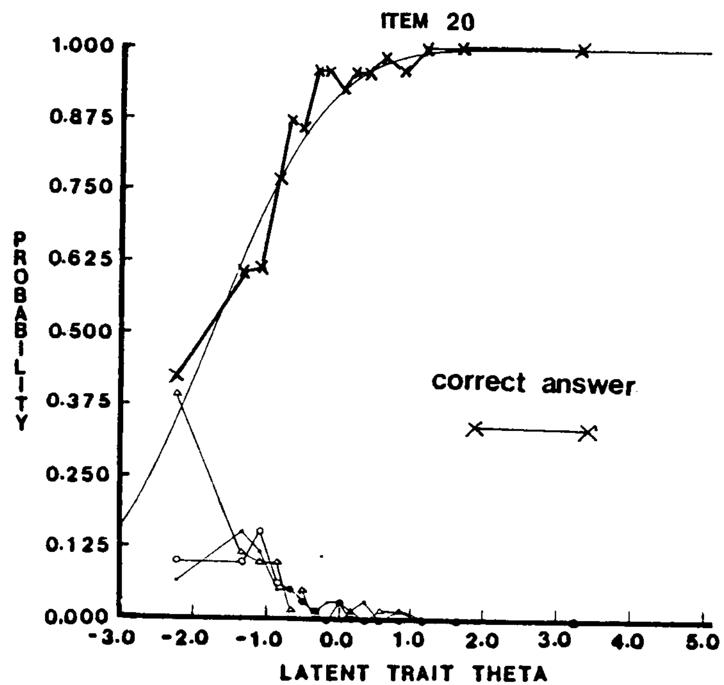


(15) Item 18: $a_g = 0.64417$
$b_g = -1.63039$

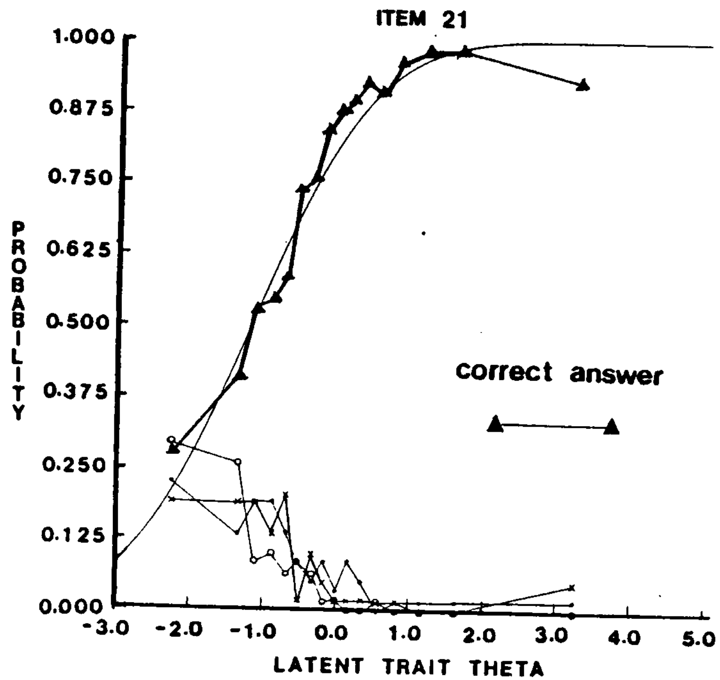


(16) Item 19: $a_g = 0.79181$
$b_g = -1.39692$

Figure D.1. (Continued)

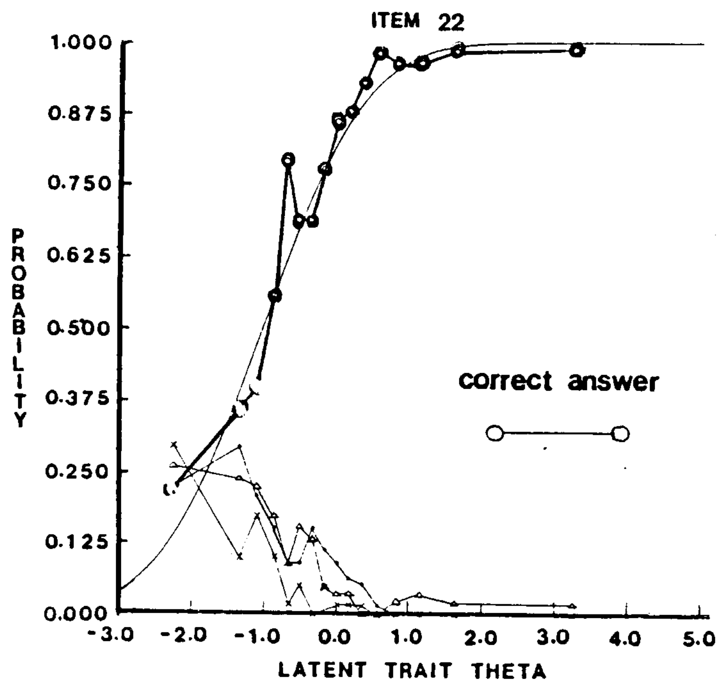


(17) Item 20: $a_g = 0.82510$
$b_g = -1.77217$

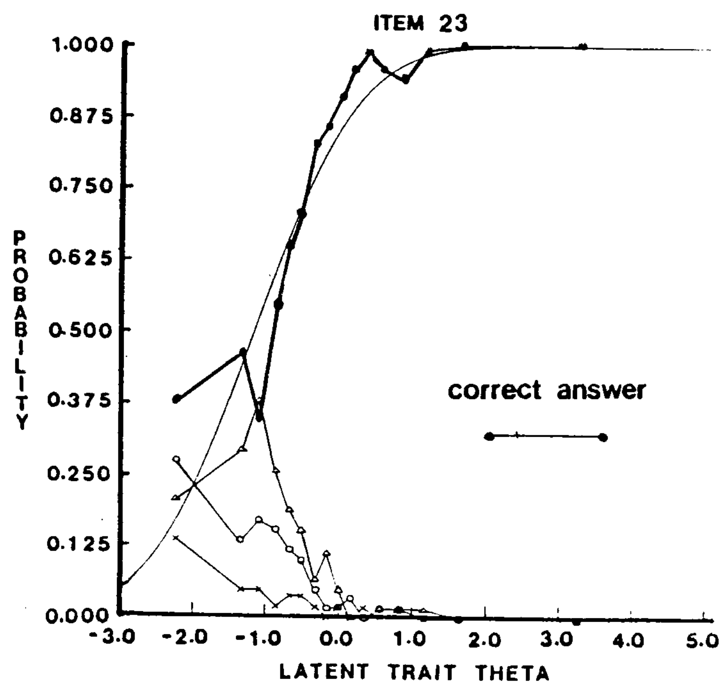


(18) Item 21: $a_g = 0.76905$
$b_g = -1.17517$

Figure D.1. (Continued)

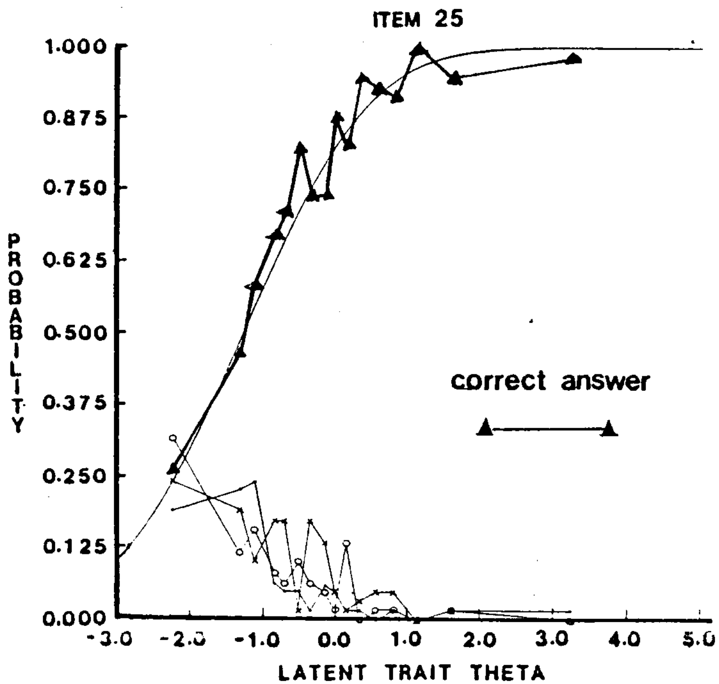


(19) Item 22: $a_g = 0.90055$
$b_g = -1.02436$

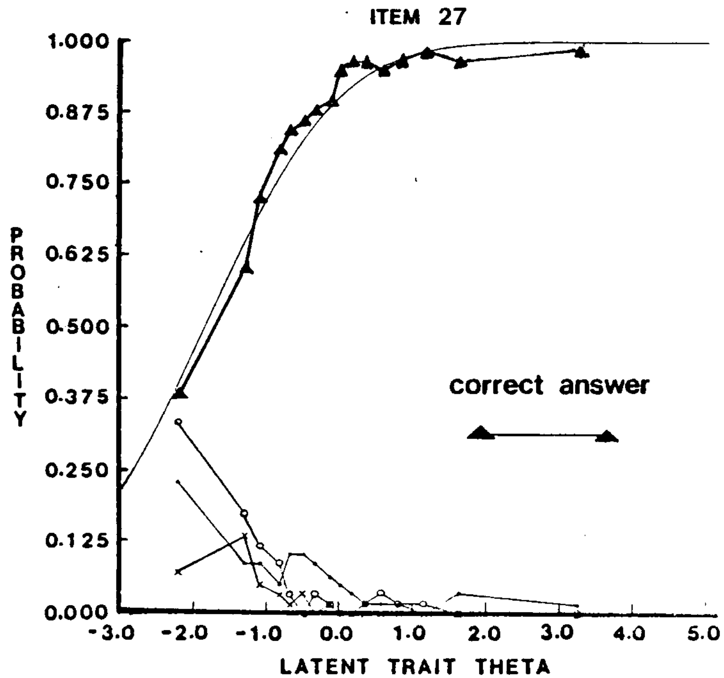


(20) Item 23: $a_g = 0.90933$
$b_g = -1.17630$

Figure D.1. (Continued)

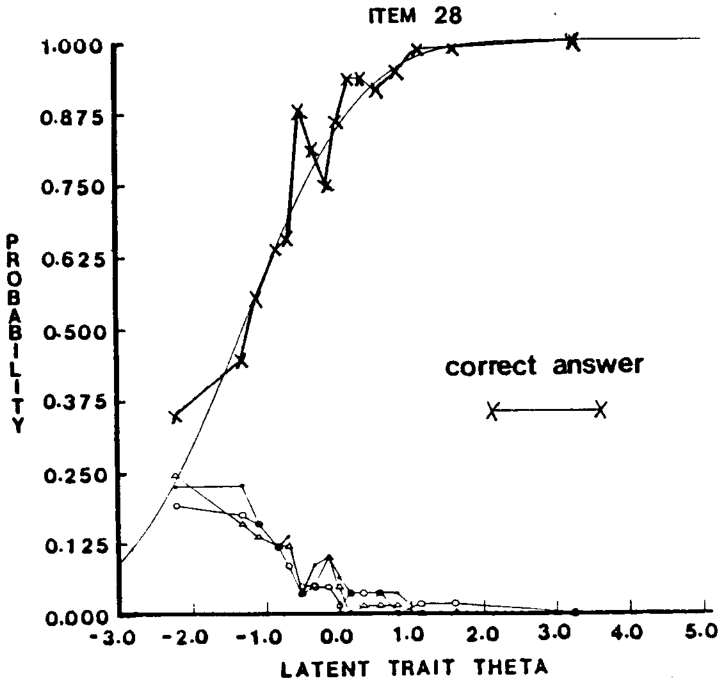


(21) Item 25: $a_g = 0.74713$
$b_g = -1.27364$

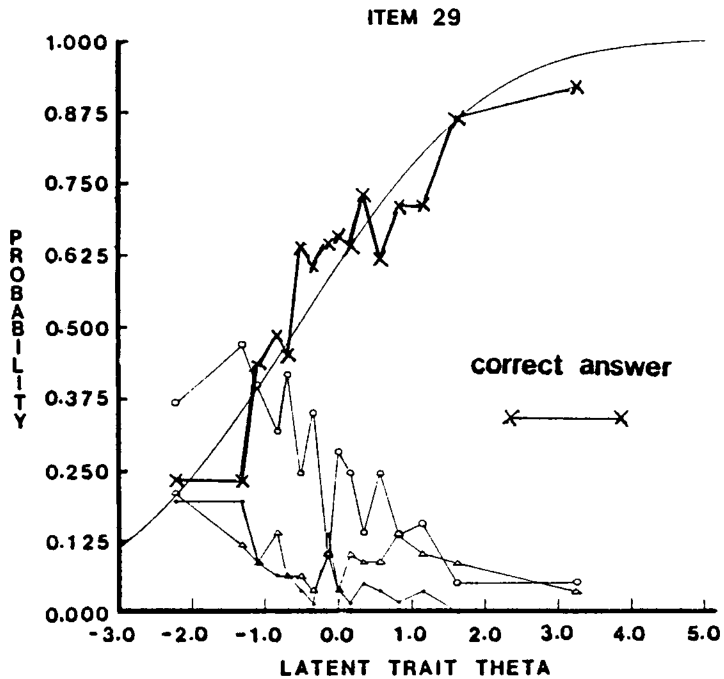


(22) Item 27: $a_g = 0.70663$
$b_g = -1.85227$

Figure D.1. (Continued)

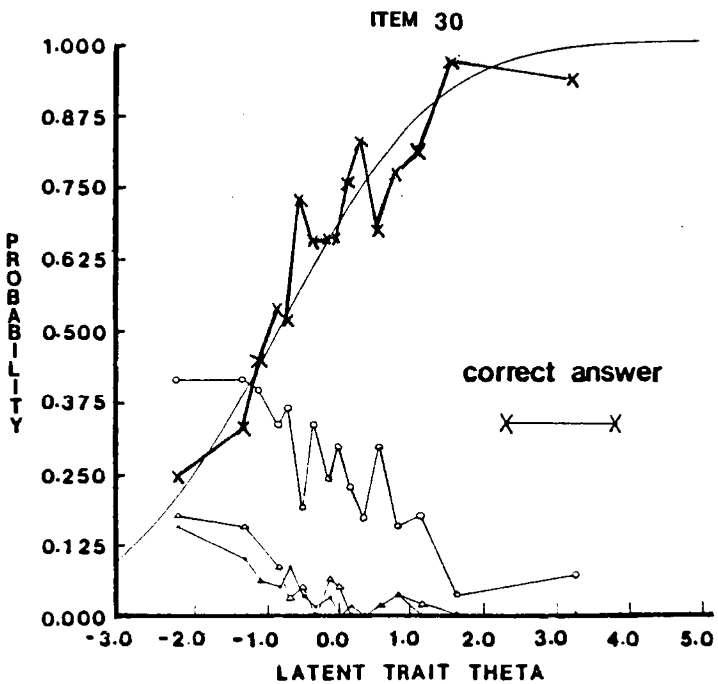


(23) Item 28: $a_g = 0.78575$, $b_g = -1.29880$

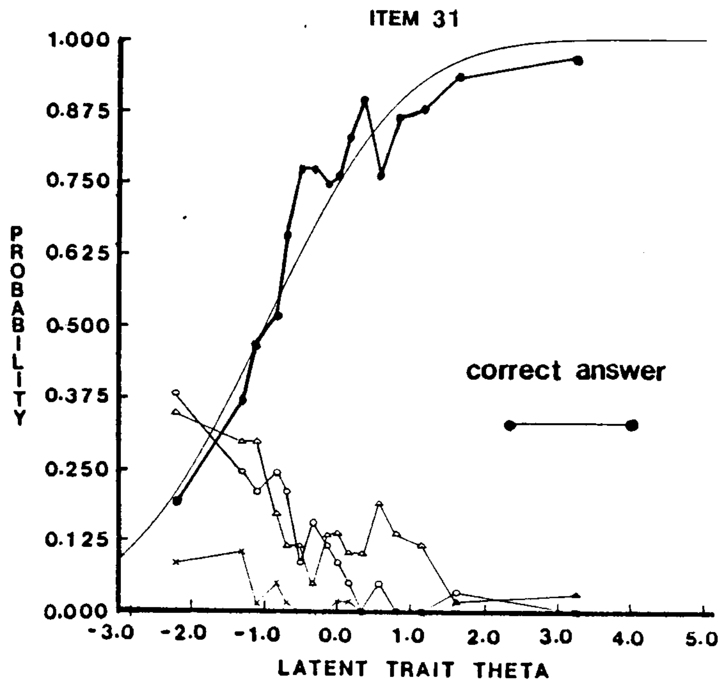


(24) Item 29: $a_g = 0.50051$, $b_g = -0.54870$

Figure D.1. (Continued)

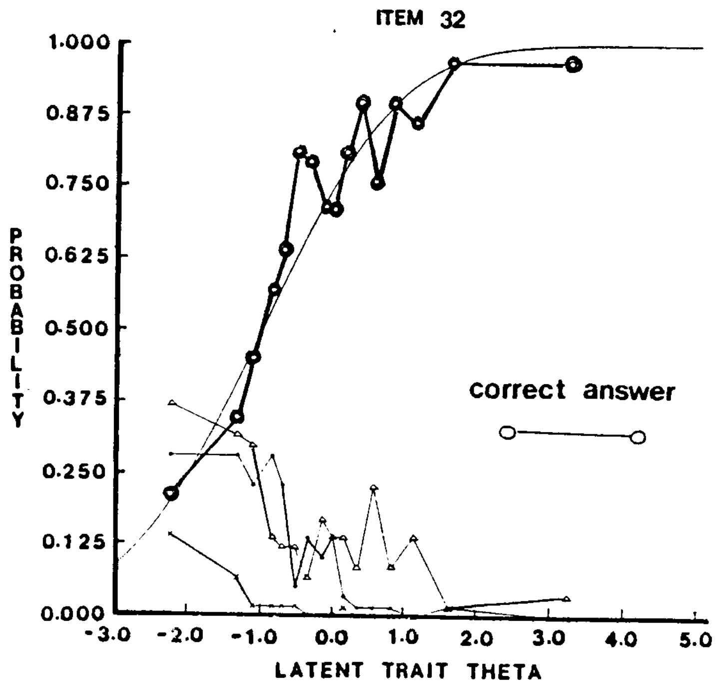


(25) Item 30: a_g = 0.57561
b_g = -0.81637

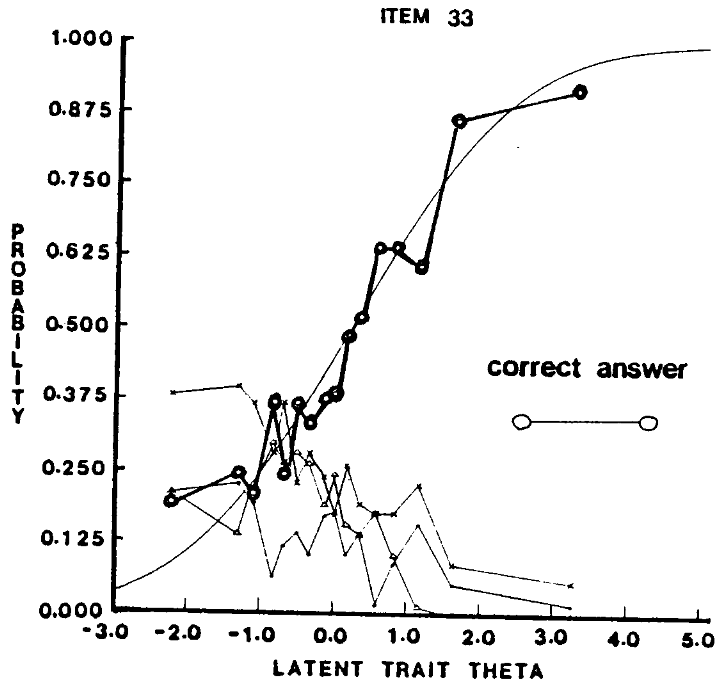


(26) Item 31: a_g = 0.67380
b_g = -0.99914

Figure D.1. (Continued)

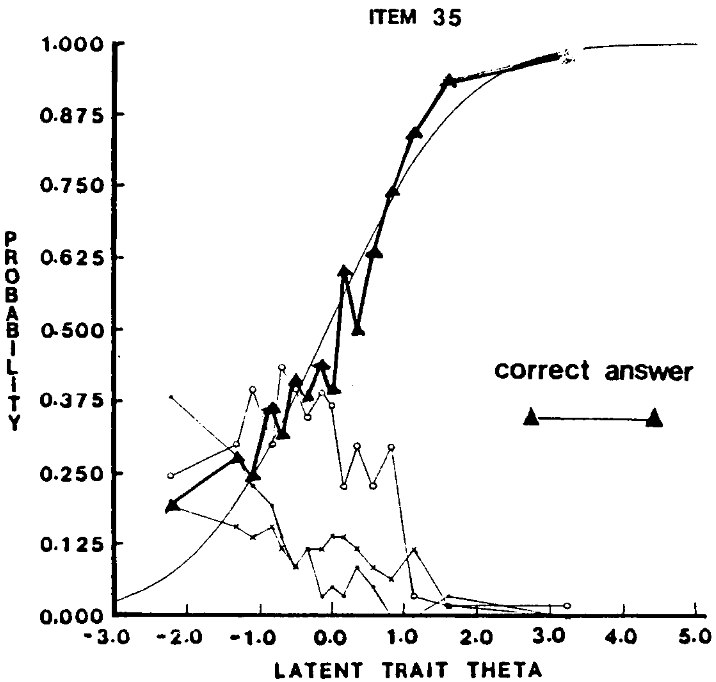


(27) Item 32: $a_g = 0.68947$
$b_g = -0.98911$

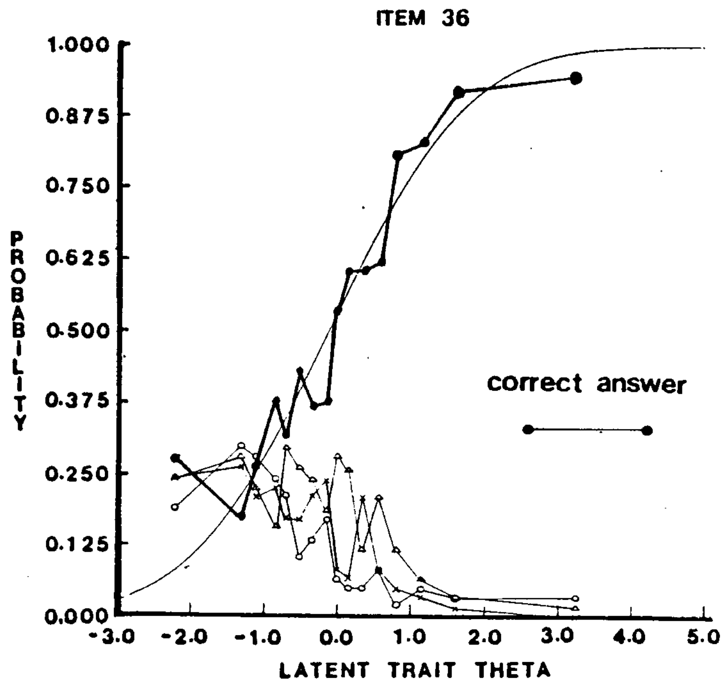


(28) Item 33: $a_g = 0.57035$
$b_g = 0.20370$

Figure D.1. (Continued)

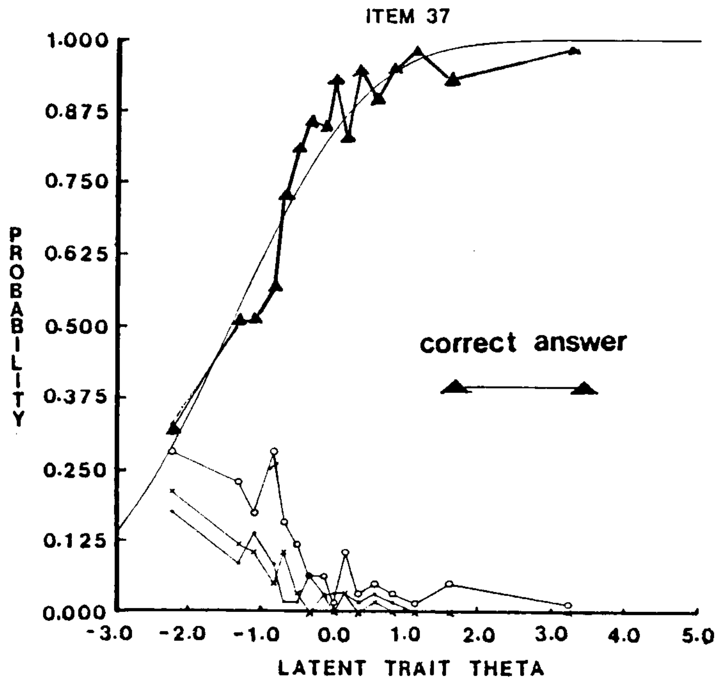


(29) Item 35: $a_g = 0.68177$
$b_g = -0.07395$

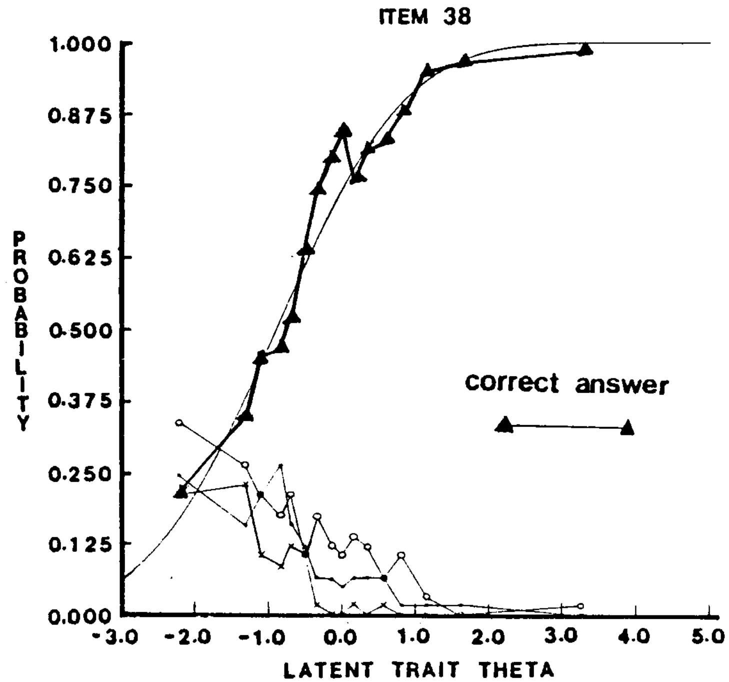


(30) Item 36: $a_g = 0.67825$
$b_g = -0.12215$

Figure D.1. (Continued)

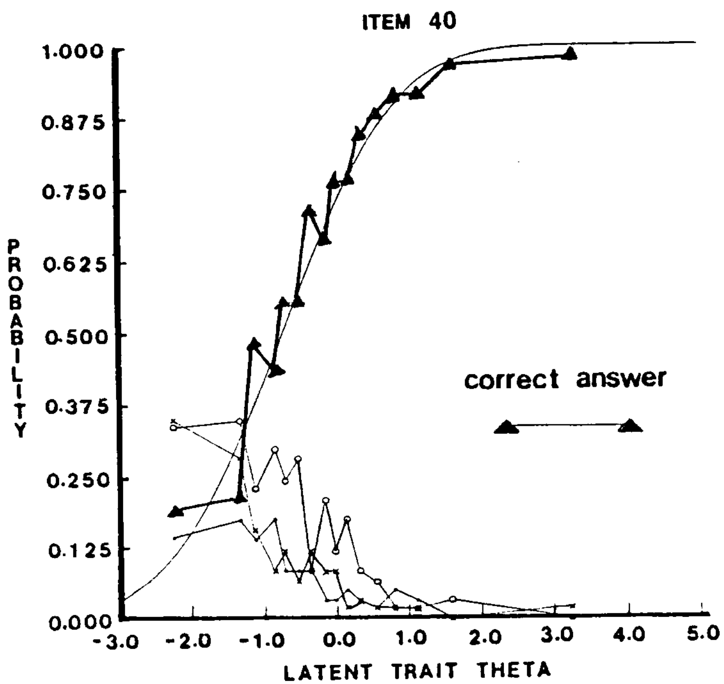


(31) Item 37: $a_g = 0.70357$
$b_g = -1.42047$

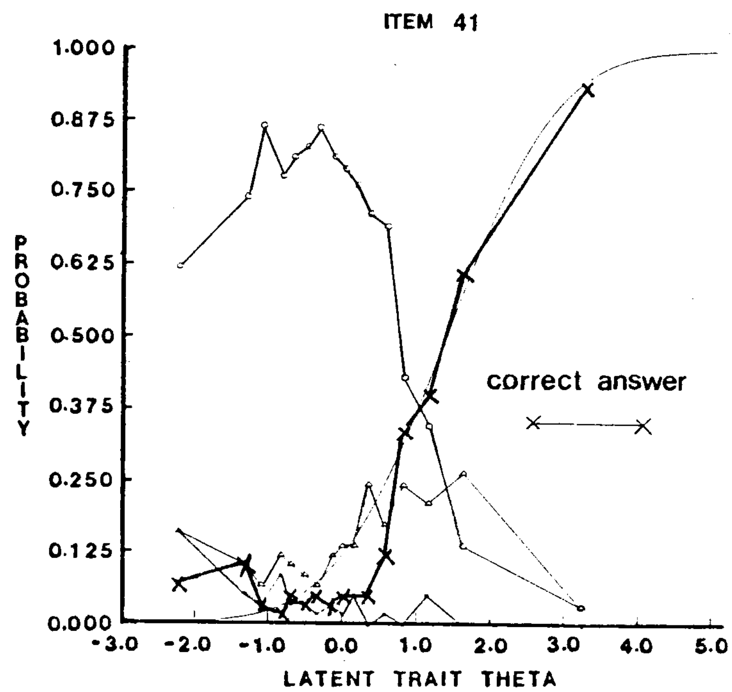


(32) Item 38: $a_g = 0.73640$
$b_g = -0.88384$

Figure D.1. (Continued)

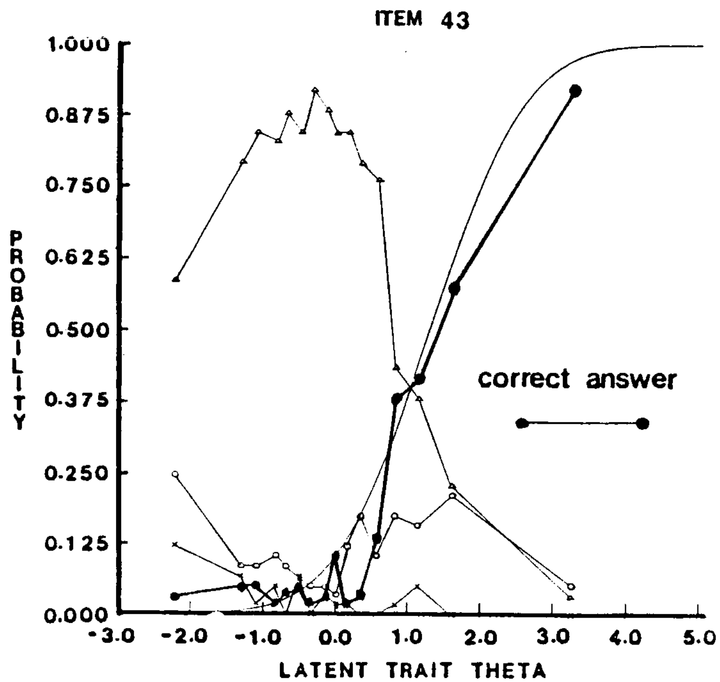


(33) Item 40: $a_g = 0.81883$
$b_g = -0.72200$

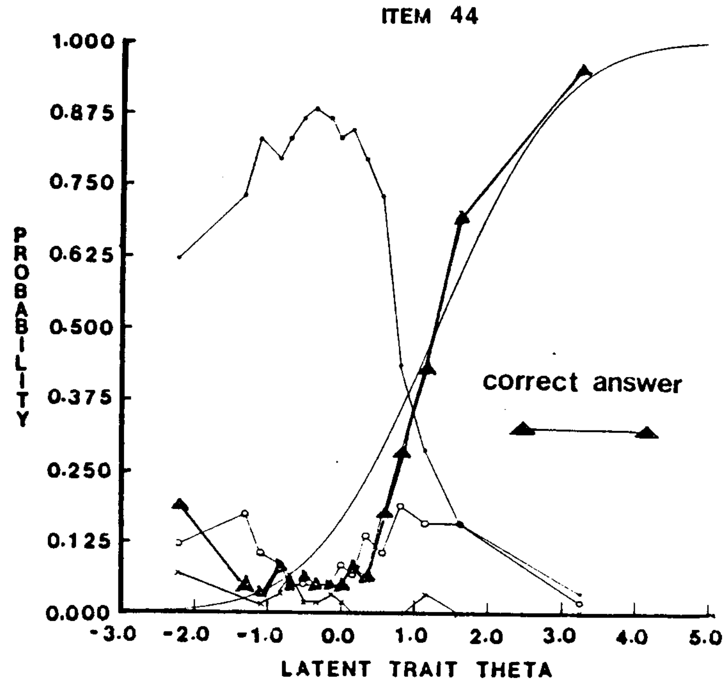


(34) Item 41: $a_g = 0.86166$
$b_g = 1.36480$

Figure D.1. (Continued)

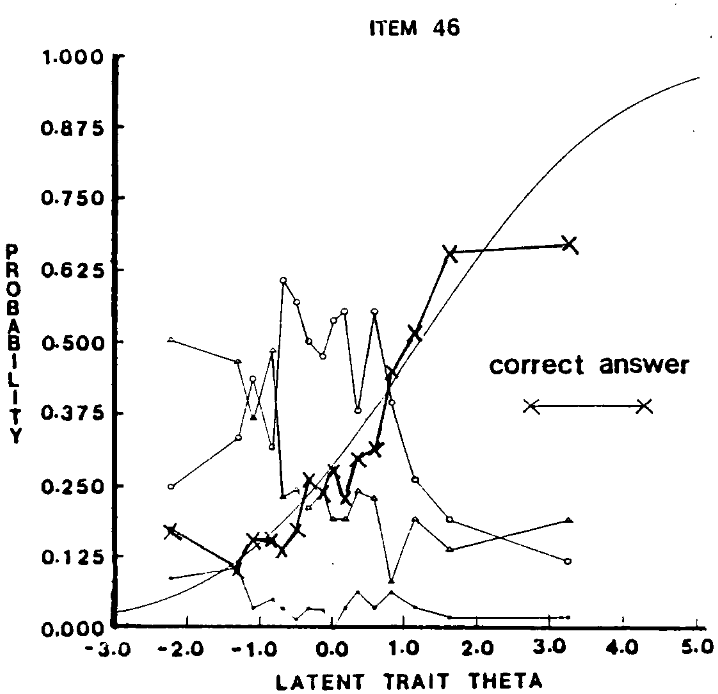


(35) Item 43: a_g = 0.98361
b_g = 1.29919

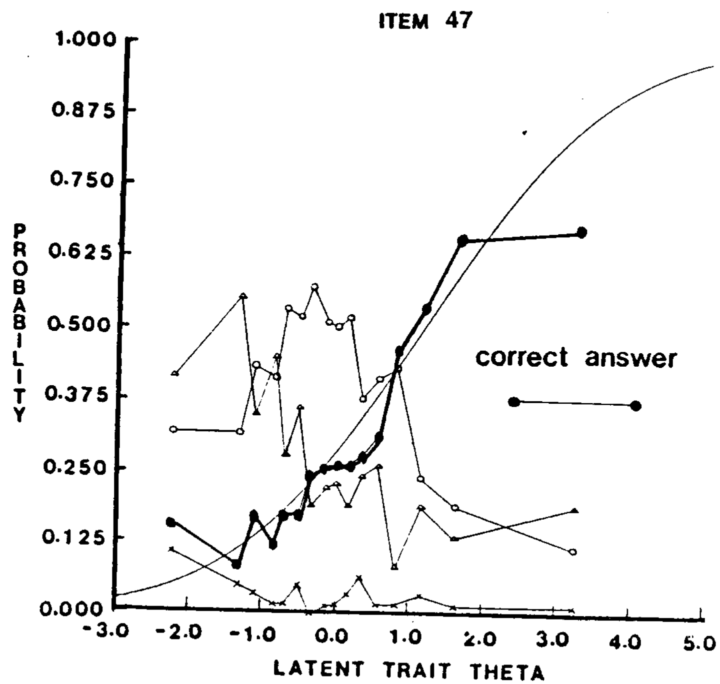


(36) Item 44: a_g = 0.75347
b_g = 1.30898

Figure D.1. (Continued)

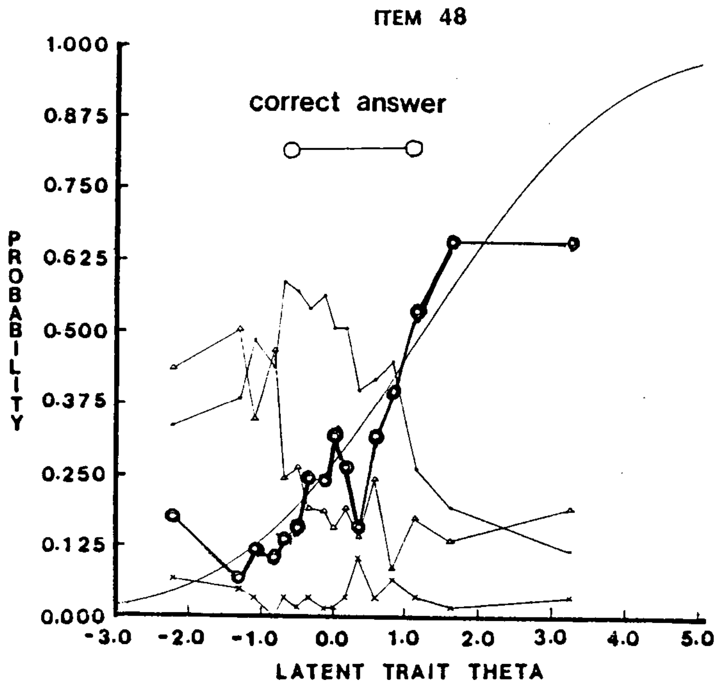


(37) Item 46: a_g = 0.47513
b_g = 1.20689

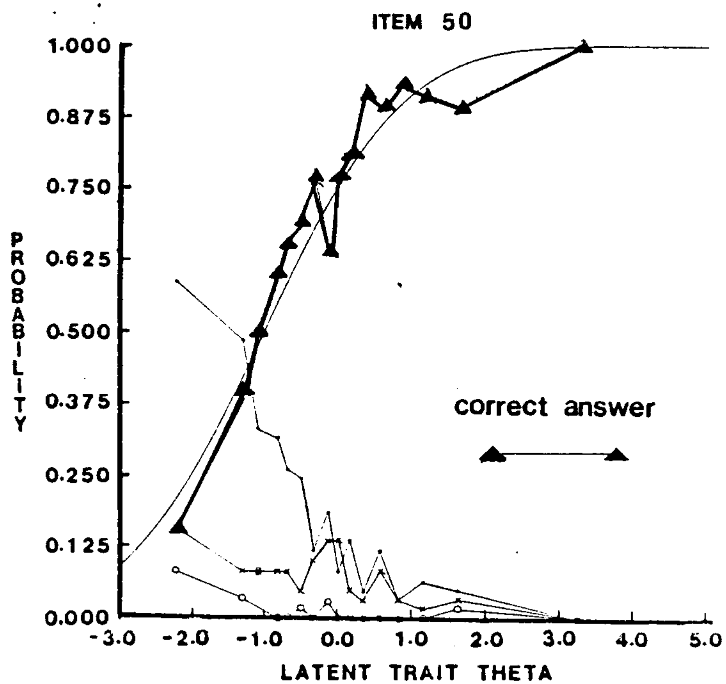


(38) Item 47: a_g = 0.49697
b_g = 1.15690

Figure D.1. (Continued)

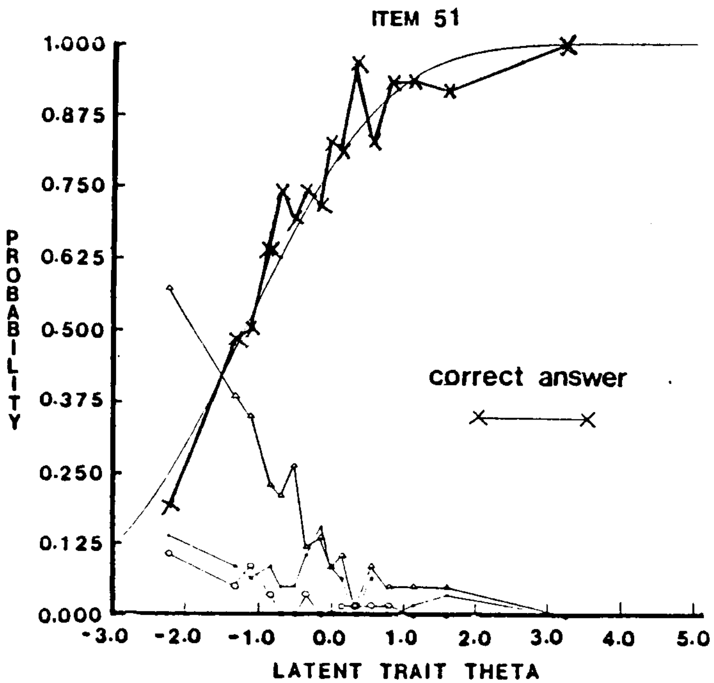


(39) Item 48: $a_g = 0.49641$
$b_g = 1.23451$

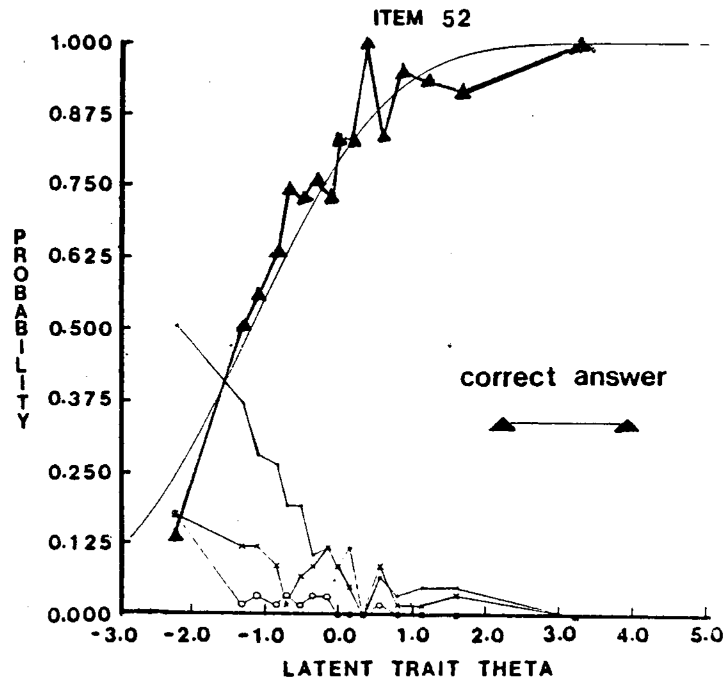


(40) Item 50: $a_g = 0.68707$
$b_g = -1.01375$

Figure D.1. (Continued)

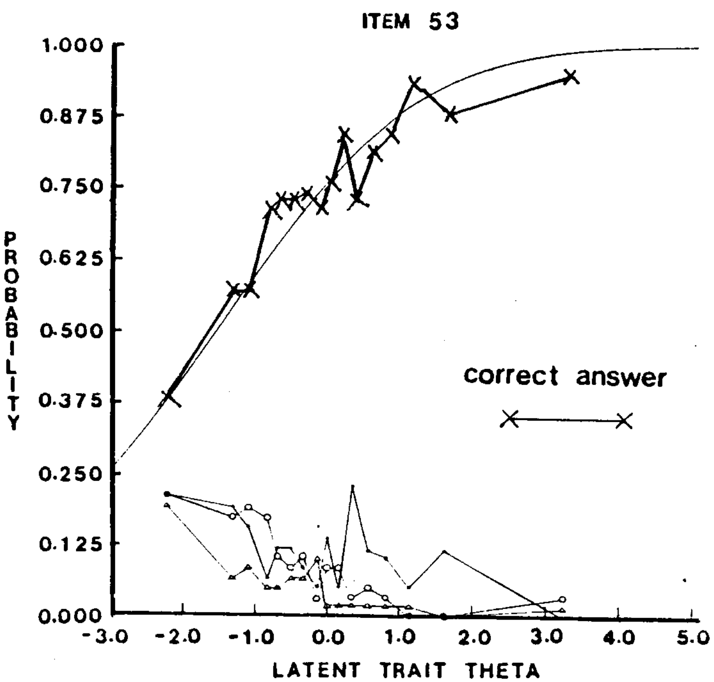


(41) Item 51: $a_g = 0.66771$
$b_g = -1.18594$

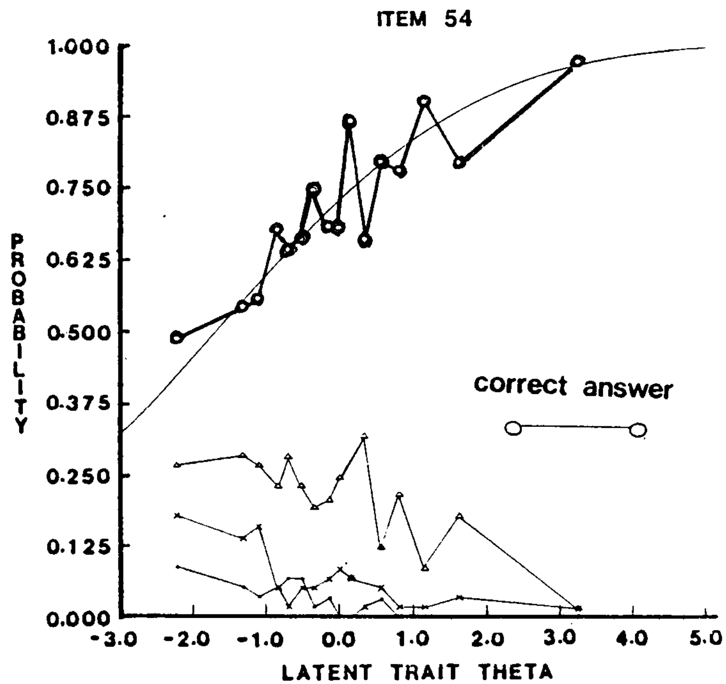


(42) Item 52: $a_g = 0.69535$
$b_g = -1.20668$

Figure D.1. (Continued)

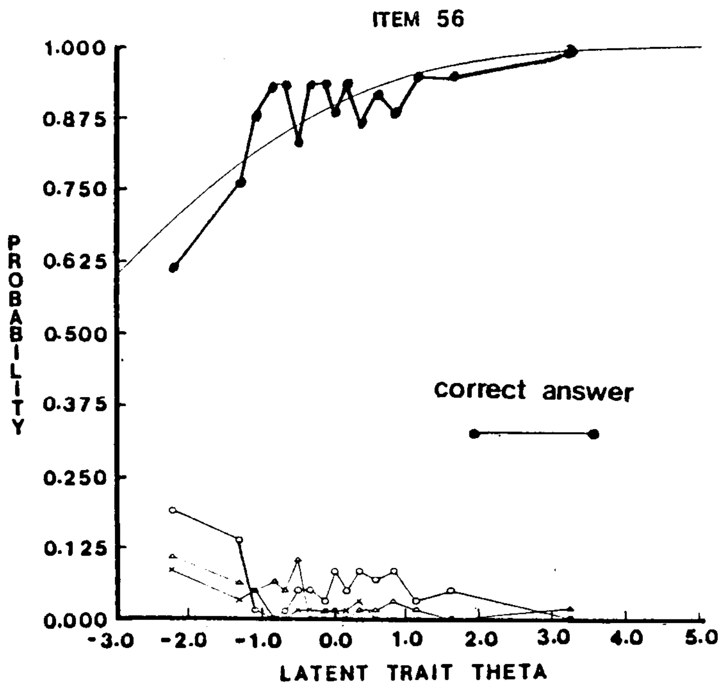


(43) Item 53: $a_g = 0.46553$
$b_g = -1.54463$

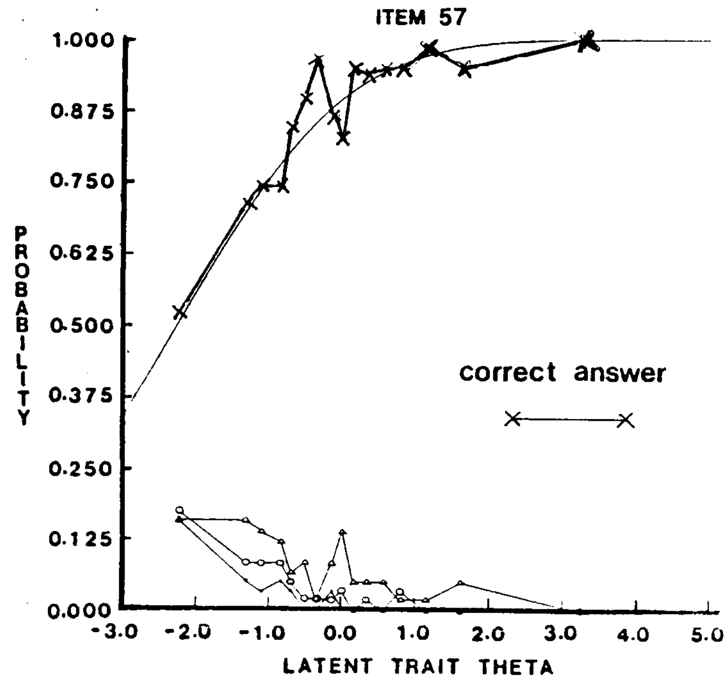


(44) Item 54: $a_g = 0.36016$
$b_g = -1.64765$

Figure D.1. (Continued)

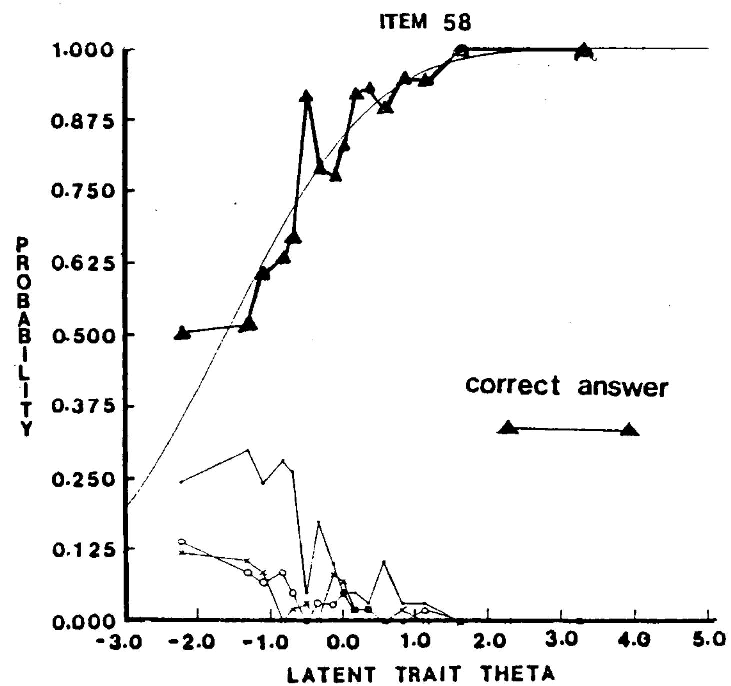


(45) Item 56: $a_g = 0.34198$
$b_g = -3.69762$

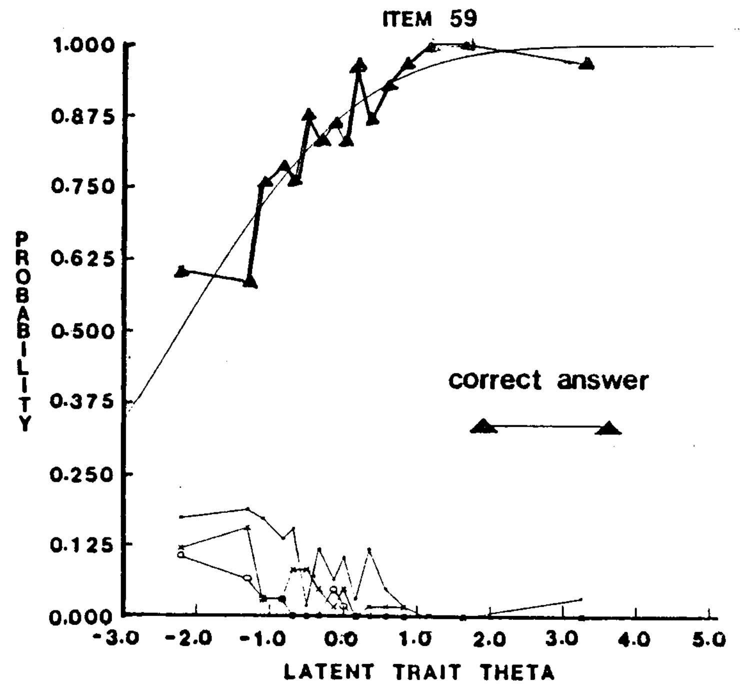


(46) Item 57: $a_g = 0.55957$
$b_g = -2.23833$

Figure D.1. (Continued)

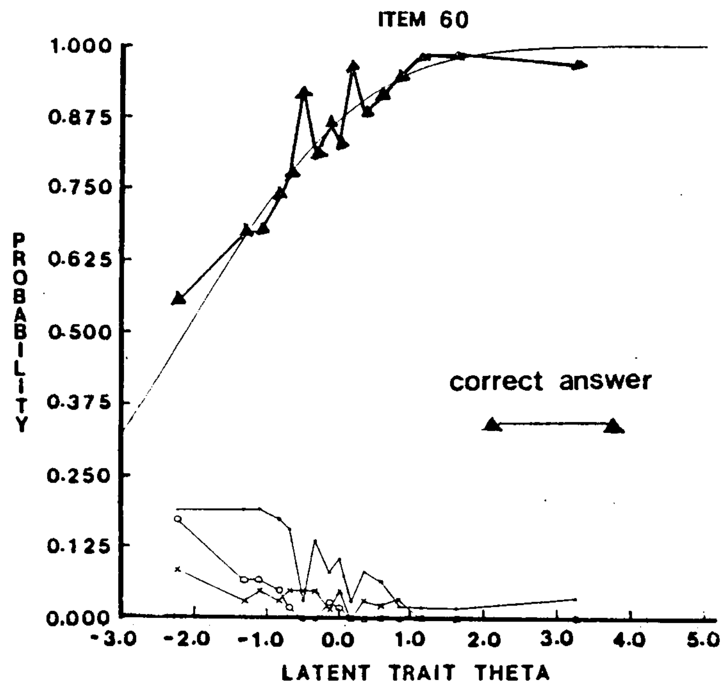


(47) Item 58: $a_g = 0.63581$
$b_g = -1.60893$

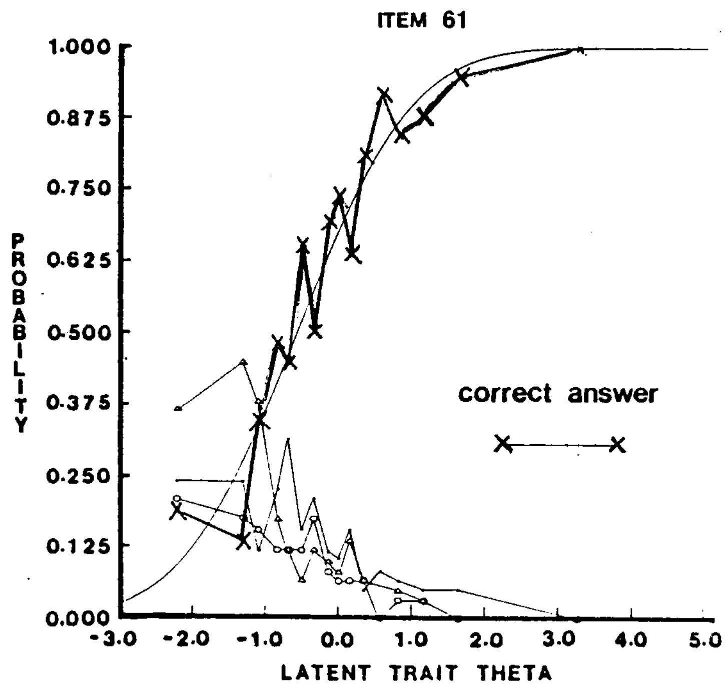


(48) Item 59: $a_g = 0.52855$
$b_g = -2.21747$

Figure D.1. (Continued)

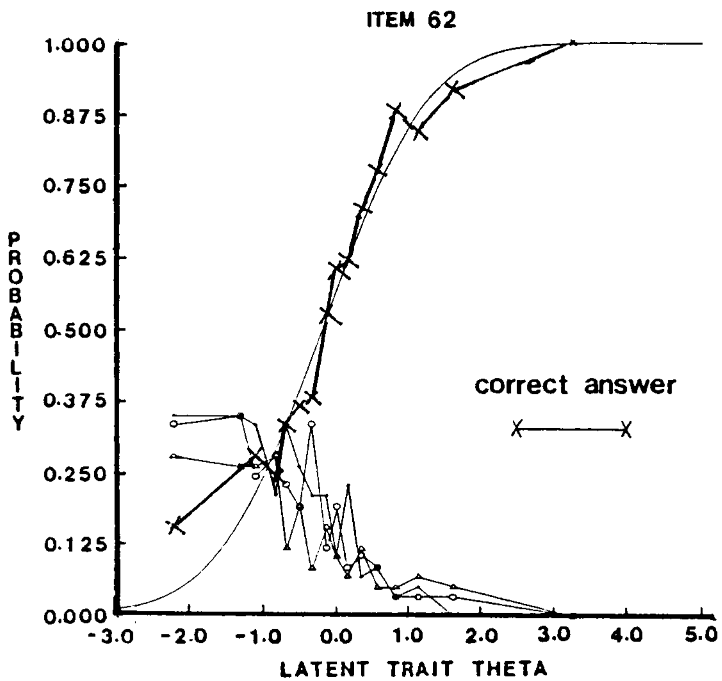


(49) Item 60: $a_g = 0.53741$
$b_g = -2.11270$

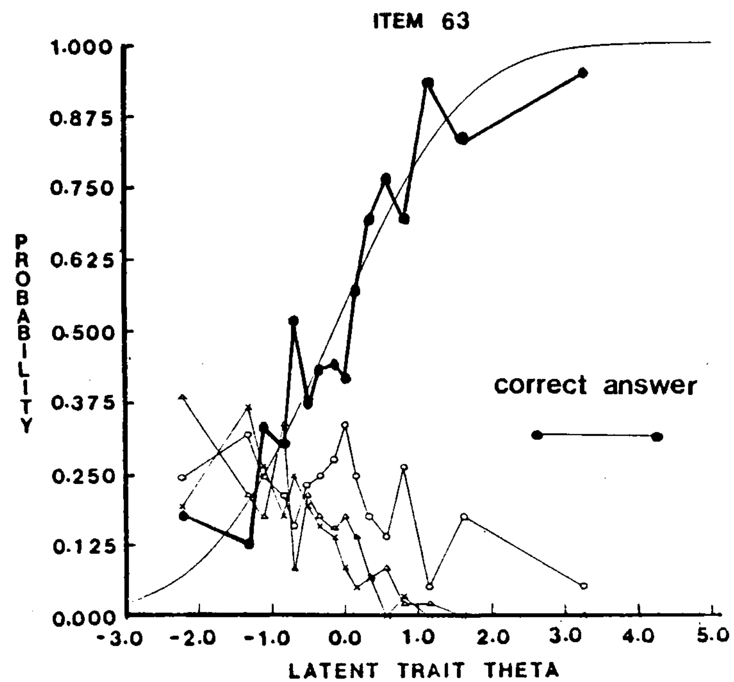


(50) Item 61: $a_g = 0.81918$
$b_g = -0.56964$

Figure D.1. (Continued)

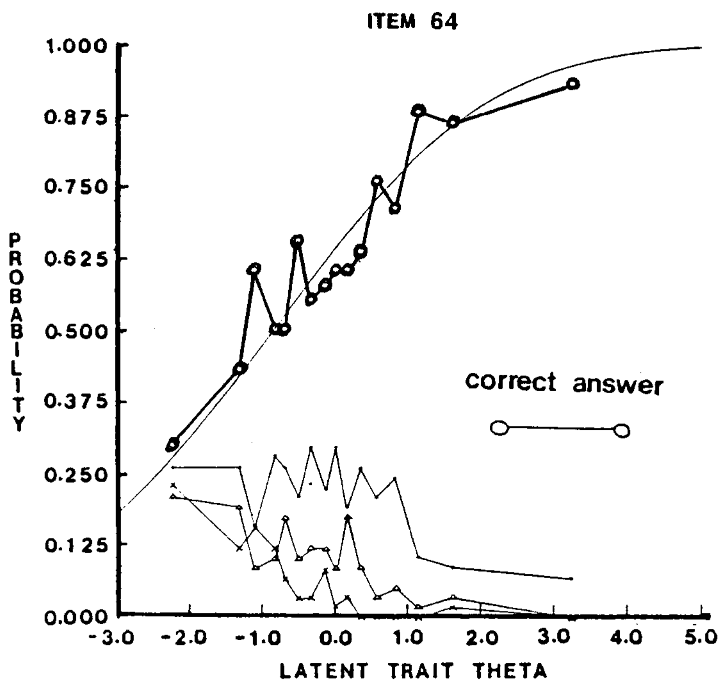


(51) Item 62: $a_g = 0.87609$
$b_g = -0.19008$

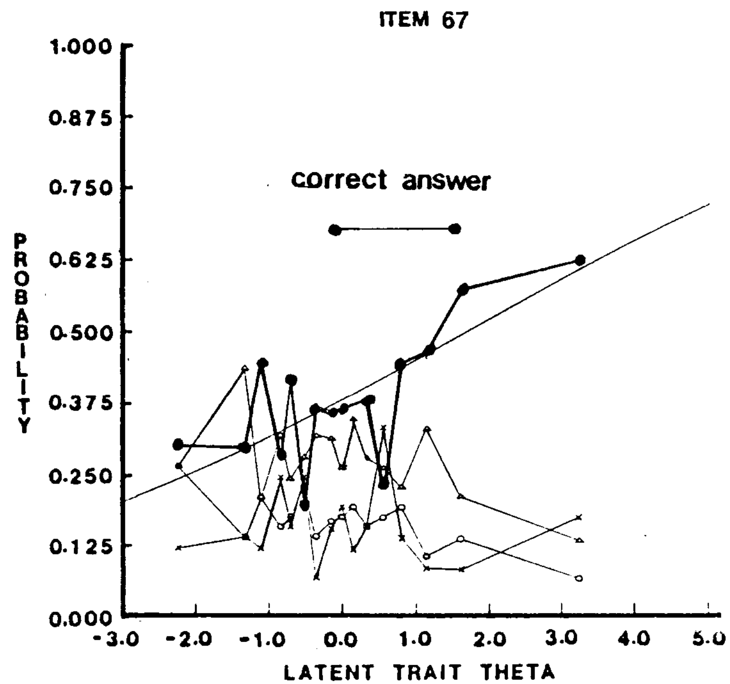


(52) Item 63: $a_g = 0.71909$
$b_g = -0.14051$

Figure D.1. (Continued)

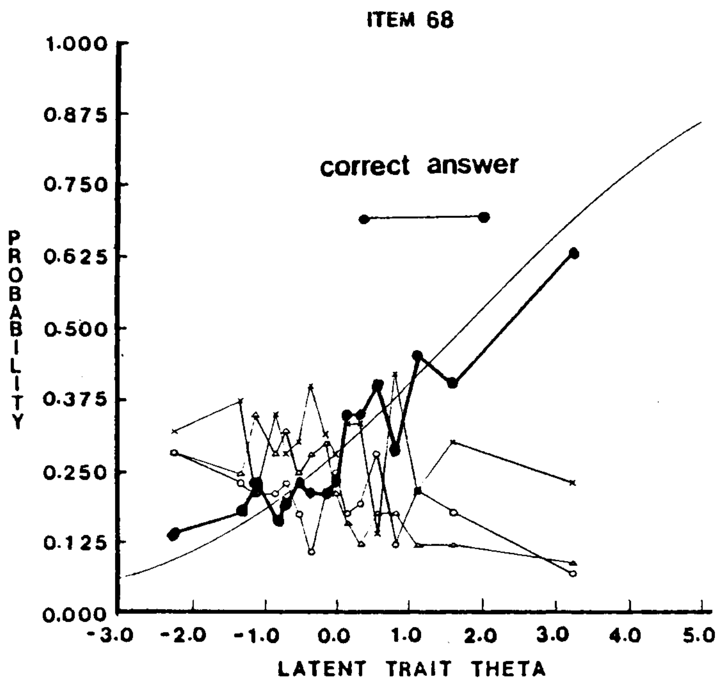


(53) Item 64: $a_g = 0.43352$
$b_g = -0.85012$

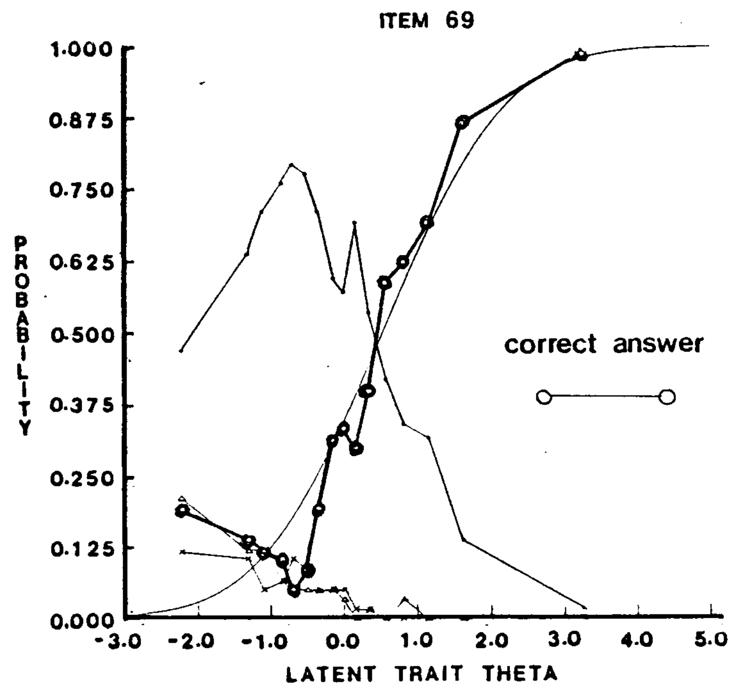


(54) Item 67: $a_g = 0.17978$
$b_g = 1.71915$

Figure D.1. (Continued)

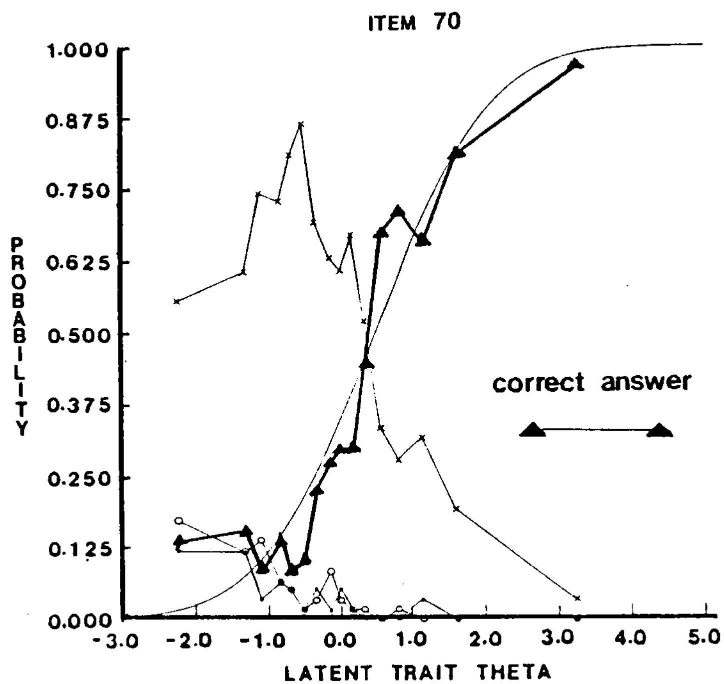


(55) Item 68: a_g = 0.33048
b_g = 1.77924

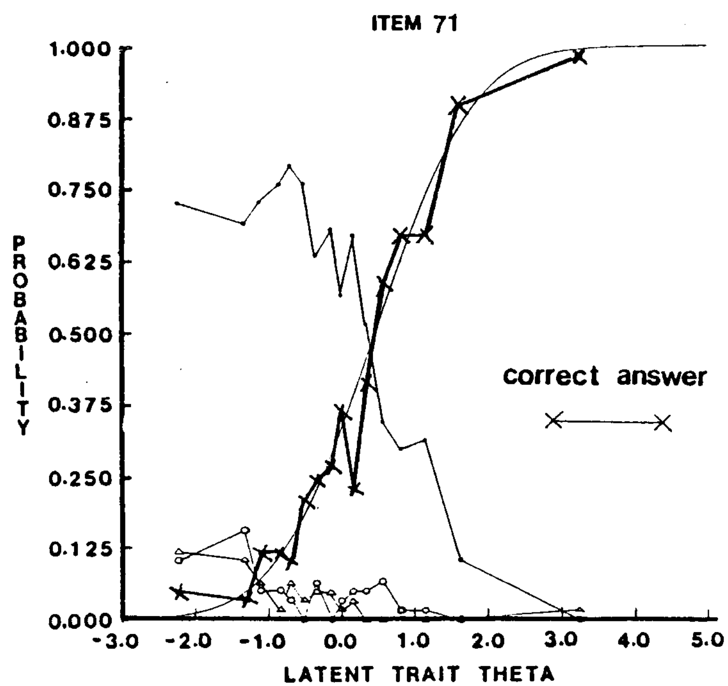


(56) Item 69: a_g = 0.75653
b_g = 0.53693

Figure D.1. (Continued)

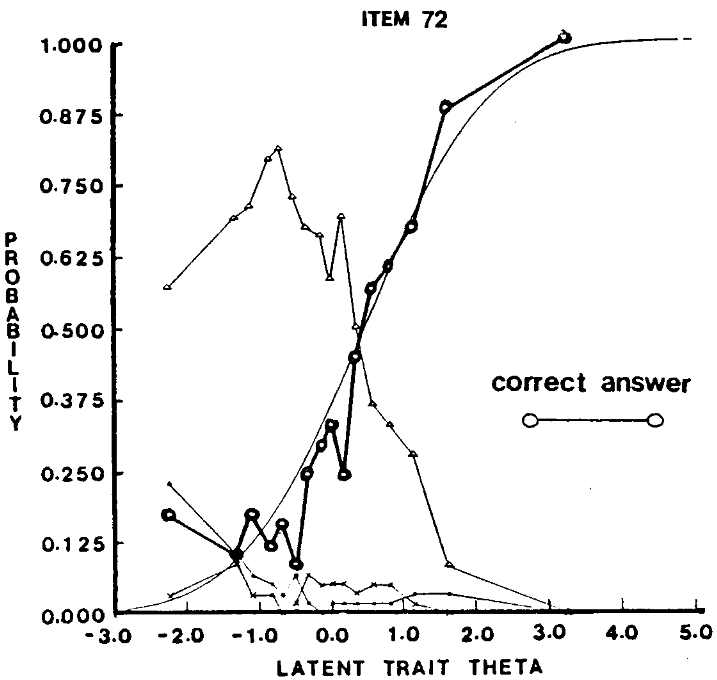


(57) Item 70: $a_g = 0.79630$
$b_g = 0.49283$

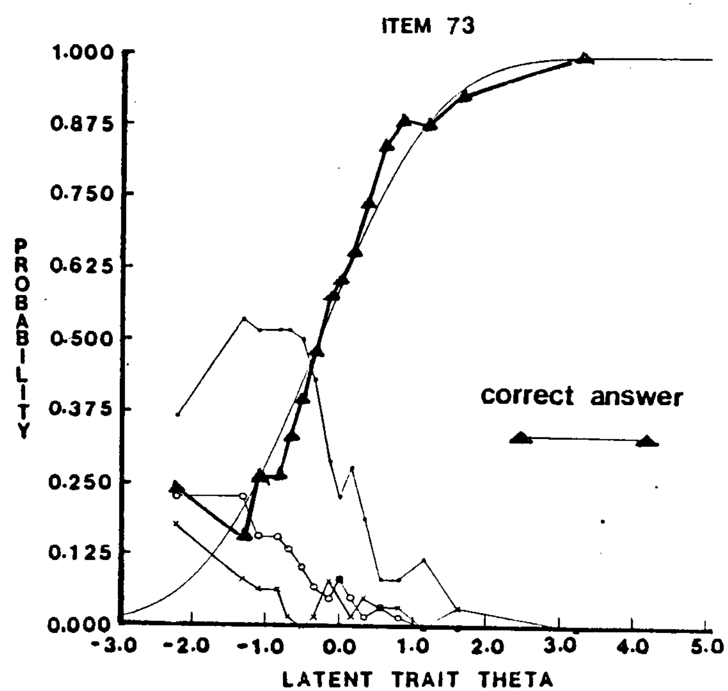


(58) Item 71: $a_g = 0.93477$
$b_g = 0.47023$

Figure D.1. (Continued)

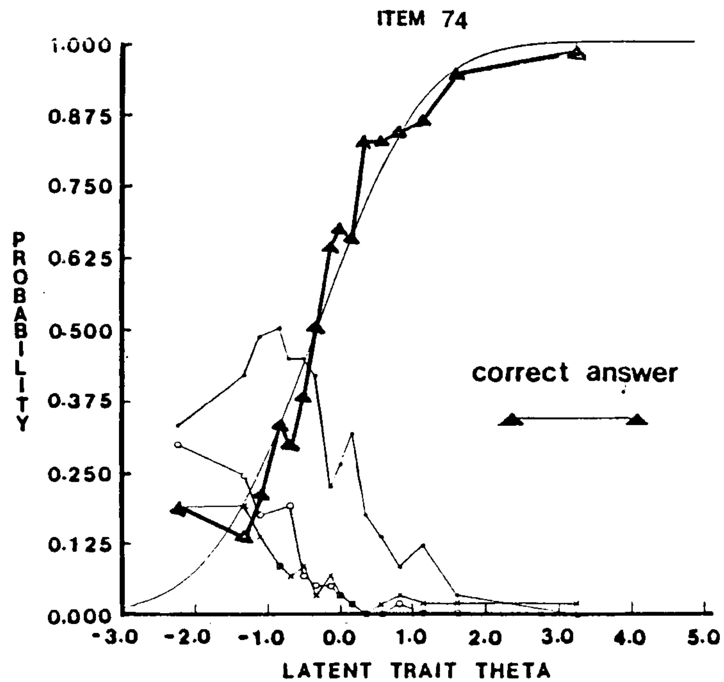


(59) Item 72: a_g = 0.73656
b_g = 0.47509

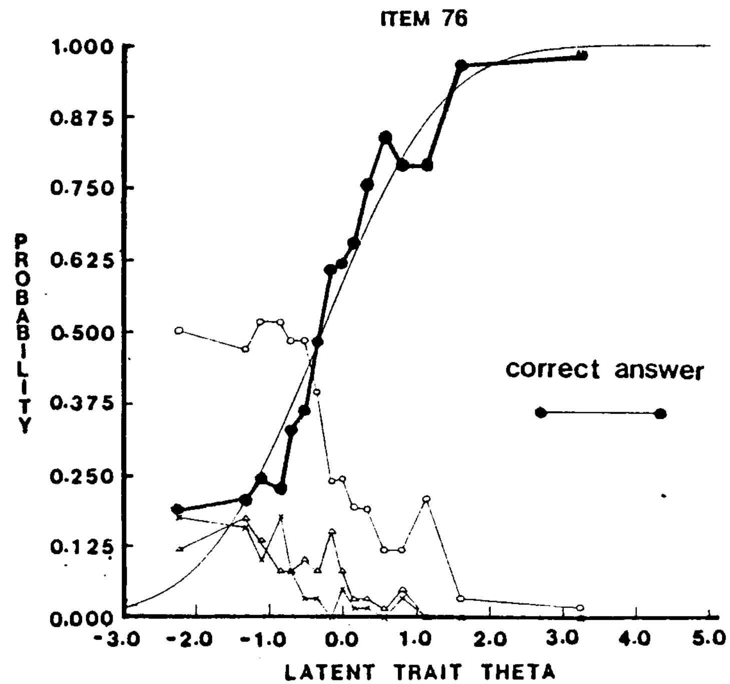


(60) Item 73: a_g = 0.82295
b_g = -0.30418

Figure D.1. (Continued)



(61) Item 74: $a_g = 0.86433$
$b_g = -0.32075$



(62) Item 76: $a_g = 0.79176$
$b_g = -0.27179$

Figure D.1. (Continued)

VITA

Megumi Asako was born in Peking, China on September 25, 1945. When she was seven months old, she moved to her country. Japan. She attended both grade school and high school in Nagasaki. Japan. She graduated with the Associate of Arts degree in Business Administration and a teaching certificate from the International College of Business. After graduation she taught at city schools in Nagasaki. Japan for two years (1966-1968). In 1968 she established and managed her own school, Asako Preparatory School for the students with ages between 10 and 16 until she started her education in the United States in 1976. After she received the Bachelor of Arts degree in Pychology from LaGrange College in LaGrange. Georgia in 1978, she began her master's program at Austin Peay State University in Clarksville, Tennessee in 1978. Upon graduation from Austin Peay State University with the Mater of Arts in Psychology in 1979, she started her doctoral program at the University of Tennessee in Knoxville. Tennessee. During her doctoral program she taught disadvantaged students from foreign countries in the Knoxville City School System from 1980 to 1981 and also has been working as a graduate assistant for the Bureau of Educational Research and Service. She is married to Yoji Takata from Japan.