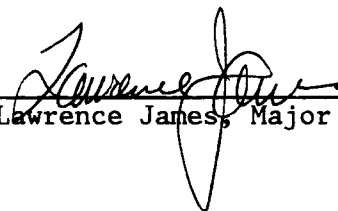
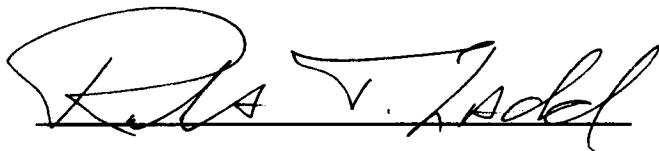


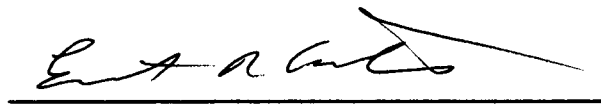
To the Graduate Council:

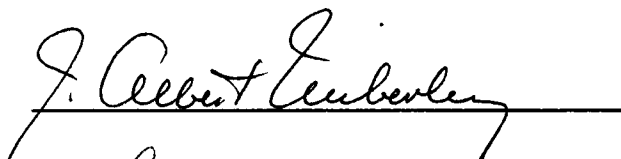
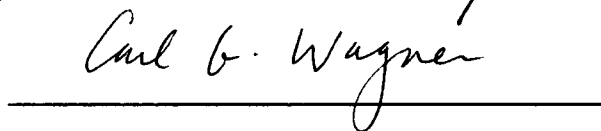
I am submitting herewith a dissertation written by Paul S. Chagas entitled "An Investigation of the Information Provided by the Distractors of Multiple-choice Test Items." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Psychology.


Lawrence James, Major Professor

We have read this dissertation
and recommend its acceptance:


Karl A. V. Isold


E. A. Smith


J. Albert Leubert

Carl B. Wagner

Accepted for the Council:


Associate Vice Chancellor
and Dean of The Graduate School

AN INVESTIGATION OF THE INFORMATION PROVIDED BY
THE DISTRACTORS OF MULTIPLE-CHOICE TEST ITEMS

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

Paul S. Chagas

December 1991

Copyright © Paul S. Chagas, 1992,

All rights reserved

DEDICATION

This study is dedicated to the memory of
my father,
who worked so hard to provide his sons
with the educational opportunities
that he was not afforded.

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to Dr. Lawrence James. It was his guidance and direction that allowed this study to reach its completion. I feel very fortunate to have had someone so capable and so concerned for students serve as my major professor.

I must thank everyone on my doctoral committee for their patience and endurance as well as their contributions to this dissertation. I am especially grateful to those committee members and other educators who participated in the evaluation of test item content. This includes Dr. Schuyler Huck, who has provided much encouragement in addition to valuable suggestions during the course of this study.

There are many individuals at the State Testing and Evaluation Center who have been supportive. I would be remiss if I did not mention Angelia Golden, Director, and Dr. Fretta "Skeeter" Bunch, Associate Director. I am very thankful to have them as friends as well as colleagues.

My family has been a tremendous source of strength for me. For all of their love and support, I must take this opportunity to thank my mother, my brother, George, and especially my wife, Kathy. I owe them all more than they will ever know.

ABSTRACT

Previous studies have indicated that the knowledge or random guessing principle is often contradicted by distractor analysis of multiple-choice test data. This study attempted to analyze item content to better describe those distractors providing differential information with regard to achievement level. Item Response Theory (IRT) methods and approaches were applied to response data for the Vocabulary subtest of the Iowa Tests of Basic Skills. Informative distractors were identified from the resultant plausibility functions, which express the relationship between achievement level and choice of distractor.

Although considerable agreement in the prediction of informative distractors was found among educators asked to review item content, there was only modest success in the accuracy of these predictions. Content-based rules or reasons that would lead to the selection of a particular distractor were identified, and indications of a relationship between these rules and achievement level were found. However, much more research is needed to understand this link between distractor content and selection. Implications for item writers and test developers were discussed.

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION.	1
Overview and Objectives	1
Background.	3
Knowledge or random guessing principle.	3
Diagnostic information from distractors	9
Testwiseness and rules of item writing.	12
Scoring and testing procedures to assess partial knowledge	17
Background of the Iowa Testing Programs	20
Iowa Tests of Basic Skills, Form 6.	23
II. DATA SOURCE AND METHODS	24
Subjects.	24
Content Analysis and Prediction of Informative Distractors	26
Latent Trait Theory and Applications to Multiple-choice Test Items.	35
Background.	35
Unidimensional latent trait theory.	36
Mathematical models for the item characteristic function.	39
Maximum likelihood estimation of ability.	41
Information functions and the error of estimation.	42
Plausibility functions.	43
Latent Trait Procedures	44
Tetrachoric method item parameter estimation.	44
Ability estimation and transformation of the latent trait.	48
Estimation of the plausibility functions.	50
III. RESULTS	52
Content Analysis.	52
Latent Trait Analysis	64
Tetrachoric correlation	64
Factor analysis	66
Item parameter computation.	67
Ability estimation.	70
Information functions and transformation of the latent trait.	73
Plausibility functions.	73
Estimated item characteristic functions	129
Frequency ratios.	130

CHAPTER	PAGE
III. (Continued)	
Comparison of Content-based and Latent Trait	
Results	134
Linking Content-based Rules to Examinee Ability . .	144
IV. CONCLUSIONS	171
Summary and Implications.	177
LIST OF REFERENCES.	184
APPENDICES.	191
APPENDIX A.	192
APPENDIX B.	206
VITA.	215

LIST OF TABLES

TABLE	PAGE
II.1. Number of Vocabulary Items in Each of Levels 11, 12, and 13, Number of Overlapping Items Between Levels, and Number of Items Appearing in Only One Level	27
II.2. Vocabulary Items Selected for Rankings of Distractors by Faculty/Teacher Evaluators	32
III.1. Predictions by Author of Informative Distractors for Each of 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest.	53
III.2. Mean and Standard Deviation of Evaluator Rankings, by Item Alternative, and Kendall's Coefficient of Concordance, W, for 16 Selected Items of Vocabulary Subtest.	60
III.3. Estimated Item Parameters for Each of 43 Items of Level 11 Vocabulary Subtest	69
III.4. Estimated Item Parameters for Each of 44 Old Test Items of Level 12 Vocabulary Subtest.	71
III.5. Estimated Item Parameters for Each of 45 Old Test Items of Level 13 Vocabulary Subtest.	72
III.6. Observed Informative Distractors for 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest. . . .	116
III.7. Informative Distractors Observed for Each Level and for Overall Results, and Identification of Levels for Which the Plausibility Functions Show a Consistent Pattern, for All Items Administered in Two or More Levels.	122
III.8. Comparison of Author's Predictions of Informative Distractors to Observed Results for 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest. . . .	135
III.9. Mean of Rankings by Author for Various Categories of Distractors with Regard to Information Provided, and Theoretical Rankings for Each Category if Perfectly Predicted	138

TABLE	PAGE
III.10. Number of Times Each Rule Identified from Content Analysis of Vocabulary Subtest Appeared as Possible Reason for Selection of Distractors Observed to Be Informative.	140
III.11. Comparison of Predicted Informative Distractors Obtained by Faculty, Teachers and Author to Observed Results for 16 Selected Items of Vocabulary Subtest.	142
III.12. Mean of Rankings by Faculty, Teachers and Author for Various Categories of Distractors with Regard to Information Provided, and Theoretical Rankings for Each Category if Perfectly Predicted	143
III.13. Rule Numbers Corresponding to Each Distractor for the 82 Items of the Vocabulary Subtest.	145
III.14. Number of Distractors Corresponding to Each Rule, and the Number and the Percent Which Are Informative, for 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest.	148
III.15. Number of Items and Mean Item Difficulty Parameter for Items Having No Informative Distractor, One Informative Distractor, and Multiple Informative Distractors, for Each of Levels 11, 12 and 13 of Vocabulary Subtest.	150
III.16. Plausibility Range and Peak of Plausibility Function of Each Informative Distractor, for Each of Levels 11, 12 and 13 of Vocabulary Subtest	152
III.17. Mean Value of Ability, θ , at the Upper Bound and at the Midpoint of the Plausibility Range and at the Peak of the Plausibility Function, by Rule, for Informative Distractors Corresponding to a Single Rule, for Levels 11, 12 and 13 of Vocabulary Subtest.	154
III.18. Mean Value of Ability, θ , at the Upper Bound and at the Midpoint of the Plausibility Range and at the Peak of the Plausibility Function, by Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest.	156

TABLE	PAGE
III.19. Multiple Regression Coefficients Corresponding to Each Rule, When the Upper Bound or the Midpoint of the Plausibility Range Is a Function of the Specific Rules Pertaining to the Informative Distractor, for Each of Levels 11, 12 and 13 of Vocabulary Subtest.	163
III.20. Percent Usage, As Determined from Hillerich's (1978) Writing Vocabulary, of the Least Frequently Used Word in Each Response Alternative, for Each of 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest	166
III.21. The Number of Items for Which the Key Word Represents a Particular Part of Speech, the Number and Percentage Having Informative Distractors, and the Mean Item Difficulty Parameter for Each of Levels 11, 12 and 13, by Part of Speech.	170
A.1. Mean of Maximum Likelihood Estimate of Ability, $\hat{\theta}$, and Number of Examinees in Each of 30 Subgroups Used in Obtaining Frequency Ratios to Approximate Plausibility Functions.	193
A.2. Comparison of Informative Distractors Identified from Plausibility Functions to Those Identified from Frequency Ratios for Level 11 Vocabulary . .	205
B.1. Interval of Latent Trait θ and Transformed Latent Trait τ for which Each of Individual Conditions 1 through 5 is Met, for Levels 11, 12 and 13	210
B.2. Interval of Latent Trait θ and Transformed Latent Trait τ for which All Conditions, 1 through 5, Are Satisfied, for Levels 11, 12 and 13.	212

LIST OF FIGURES

FIGURE	PAGE
III.1. Square Root of the Test Information Function, $[I(\theta)]^{1/2}$, and Its Approximation by the Polynomial of Degree 7 Obtained by the Method of Moments for Each of Levels 11, 12 and 13 of the Vocabulary Subtest.	74
III.2. Estimated Item Characteristic Function Following the Normal Ogive Model (Solid Curve) and the Non-Parametric Estimated Item Characteristic Function (Dotted Curve) and Plausibility Functions (Dashed Curves) by the Simple Sum Procedure/Normal Approach Method, for Each of 82 Items of Vocabulary Subtest.	77
III.3. Mean Value of Ability, θ , at the Upper Bound of the Plausibility Range Corresponding to Each Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest. . . .	158
III.4. Mean Value of Ability, θ , at the Midpoint of the Plausibility Range Corresponding to Each Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest. . . .	159
III.5. Mean Value of Ability, θ , at the Peak of the Plausibility Function Corresponding to Each Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest. . . .	160
A.1. Estimated Item Characteristic Function Following the Normal Ogive Model (Solid Curve) and the Frequency Ratios Roughly Approximating the Non- parametric Item Characteristic Function (Dotted Curve) and Plausibility Functions (Dashed Curve), for Each of 43 Items of Level 11 Vocabulary.	194

CHAPTER I

INTRODUCTION

Overview and Objectives

Multiple-choice tests have been widely used for many years, due in large part to their relative ease in administration and scoring. However, specific information provided by the incorrect alternatives of multiple-choice items, which are called distractors or foils, has rarely been utilized (Samejima, 1984a). In general, scoring of multiple-choice items has involved a strict dichotomy between the correct answer and the distractors, with the distractors treated as being equivalent to each other. As Powell (1968, p. 403) points out:

Much time is spent by the examiner in the preparation of foils for multiple choice tests. A proportionally large time is spent by the examinee in making his selection decisions among the alternatives. In spite of the time thus spent, the foils are generally treated as a mask to the right answer and are lumped together in a general "wrong category."

In this study an attempt is made to determine whether distractors of a multiple-choice test of Vocabulary achievement provide useful, specific information with regard to the ability level of the examinee. In the process of reaching this goal, there are three main objectives of this study:

1. The first objective is to analyze the content of the items and their alternatives, with the purpose of identifying specific rules or reasons that would lead to the selection of particular distractors and of predicting which distractors will be informative. The prediction of informative distractors will include the ranking by knowledgeable judges of distractors from high to low, according to the

ability level of the examinees who are expected to find each distractor most plausible. It is hoped to demonstrate that informative distractors can be accurately identified in many cases from an analysis of their content. An informative distractor is one for which the conditional probability of selection, given ability, is above the chance level (assuming random guessing) and increases with ability over some meaningful interval.

2. The second objective of this study is to apply latent trait methods and approaches developed by Samejima (1981) to the items in order to estimate the plausibility functions, or the conditional probability of selecting a particular distractor, given ability. It is necessary to obtain the plausibility functions (described in detail in Chapter II) in order to determine which distractors are informative, and for what ranges of ability they are informative.

3. The third objective is to integrate the results of the content analysis of the distractors (Objective 1) and of the empirical analysis (Objective 2), in order to better describe the characteristics of informative distractors. It is expected that the data analysis will reveal at least some informative distractors that were not predicted from the content, and that the interpretation of the distractors will change somewhat in light of this additional information. The eventual goal is to identify rules or tendencies that will have practical implications for the construction of informative distractors, and thus more efficient multiple-choice items.

Background

Knowledge or random guessing principle. The theoretical basis for treating the distractors of multiple-choice test items as being equivalent is the knowledge or random guessing principle, which states that an examinee either knows the correct answer to an item or else guesses randomly. Following this logic, we should expect each distractor for an item to be selected by an approximately equal number of examinees, since the theoretical conditional probability, given ability, of choosing any particular distractor when guessing randomly is $1/m$, where m is the number of alternatives for the item. In addition, we should expect considerable noise due to random guessing, since the probability of choosing the correct alternative when the correct answer is not known is theoretically the same as that for choosing any of the distractors. Based upon the knowledge or random guessing principle, scoring formulas have been developed to adjust test scores and specific mathematical models in modern test theory have been developed in an attempt to account for the effect of guessing on test performance.

Probably the most conventional scoring formula to adjust test scores when the knowledge or random guessing principle is assumed is to subtract the number of incorrect responses divided by the number of distractors (e.g., Lord and Novick, 1968; Abu-Sayf, 1979), so that

$$S = R - [W / (m-1)] \quad , \quad (1.1)$$

where S is the formula score, R is the number of correct responses, W is the number of incorrect responses, and m is the

number of alternatives for each item. Given that an examinee knows the correct answer to k items out of the n items on a test and guesses randomly for the remaining $(n-k)$ items, and that the probability of randomly guessing correctly for each of these $(n-k)$ items is $1/m$, the expectation of the number of items correctly answered, R , is as follows:

$$\begin{aligned} E(R) &= k + (n-k) (1/m) \\ &= k + (n-k) / m \end{aligned} \quad (1.2)$$

From equations (1.1) and (1.2) the expectation of the formula score, S , is derived as follows:

$$\begin{aligned} E(S) &= E[R - W/(m-1)] \\ &= E[R - (n-R)/(m-1)] \\ &= E(R) - E[n/(m-1)] + E[R/(m-1)] \\ &= [k + (n-k)/m] - n/(m-1) + [k/(m-1) + (n-k)/\{m(m-1)\}] \\ &= k + [(n-k)(m-1) - mn + mk + (n-k)] / [m(m-1)] \\ &= k + 0 \\ &= k \end{aligned} \quad (1.3)$$

Thus we see from equation (1.3) that the expectation of the formula score S is equal to the number of items whose correct answers are known, k , and therefore S is an unbiased estimator of k (Lord and Novick, 1968, p. 307).

Some researchers have stated that even with formula scoring, examinees should be advised to guess since those who attempt, and guess at, more items tend to obtain higher formula scores (e.g., Cross

and Frary, 1977). Their logic is that examinees often know the material better than they think they do and can eliminate one or more distractors. Angoff (1989) investigated this claim, however, and did not find that guessing is generally advantageous to examinees when formula scoring is used. In a study of college applicants taking Verbal sections of the SAT, Angoff found that guessing might give a very slight advantage to high-ability examinees, but may actually be disadvantageous to low-ability examinees. His explanation of this finding is that partial information may lead to misinformed guessing, especially when test items were written so that distractors should attract examinees whose ability or knowledge is less than fully adequate. Such distractors may be most plausible to low-ability examinees.

Lord (1975, p. 7) has supported formula scoring while acknowledging that the assumption upon which it is based is not very sound:

Typically, examinees have some partial information about an item. For most multiple-choice items, they very likely can rule out one or more of the alternative responses with greater or less assurance. It is very difficult to be content with any kind of scoring based on an assumption of random selection. . . .

.
The writer does not assert the universal truth of this assumption. Only that it may often be correct as a reasonable approximation. Also that its correctness in common assumptions should be determined by research.

Lord (1980) has used a similar argument in support of the three-parameter logistic model, a latent trait model in which the knowledge or random guessing principle is assumed conditionally, given ability, and that will be discussed in more detail in Chapter II.

According to Lord (1980, p. 14), the model should be judged based upon empirical results rather than theoretical grounds.

Several researchers have reported evidence that contradicts the knowledge or random guessing principle (Ebel, 1968; Powell, 1968; Powell and Isbister, 1974; Samejima, 1980; Samejima and Trestman, 1980; Levine and Drasgow, 1983; Samejima, 1984a). Ebel (1968) found that blind or random guessing was reported to occur very infrequently on true-false test items administered to college students in measurement courses. Students were asked to indicate those items on which they guessed randomly, and, despite being guaranteed at least a chance level score for those items, the proportion of responses which were reported to be random guesses ranged from only 3 percent to 8 percent for four tests.

Powell factor analyzed the responses of college students to the distractors of Gorham's (1956) Proverbs Test. Students were asked to give reasons for their selection of alternatives on the test, and when these reasons were analyzed with respect to the observed factor structure, "a fairly clear psychological interpretation of each factor occurred" (Powell, 1968, p. 408). The interpretation of the factors was actually a classification of the types of errors that would result in the selection of the particular distractors. A chi-square test for independence showed that the frequency of selection for the 15 distractors differed significantly from the chance or random guessing level. Powell (1968, p. 412) concluded from this study that the random assumption for the selection of distractors is not acceptable and that "wrong answers clearly have psychological and statistical

meaning."

Powell and Isbister (1974) found that certain distractors and correct alternatives loaded highly on the same factor, leading them to conclude that the use of distractors could possibly improve the reliability of the subtests. They also found contradictory evidence to the usual assumption that the incorrect answers are opposite to and thus linearly related to the correct answers, indicating to them that the distractors contained information not provided by the correct answers.

Samejima (1980) introduced an index that can be used with multiple-choice test items to test the assumption of the knowledge or random guessing principle. This index, Index k^* , was derived from an index for the number of hypothetical, equivalent alternatives developed by Sato (1977), index k , which is based upon information theory (e.g., Goldman, 1953). An informative item is one for which there is considerable uncertainty in the selection of an item alternative. Sato's index k increases as the uncertainty of selection increases, with a maximum equal to the number of item alternatives, which occurs only if each alternative is equally probable.

Index k^* should be very close to the number of multiple-choice alternatives for the item following the knowledge or random guessing principle. However, in analyzing data from junior high students in Japan taking word comprehension tests constructed by Shiba (1978), Samejima (1980) found that for many items Index k^* was significantly smaller than the number of item alternatives. This finding and the finding that the mean estimate of ability for examinees selecting

different distractors of an item usually differed substantially lead to the conclusion that the knowledge or random guessing principle was not valid for many of these items (Samejima, 1980).

Samejima's Index k^* was also applied to data collected for Levels 11, 12 and 13 of the Iowa Tests of Basic Skills, Form 6, Levels 9-14 (Hieronymus and Lindquist, 1974). The results of this study, presented in a research report by Samejima and Trestman (1980), indicate that most of the items on these tests are not satisfactorily explained by random guessing. In addition to estimating Index k^* , the chi-square test for the goodness of fit was applied to the frequencies of examinees selecting the distractors of each item. These items appear to belong to a family of mathematical models that Samejima (1979) has labeled the Informative Distractor Model, defined as the family of models in which it is assumed that specific information can be obtained from separate alternatives of the multiple-choice item. The Informative Distractor Model is contrasted to the Equivalent Distractor Model, in which the distractors of an item are considered to have equivalent plausibility functions, or the same conditional probability of being selected, given ability.

Levine and Drasgow (1983) examined the relationship between estimated ability level and choice of distractor for the Verbal sections of two well-known multiple-choice tests, the Scholastic Aptitude Test and the Graduate Record Examination. Examinees for each test were divided into subgroups based upon their ability estimates. The frequency of selection and pattern of responding often differed for an item's distractors over the range of estimated ability, which

contradicts the knowledge or random guessing principle. It was found that for several items one or two distractors might be frequently chosen by high ability groups while another distractor might be chosen by low ability groups.

Further evidence that examinees of different ability levels select distractors with different probability was presented by Samejima (1984a). This analysis involved multiple-choice test data obtained from the Educational Testing Service in which examinees were divided into ability subgroups on the basis of their total test score. Similar results were obtained from data for one of Shiba's word/phrase comprehension tests, which were referred to earlier in this chapter. In this case, however, ability subgroups were determined from maximum likelihood estimates of ability rather than from test scores. In addition, Samejima (1984a) used various methods and approaches that she has developed to estimate the plausibility functions for one of Shiba's tests, and the results showed that the plausibility functions for the distractors of most of the items were clearly different, a result that again contradicts the random guessing principle.

Diagnostic information from distractors. The belief that errors on test items are usually not random but instead reflect systematic tendencies has been fundamental to researchers who have studied such errors with respect to the diagnostic or developmental information they provide. To illustrate, Glaser (1981, p. 10) stated:

An important skill of teaching is to identify the nature of the concept or rule that the student is employing that governs his performance in some systematic way (in most cases, the student's behavior is not random or careless, but is driven by some underlying misconceptions or by incomplete knowledge).

Much diagnostic research in error analysis has been done in mathematics, probably due in large part to the relatively systematic and logical progression of skills required in this area. Several researchers have investigated incorrect mathematical rules or procedures that can possibly be identified from the errors of examinees (Brown and Burton, 1978; Birenbaum and Tatsuoka, 1980, 1981; Tatsuoka et al., 1980; Tatsuoka, 1981; Mayberry and Ory, 1985). Brown and Burton (1978) have borrowed the term "bug" from computer science to describe the incorrect procedure within a series of procedures that can lead to an error on a task. Based upon their experience that "students are remarkably competent procedure followers, but that they often follow the wrong procedures" (Brown and Burton, 1978, p. 157), they have developed a network of diagnostic models of arithmetic computation procedural skills. Various procedures necessary to carry out a task are linked together in what are called a "procedural network" (p. 159), with each procedure consisting of two parts: the intent of the procedure and the actual methods by which that intention is carried out. A procedure can be carried out by more than one method in most cases, and networks describing complex behaviors can become quite involved. However, in order to isolate the incorrect methods or bugs by which procedures are carried out, the task must be broken down into very basic behaviors or skills. Diagnosis of specific misconceptions can be inferred from errors that are explained by consistent application of this bug.

Tatsuoka et al. (1980) introduced a method of diagnosing misconceptions using "error vectors," which consist of all possible

operations in doing a particular problem. Based upon the pattern of responses of an examinee to the items on a test, particular types of errors that occur with some consistency can be identified. A vector is used to represent each significant step in the problem, and each action involved in this step corresponds to an element in this vector. A value of 0 or 1 is assigned to each element in this vector depending upon whether or not the occurrence of this action is indicated by the response to the item. Multiplication of vectors corresponding to the same type of problems is used to determine whether or not specific actions or behaviors were followed consistently by the examinee. This approach requires careful and detailed analysis of the content and the tasks being tested, as does the approach by Brown and Burton (1978), but does not depend upon a universal procedural network to explain errors.

Tatsuoka et al. (1980) state that diagnosing the misconceptions of students is not only important to improving their learning, but is also beneficial to the scoring of tests, since some items can be answered correctly by an incorrect rule or operation. These "false corrects" (Tatsuoka et al., 1980), or "answering anomalies" (Powell and Miki, 1985), that can lead to measurement error, are most likely to be identified when the rules which an examinee follows are identified. Powell and Miki (1985) found through interviews that a significant percentage of mathematics items were correctly answered when the concept was not understood or were incorrectly answered when the concept was understood.

More recently Tatsuoka, Birenbaum and Arnold (1989) have found

that procedural rules may not be applied consistently by all examinees. Students who had not mastered the skills being tested appeared to shift between rules in identical situations and to develop new rules when the old ones did not lead to a satisfactory solution. Students who had achieved mastery, however, tended to remain consistent in their application of rules during the test sessions.

Testwiseness and rules of item writing. If it is indeed true that distractors are generally selected for reasons other than random guessing, and that one particular distractor of an item may be more likely to be chosen than the others, the question may then be raised: What are the characteristics of a distractor that make it more plausible to certain examinees? The majority of studies which have investigated the characteristics of multiple-choice items and their distractors have looked for variables that are considered unrelated to the ability that is supposed to be measured, yet which do affect the way examinees respond to items. The hypothetical construct "test-wiseness" can be defined as the ability to improve one's test score by using extraneous clues from multiple-choice test items (Diamond and Evans, 1973). Since test-wiseness involves the use of extraneous clues in the items themselves, it is unrelated to the content area of interest, and items that are advantageous to the testwise examinee are often seen as being flawed, or poorly constructed in some way. A number of studies involving test-wiseness have investigated specific types of item flaws.

Haladyna and Downing (1984) reviewed the literature relating to multiple-choice item writing to identify specific principles or

practices suggested in item construction. They identified 45 rules organized into 6 different categories: (a) item content in general (i.e., "avoid opinion-based items"), (b) item construction in general (i.e., "use good grammar consistently, making sure that the item and the options agree grammatically"), (c) stem construction (i.e., "word the stem positively; avoid negatives"), (d) option construction in general (i.e., "keep the length of options fairly consistent"), (e) construction of the correct alternative (i.e., "make sure there is one and only one correct option"), and (f) construction of the distractors (i.e., "avoid technically phrased distractors"). In a later study (Haladyna and Downing, 1985), they examined the research supporting these various rules or principles and found that there had been little research to support most of these rules. Although many of these rules appeared to be supported by "common sense and face validity" (Haladyna and Downing, 1985, p. 20), others were found to be in need of further investigation. One rule that was identified by Haladyna and Downing (1984, 1985) that is especially pertinent here states: "Avoid illogical distractors; use plausible distractors." Three different studies involving this principle were found in their review of the literature.

The earliest of these three studies, by Owens, Hanna and Coppedge (1970), compared three methods of generating distractors with respect to the reliability and the validity of the test. The three methods compared were the judgmental method, in which authors generate distractors that they believe will be plausible, the frequency method, in which the most frequent incorrect responses given by examinees when

the test is given in completion form are used as distractors, and the discrimination method, in which the distractors selected are those incorrect responses to the items given in completion form that give the largest discrimination index, or in other words, that best discriminate between high and low scoring examinees. It was found that the frequency method and the discrimination method were not significantly more reliable or valid than the simpler judgmental method, although it was pointed out that the judgmental test was composed from the pooled judgments of teachers, rather than one particular judge, thereby making it a more laborious, costly procedure.

A study by Smith (1982) indicates that the use of plausible distractors can sometimes cue the testwise examinee to the correct answer, as the different concepts or dimensions represented by the distractors can converge to the correct answer. If distractors are chosen that are similar to the correct answer but which involve different concepts, the correct answer might be found by a testwise examinee who finds the alternative that is related to two or more of the concepts presented. For example, suppose that an item on a multiple-choice test had the following four alternatives: (a) Winter, (b) October 22, (c) Easter, and (d) Fourth of July. An examinee not knowing the correct answer might select alternative d following the strategy discussed in his paper. If we look at familiar concepts or dimensions associated with these alternatives, we see that alternative a is a season, alternative b is a day of a month, alternative c is a holiday, and alternative d is both a day of a month and a

holiday. From this viewpoint, alternative d would be considered the most likely choice as the correct answer since it involves two different dimensions that are similar to the other alternatives. Smith (1982) found that this strategy can be learned by examinees and that it can result in higher test scores. This study indicates that care must be taken when constructing plausible distractors so that the correct answer is not cued to testwise examinees.

Weiten (1984) examined the effects of four types of violations of generally regarded item writing principles, such as resemblance between the item stem and only the correct option, grammatical inconsistency between the stem and the distractors, implausible distractors, and consistently increased length of the correct alternative in comparison with the distractors. He found that except for the case of grammatical inconsistency, items having these flaws were significantly less difficult than items that did not violate these principles. However, no significant differences were found between flawed and unflawed items for any of these four principles with respect to item discrimination or test reliability or validity.

Board and Whitney (1972) did not find significant differences in test difficulty when the correct alternative was the longest alternative, but they did find that examinees of lower ability were more likely to use this clue than were examinees who obtained higher test scores. Chase (1964) also found no relationship between correct option length and test difficulty, but did find that examinees were more likely to use this clue for more difficult items than for easier items. A number of studies, however, indicate that longer correct

options lead to easier items (McMorris et al., 1972; Jones and Kaufman, 1975; Strang, 1977; Evans, 1984; Weiten, 1984). Jones and Kaufman obtained results contradictory to those obtained by Board and Whitney (1972), in that higher ability examinees were found to use the length clue more often than lower ability examinees.

Green (1984) found that small variations in the phrasing of general information items administered to undergraduates did not have a significant effect on item difficulty. The variations included differences in the length of the stem, the syntactic complexity of the language used, and substitution of familiar terms in the stem with more uncommon terms. Item difficulty was found to be significantly affected, however, by the similarity between item alternatives, with items having more similar options being more difficult.

Strang (1977) found that when multiple-choice alternatives varied along the dimensions of length and of technicality of terms, long options were selected more often than short options and non-technical options were more frequently selected than technical options. When the familiarity of the terms was considered, it was found that the less familiar a technical alternative was, the less frequently it was selected (Strang, 1977, p. 258).

A concept sometimes confused with testwiseness is response sets, which refers to a bias in response selection by the examinee (Sarnacki, 1979). An example of a positional response set would be that of an examinee who frequently selects the third alternative when he/she does not know the correct answer to a multiple-choice item. An important distinction between response sets and test-wiseness is that

the former does not necessarily lead to a higher test score, and in some cases may result in lower scores.

Scoring and testing procedures to assess partial knowledge.

Attempts to improve test reliability and validity by assessing partial or incomplete knowledge of examinees have resulted in a number of different scoring procedures and testing procedures with multiple-choice items (Wang and Stanley, 1970; Kansup and Hakstian, 1975).

Scoring procedures have involved assigning different weights to the different item alternatives and taking an examinee's test score as the sum of the weights of those alternatives selected. The weights can be determined through logical weighting, in which distractors are judged subjectively with regard to their closeness to the correct answer, or through empirical weighting, based upon the psychometric properties of the alternatives (Kansup and Hakstian, 1975). Studies of differential weighting have generally not shown significant increases in validity (Kansup and Hakstian, 1975), although a few have shown improved reliability (Nedelsky, 1954a; Patnaik and Traub, 1973).

Nedelsky (1954a) used expert opinions to evaluate distractors of items on a comprehensive evaluation in order to identify those that were considered so implausible as to appeal only to students who were likely to fail. A composite score, denoted by C , was computed for each examinee in which the number of such implausible distractors selected, S_i , divided by the average number of implausible distractors per item, A_i , was subtracted from the number of items correctly answered, R , so that

$$C = R - S_i/A_i \quad . \quad (1.4)$$

It was found that the composite score was much more reliable than the usual total correct score for the group of examinees, and that for students who obtained a grade of either D or F on the test, the number of distractors identified as being implausible, S_i , was the most reliable score.

Davis and Fifer (1959) used both expert evaluators and empirical results to determine weights for different alternatives on a test of arithmetic reasoning. From the responses of a sample group of examinees, weights for the alternatives were obtained from the correlation between the selection of the alternatives and the total test score. These weights were modified, however, by a priori ratings given by two mathematicians. The authors concluded from the study that the weighted scores were significantly more reliable than unweighted scores, but were not more valid.

Hendrickson (1971) found improved reliability with Scholastic Aptitude Test (SAT) data using a variation of a method developed by Guttman (1941) which determines option weights that maximize internal consistency. Initial weights were obtained from the mean total correct score on the remaining items of the test for those examinees who selected that particular category. An iterative method was then used to find the weights that produced maximal internal consistency reliabilities.

Two testing procedures that are designed to measure partial knowledge with multiple-choice items are elimination testing (Coombs, Milholland and Womer, 1956) and confidence testing (Dressel and

Schmid, 1953). Elimination testing requires that the examinee cross out all item alternatives that are believed to be incorrect. One point is given to the examinee for each incorrect alternative crossed out and $(m-1)$ points are subtracted, where m is the number of response alternatives, for a correct answer that is crossed out. This technique requires that the examinee make some discrimination between the alternatives in order to receive credit, distinguishing between one or more plausible alternatives and at least one implausible alternative. If the examinee can not make any discrimination and fails to eliminate any of the alternatives, or instead eliminates all of the alternatives, including the correct answer, a score of zero is received for this item. Studies using this method have not demonstrated significant increases in reliability and have shown little evidence of improved validity (Hakstian and Kansup, 1975).

Equally unsuccessful in this regard have been studies involving confidence testing, which requires that that the examinee rank or assign weights to the alternatives of an item, such that the more confidence he/she has that an alternative is correct, the more highly it is weighted. Dressel and Schmid (1953) had examinees assign weights of one to four to the multiple-choice alternatives that they thought were correct. The score for the test was obtained by subtracting the sum of the assigned weights for incorrect alternatives from the sum of the weights for the correctly identified alternatives. A small increase in test reliability was reported. Shuford, Albert and Massengill (1966) used a similar approach in which a number between zero and one was assigned to each alternative to reflect the

examinee's perceived probability of the response being correct.

Thissen (1976) applied a latent trait model developed by Bock (1972), for use with two or more nominal categories, to the Raven Progressive Matrices. The model utilized information obtained from incorrect responses to reduce the error estimation of the test. Thus ability estimation could be improved without increasing the number of test items. While the items of the Progressive Matrices were developed so that distractors have varying degrees of similarity to the correct response, Thissen (1976, p. 213) states that distractor analysis is useful in other cases:

Where distractors are not constructed with a known relationship to ability, multiple category analysis of the test items would be useful in determining the performance of individual distractors, both for purposes of item and distractor selection in test construction and for the study of the cognitive processes contributing to the item responses by individuals. It may even be possible to design distractors so that items are answered correctly only by persons of high ability, but have distractors chosen differentially by those of lower ability.

Background of the Iowa Testing Programs

The Iowa Testing Programs developed out of a project sponsored by the College of Education and the Extension Division of the University of Iowa that was designed to improve educational measurement and to stimulate interest among high school students within the state of Iowa (Peterson, 1983). This project was the first statewide achievement testing program of its kind, with the first tests being administered in 1929. One purpose of the Iowa program was to provide structured competition in academics, with local, district and state contests.

Assistant Professor E.F. Lindquist of the University of Iowa

was responsible for the supervision of committees of his colleagues in the construction and administration of the early tests. These tests covered several different content areas, including English Mechanics, English and American Literature, World History, American History, Algebra, Plane Geometry, General Science and Physics. A variety of item formats were used, such as multiple-choice, completion, matching, true-false, and free-response.

After the high school testing program became established, a project was initiated to develop tests for the junior high grades. This project, which was called the Iowa Every-Pupil Elementary School Testing Program, led to the development of a battery of tests for grades 6, 7 and 8 in 1935. The first battery consisted of fourteen subtests making up three main tests: Silent Reading Comprehension, Basic Study Skills and Basic Language Skills. A fourth test, Basic Arithmetic Skills, was added in 1936.

The administration times for the tests were longer than those of most other existing tests, ranging from 50 to 80 minutes. This time was necessary in constructing diagnostic tests that could be useful in remedial and individualized instruction and in educational guidance. In general, the tests were designed to measure how students used their academic skills in practical applications.

The Iowa Basic Skills Program was expanded to grades three through five in 1940. The 1940 edition of the Basic Skills Tests was the first in a series of annual forms published over the next eleven years. Nine parallel forms were developed during this time that served the statewide testing programs for fifteen years. Rotation of

forms and reuse only at broadly spaced intervals helped insure that the test content was not previously familiar to examinees.

Over the years a great number of changes have occurred in the skills tested and the methods and formats used to measure achievement in different content areas, but the Iowa Testing Programs have continued to serve as a leader in the field. A description of many of these changes in the tests, which were eventually expanded to test skills as early as kindergarten, and of many of the technical and engineering innovations which have accompanied the development of the tests over the past five decades has been given by Peterson (1983).

The Iowa Tests of Basic Skills (Hieronymus and Lindquist, 1974), as they are currently called, are now designed for students in grades three through nine. In addition, a Primary Battery and Early Primary Battery have been developed for children in kindergarten to the third grade. A technical advancement was made around 1970 that provided for fast processing of mixed answer sheets, allowing individualized or out-of-level testing to become a reality. It was no longer necessary for all students in the same grade to take the same test. Since the third edition of the Iowa Tests of Basic Skills was being assembled when this achievement was made, the grade designations that had been placed on previous editions were replaced with level designations corresponding to chronological age. Thus the six levels of the test battery were numbered 9 to 14, representing typical ages of pupils in grades three to eight.

Iowa Tests of Basic Skills, Form 6

The test battery used in this study is the Iowa Tests of Basic Skills, Form 6, Levels 9-14. These tests have been designed with the purpose of assessing "generalized intellectual skills and abilities" (Hieronymus and Lindquist, 1974, p.6) rather than achievement in specific content subjects, such as geography, science or literature.

There are five major content areas on the Iowa Tests, covering vocabulary, reading comprehension, language skills, work-study skills, and mathematics. These five major areas are broken down into 11 separate tests, or subtests, each of which concentrates on a different basic skill. All items are in a multiple-choice format, and for each of Levels 9 through 14, only a subset of the items in each subtest is administered.

The subtest specifically investigated in this study is Vocabulary. This subtest assesses the examinee's knowledge of the meanings of words by requiring identification of the response alternative that is closest in meaning to the stimulus word, or key word, for each item. A phrase or stem is presented in each item in which one word, the key word, is in heavy type, and four alternatives are presented underneath. The directions ask that the examinee decide which of the four answers has most nearly the same meaning as the word in heavy type above them, and then to mark the number of that answer on the answer sheet.

CHAPTER II

DATA SOURCE AND METHODS

The test data used in this study was provided by Dr. William Coffman of the University of Iowa, who also served as Director of the Iowa Testing Programs, which were discussed in Chapter I. The initial investigations of this data were conducted by Samejima and are described in two research reports produced under contract with the Office of Naval Research (Samejima and Trestman, 1980; Samejima, 1984b). The author is indebted both to Dr. Coffman and to Dr. Samejima for their permission to use the data in this dissertation research.

Subjects

Data were collected in three different school systems in the state of Iowa, during the years 1971 through 1977, and consist of the item responses of 7,581 students in grades three through seven to the Iowa Tests of Basic Skills, Form 6, which were described in Chapter I. Students in each grade took the level of the test appropriate for their age group, so that students in grade three took Level 9 Test, students in grade four took Level 10 Test, and so forth. Thus the original data included responses to Levels 9 through 13 of the Iowa Tests. Out of the original 7,581 examinees, 28 took Level 9 Test, 114 took Level 10 Test, 2,460 took Level 11 Test, 2,452 took Level 12 Test, and 2,527 took Level 13 Test. Since such a small number of examinees took Levels 9 and 10, Samejima excluded them from the

original group and referred to the remaining 7,439 examinees as the "original data" (Samejima and Trestman, 1980).

It was found during the initial investigation by Samejima and Trestman (1980) that a relatively small portion of the examinees in the original data did not respond to a large number of the test items. Of the 7,439 examinees for Levels 11, 12 and 13, only 162 left 100 or more of the test items unanswered. Some examinees skipped one or more entire subtests. After close analysis of the data, Samejima decided to exclude 193 examinees who left one half of the items of a subtest, or more, unanswered, leaving 7,246 examinees in the "revised data" (Samejima and Trestman, 1980). By excluding these 193 examinees, there were only 28 who left more than 100 items in the total test unanswered, and no examinees who omitted more than 200 items. The revised data consists of 2,364 examinees for Level 11, 2,413 for Level 12, and 2,469 for Level 13. It is these revised data that are used in this dissertation study.

As discussed, the Iowa Tests are composed of 11 subtests of academic achievement. In this study one particular subtest, Vocabulary, is investigated. This subtest measures a content area that is considered to be very important to academic success and is covered to some extent in most general achievement test batteries. A description of this subtest and its item format was presented in Chapter I. Items are in a multiple-choice format with four alternatives each. The content analysis and latent trait analysis conducted in this study included items from all three Levels, 11, 12 and 13. There are a total of 82 Vocabulary items for these three

levels. Table II.1 presents the number of items administered to each level for this subtest, as well as the number of items that overlap, or are administered to more than one level.

The Vocabulary subtest was designed to be a power test rather than a speed test, as were all of the subtests of the Iowa Tests. However, as Samejima and Trestman (1980) pointed out in their initial investigation, there is a definite tendency that items presented later within a subtest are omitted by larger numbers of examinees. Thus the proportion of examinees correctly answering these later items is less than would be expected under ideal testing conditions.

Content Analysis and Prediction of Informative Distractors

There are two distinct approaches that are taken in this study in analyzing the distractors of the multiple-choice items. The first approach is to look solely at the items, independent of any response data, and to attempt to identify informative distractors based upon their content. The second approach, which will be described later in this chapter, is to identify informative distractors empirically, utilizing latent trait methods to estimate the plausibility functions from the item responses of a group of examinees.

The primary purposes of the content analysis are to predict which distractors will be informative, and more importantly, to identify reasons why these distractors might attract examinees of different abilities. By identifying general tendencies or perhaps specific rules that appear to be affecting the examinees' selection of alternatives, it becomes possible to make recommendations that can hopefully lead to the construction of informative distractors and thus

TABLE II.1 Number of Vocabulary Items in Each of Levels 11, 12, and 13, Number of Overlapping Items Between Levels, and Number of Items Appearing in Only One Level.

Level	Number of Items
11	43
12	46
13	48
11,12,13	9
11,12	27
12,13	28
11 Only	16
13 Only	20
Total	82

more efficient multiple-choice items in the future.

The major part of the content analysis was done by the author, although certain items were also presented to selected university faculty members and to teachers in the local public school system to obtain their expertise and their perception of informative distractors. The content analysis by the author was conducted in five major steps:

1. Each item for Levels 11, 12 and 13 of the Vocabulary subtest was read carefully by the author and the alternatives were ranked according to the ability level of the examinees who were predicted to find a particular alternative more plausible than the others. The number 1 was assigned to the answer considered to be correct, the number 2 was given to the distractor that should attract examinees of relatively high ability, the number 3 was given to the distractor appealing to the next-highest level of ability, and 4 was assigned to the distractor expected to appeal to the lowest ability group. It should be made clear that the numbers assigned to the distractors are strictly rankings and indicate only the relative order predicted with regard to ability among subgroups of examinees not selecting the correct answer. Due to this fact and because items are of different difficulty, rankings for different items are not directly comparable. It may be more informative to think of the rankings in terms of plausibility to high-ability examinees, with 1 being the most plausible and 4 the least plausible.

Distractors that were not predicted to be informative were still ranked but were separately identified as not being informative.

Thus each item on these two subtests had its alternatives numbered from 1 to 4 by the author, although in a few cases none of the distractors was predicted to be informative. For some items, only one of the distractors, ranked as number 2, might be predicted to be informative, while for other items either two or all three distractors might be judged to be so.

2. The basis or reasoning behind the author's predictions of informative distractors was written down for each distractor. Some of the reasons were quite specific to the particular distractor, such as a reference to a part of the wording of the distractor that might cause the examinee to associate that alternative with the correct answer. For certain selections the reasons or comments given were rather general and could be applied to other similar types of distractors.

3. Following the order or ranking of alternatives obtained in step 1, the predicted plausibility functions and the item characteristic function for each item of Levels 11, 12 and 13 of Vocabulary subtest was sketched, or roughly drawn, by the author. The purpose of this step was to provide additional information regarding the author's perception of the distractors, including the relative frequency with which alternatives were expected to be chosen and the position of the plausibility curves relative to each other. For instance, the plausibility function for a hypothetical item distractor, A, could be higher than the plausibility function for distractor B of the same item at every level of ability, and yet B could be ranked higher than A in terms of the relative ability level

of examinees selecting each. This situation would occur when the ability level corresponding to the peak of the plausibility function of distractor B was greater than the ability level for which the plausibility function of distractor A was at its peak.

4. The various rules and reasons behind the predictions of informative distractors obtained in step 2 were reviewed in an effort to find similarities and general tendencies by which these rules and reasons could be classified. It was expected that certain of the rules might be so specific as to be applicable only to one or two items while other strategies or reasons might apply to several of the items. After identifying these rules, the distractors of the items within each subtest were categorized accordingly.

5. After the empirical plausibility functions were generated, as will be discussed later in this chapter, the author's predictions of informative distractors were compared to the actual results. The percentage of items in agreement between the predicted and the observed informative distractors was computed for each subtest. Items were reviewed in light of the information provided by the empirical results and an attempt was made to identify those rules that appeared to be most related to the informative distractors.

Predictions of informative distractors were also made for some of the items of each subtest by university faculty members and by public school personnel. This part of the content analysis was completed in the following three steps:

1. Sixteen items were selected from the Vocabulary subtest to be evaluated by faculty members and teachers. These items came from

each of Levels 11, 12 and 13 and were considered to be representative of the subtest in terms of difficulty and of the content tested. Another consideration in the selection of these items was the overall frequency distribution of examinee responses to the alternatives for each level, that was obtained previously in the initial investigation of the data by Samejima and Trestman (1980). Items for which relatively high frequencies of examinees selected various distractors were generally preferred since they were more likely to indicate informative distractors. In addition, certain items were selected because their content was considered to be particularly interesting or informative. Table II.2 gives the actual item numbers of the items selected to be presented to the faculty/teacher evaluators and specifies the level or levels in which each item appears.

A separate form containing the 16 Vocabulary items was presented to the evaluators. The first two pages contained the instructions for completing the form and an example item. The concept of the informative distractor was briefly explained and the evaluator was asked to number the alternatives for each item from 1 to 4, with 1 being given to the correct answer, 2 to the distractor predicted to be selected by examinees of relatively high ability, 3 to the distractor attracting examinees of lower ability, and 4 to the distractor likely to be selected by examinees of the lowest ability level. In addition, evaluators were asked to write their reason for ranking the distractors as they did. The selected items were numbered 1 through 16, with two items presented on each page. The items were actual items taken from the Iowa Tests except that the order of the

TABLE II.2 Vocabulary Items Selected for Rankings of
Distractors by Faculty/Teacher Evaluators.

	Item	Levels
1	27	11
2	37	11
3	40	11, 12
4	53	11, 12
5	56	11, 12
6	59	11, 12
7	60	11, 12, 13
8	64	11, 12, 13
9	65	11, 12, 13
10	67	12, 13
11	77	12, 13
12	84	12, 13
13	86	13
14	94	13
15	100	13
16	104	13

alternatives was randomized for each item and the alternatives were lettered, a through d, rather than numbered, 1 through 4. Items were renumbered and alternatives randomized in order to help maintain test security. The lettering of alternatives helped distinguish the randomized version from the original version of the items. A blank on which the evaluator was to put his/her ranking was left in front of each alternative. Several blank lines corresponding to each distractor appeared to the right of each item for comments from the evaluator regarding the rankings. An additional page for general comments was given at the end of each form.

2. The form containing 16 selected items was completed by four members of the faculty of the University of Tennessee and by three teachers in the Knoxville City School system. The faculty members came from the departments of Psychology, Educational and Counseling Psychology, and Marketing and Transportation. Two of the teachers were fifth grade teachers and the other was a resource teacher who teaches elementary school children receiving special education assistance outside of the regular classroom.

3. The predictions of informative distractors made by the author were combined with faculty/teacher rankings for the 16 selected Vocabulary items. The average ranking for each distractor was computed from the combined data. The distractors were then ranked on the basis of these averages, from highest to lowest ability of the examinees expected to be attracted to each. The percentage of items in agreement between these average rankings and the observed rankings was computed. The predicted rankings were compared to the observed

results for all 16 selected items. As an index of the agreement among the faculty/teacher evaluators in their predictions, the Horst reliability coefficient (Horst, 1949) was computed for each subtest. This reliability coefficient, that was initially used with multiple ratings or evaluations of job proficiency, can be used in any situation in which multiple ratings exist, all purporting to measure the same thing, for each of a group of individuals or objects to be evaluated. The Horst reliability coefficient, denoted by r , is given by

$$r = 1 - \frac{\frac{\sum_{k=1}^N \frac{\sigma_k^2}{n_k - 1}}{N}}{\sigma_M^2}, \quad (2.1)$$

where σ_k is the standard deviation of the ratings for object k (distractor k in this case), N is the number of objects (distractors) evaluated, n_k is the number of ratings for object k , and σ_M is the standard deviation of the means, denoted by M_k , of the ratings for the N objects. This coefficient increases as the variation in ratings for specific objects decrease, or account for a smaller portion of the total variation. This equation (2.1) is based upon the generalized equation for the reliability coefficient,

$$r = 1 - \frac{\sigma_e^2}{\sigma_o^2}, \quad (2.2)$$

where σ_e is the error variance of the measures or ratings and σ_o is the observed variance. The derivation of the equation (2.1) and

the proof that the Spearman-Brown reliability index and the Kuder-Richardson formula 21 for estimating test reliability are special cases of this equation are presented by Horst (1949).

Latent Trait Theory and Applications to Multiple-choice Test Items

Background. Latent trait theory, or item response theory, developed out of the early work of researchers such as Lawley (1943, 1944), Tucker (1946) and Lord (1952). Latent trait theory is sometimes referred to as modern test theory, to distinguish it from classical test theory, which has been unable to provide appropriate solutions to a number of the problems that are faced in testing situations (Samejima, 1977; Hambleton, 1978).

Latent trait theory assumes that behavior can be explained to a significant extent by the existence of hypothetical constructs, which are called latent traits. As latent traits, or abilities, are thought to largely determine how an individual behaves in certain situations, knowledge of the degree to which these traits are possessed by the individual can help in the prediction of future performance in these situations. Like many other constructs in psychology, these traits are not directly observable and measurable, but must be estimated from observable behavior. It is not implied that latent traits exist in any physiological sense (Lord and Novick, 1968, p. 358), but only that they can be used to make meaningful predictions.

The relationship between the latent trait and some specific behavior of interest is often expressed by a mathematical model. In mental testing, this behavior is typically an individual's response to

a particular test item. As will be discussed later in this chapter, latent trait models allow for the outcome of this behavior, such as whether the item was answered correctly or incorrectly, to be specified probabilistically and conditionally, given the ability level of the individual. This conditional probability does not depend upon any specific group of examinees and, therefore, unlike classical test theory, the parameters in latent trait models are "population-free" (Samejima, 1980).

Unidimensional latent trait theory. As was stated earlier in this chapter, latent trait theory assumes that particular behaviors can be explained by hypothetical latent traits. The number of different latent traits necessary to account for a certain behavior depends upon both the complexity of the behavior and the way in which the latent traits are defined. For example, one could define a single construct, such as intelligence, which includes many different types of abilities, or these component abilities could be defined as separate constructs. More complex behavior that requires a variety of skills to accomplish would generally be less likely to be expressed by just one latent trait than would a relatively simple behavior.

The set of latent traits necessary to account for the behavior of interest is called the latent space. If the number of latent traits required to describe the behavior is denoted by k , then the latent space is said to be k -dimensional. While latent trait models have been developed for the multidimensional latent space (Samejima, 1974), applications have generally been limited to the unidimensional case, where one latent trait is assumed to be sufficient. For this

reason, and particularly due to the complexity of multidimensional models, this study will deal only with unidimensional latent trait theory. In this research the behavior of interest is performance on a test of vocabulary, and the latent trait could be defined as vocabulary achievement.

In order to determine the dimensionality of the latent space underlying the items of a test, factor analysis can be applied to the tetrachoric correlation coefficient matrix obtained from the observed interitem contingency tables (Samejima, 1984a). This procedure can be justified theoretically on the basis of certain assumptions (Lord and Novick, 1968). First we assume that there is a continuous variable, which is called the response tendency, or item variable, that underlies the observable response to an item. For an arbitrary item, which we will denote as item g , the response tendency will be given by the symbol X_g . Item g is scored dichotomously, as either correct or incorrect, and we assume that there is a cutting point of the response tendency, denoted by γ_g , such that the item is answered correctly if the response tendency X_g is greater than or equal to the cutting point γ_g , and the item is incorrectly answered if X_g is less than γ_g . Denoting the item score by u_g , and assigning a score of zero to an incorrect item and a score of one to a correct item, we have:

$$u_g \begin{cases} = 0 & \text{when } X_g < \gamma_g, \\ = 1 & \text{when } X_g \geq \gamma_g. \end{cases} \quad (2.3)$$

Let us further assume that for a test of n items, the n different response tendencies for these items have an n -variate normal distribution, and thus each pair of items has a bivariate normal distribution. The tetrachoric correlation coefficient obtained from the interitem two-by-two contingency table for items g and h under these conditions will express the correlation between the two underlying response tendencies X_g and X_h (Lord, 1980, p. 39).

If the latent space is unidimensional and a single common factor, that we define as the latent trait and denote by θ , underlies the n items, then from the factor analysis model we have the following:

$$\begin{aligned} X_1 &= \lambda_1 \theta + e_1 \\ X_2 &= \lambda_2 \theta + e_2 \\ &\vdots \\ X_g &= \lambda_g \theta + e_g \\ &\vdots \\ X_n &= \lambda_n \theta + e_n \end{aligned} \quad (2.4)$$

where λ_g is the factor loading on the latent trait θ and e_g is the unique factor corresponding to item g . Thus the factor loading λ_g obtained from the factor analysis of the interitem tetrachoric correlation matrix indicates the correlation between item variable X_g and latent trait, or ability, θ . In practice, if the principal factor solution of factor analysis yields one eigenvalue that is much larger than all the others, indicating a dominant factor, the latent space can be treated as being unidimensional (Lord and Novick, 1968, p. 382).

In order to estimate an examinee's ability from his/her pattern of responses to a set of n items, it is necessary to make the assumption of local independence. Local independence is defined conditionally, for all examinees having the same level of ability, or abilities in the multidimensional latent space, and states that, when ability is fixed, the probability of correctly answering any item is independent of the responses to all other items. This does not mean that the item scores are independent of each other for the total group of examinees (Lord and Novick, 1968, p. 361), but only for those examinees having the same position in the latent space.

For dichotomous items in the unidimensional latent space, local independence means that the joint conditional probability, given ability θ , for a particular pattern of item scores for n items is equal to the product of the n marginal conditional probabilities for the individual items, so that

$$\text{prob.}(u_1, u_2, \dots, u_g, \dots, u_n | \theta) = \prod_{g=1}^n \text{prob.}(u_g | \theta) \quad , \quad (2.5)$$

where u_g is the dichotomous item score for item g , and $\text{prob.}(u_g | \theta)$ is the conditional probability of obtaining u_g , given ability θ .

Mathematical models for the item characteristic function. One of the most important concepts in latent trait theory is the item characteristic function, which is defined as the conditional probability of correctly answering an item, given ability θ . The mathematical models used by researchers and practitioners to represent the item characteristic function can be categorized based upon the number of parameters they contain. The most frequently used models

involve one, two or three item parameters.

The item characteristic function, denoted by $P_g(\theta)$, is defined for the three-parameter logistic model as follows:

$$P_g(\theta) = c_g + (1 - c_g) [1 + \exp\{-Da_g(\theta - b_g)\}]^{-1}, \quad (2.6)$$

where a_g is the item discrimination parameter, b_g is the item difficulty parameter, c_g is the guessing parameter, and D is a scaling factor, which is set equal to 1.7 when the logistic is used as an approximation to the normal ogive model. The item discrimination parameter is proportional to the slope of the item characteristic function at its point of inflexion. The item difficulty parameter is a location parameter which is given as the value of the latent trait corresponding to this point of inflexion. The third parameter, the item guessing parameter, is a positive value that indicates the probability of responding to the item correctly when guessing randomly. The guessing parameter defines the lower asymptote of the three-parameter logistic item characteristic function. The upper asymptote is equal to one. The item characteristic function in this model is equal to $(1 + c_g)/2$ at its inflexion point.

Another model frequently used with dichotomously scored items is the normal ogive model. The normal ogive is a two-parameter model for which the item characteristic function is given by

$$P_g(\theta) = (2\pi)^{-1/2} \int_{-\infty}^{a_g(\theta - b_g)} e^{-\frac{u^2}{2}} du, \quad (2.7)$$

Unlike the three-parameter logistic model, the normal ogive does not contain a guessing parameter. The item characteristic function in the

normal ogive model has a lower asymptote of zero, an upper asymptote of one, and is equal to 0.5 at its point of inflexion.

The theoretical justification for the normal ogive model has been discussed in detail by Lord and Novick (1968). To derive this model, we must make three assumptions regarding the relationship between the item response tendency, X_g , and the latent trait, θ : (a) the conditional distribution of X_g , given θ , is normal; (b) the conditional variance of X_g , given θ , is constant; and (c) the regression of X_g on θ is linear.

In order to estimate the item parameters based upon the responses of a group of examinees, we assume that ability is normally distributed for the group, and we then select the scale of measurement for X_g and for θ so that both variables are standardized. From the proportion of examinees correctly responding to each item and from the factor loadings obtained from the principal factor solution of factor analysis, the estimated item discrimination and item difficulty parameters can be obtained. The reader is referred to Lord and Novick (1968) for further discussion of these procedures.

Maximum likelihood estimation of ability. In classical test theory, the examinee's responses to the items of a test are generally summated into an overall test score, or raw score, that is considered to reflect the examinee's level of achievement or ability. In latent trait theory, on the other hand, information available from the complete pattern of responses to a set of items is generally used to estimate the ability of the examinee. Thus two examinees could have the same raw score yet have different ability estimates if their

pattern of responses are different.

Maximum likelihood estimation is typically applied in latent trait theory in the process of estimating the ability levels of the examinees. In order to obtain maximum likelihood estimates of ability, one must have an Old Test (Samejima, 1981), or a set of test items whose item characteristic functions are known. If the item characteristic functions are not known for the items administered, they can be estimated, assuming a particular mathematical model or models. For example, if the item characteristic function of each item in the Old Test is assumed to follow the normal ogive model, then the item discrimination parameter, a_g , and the item difficulty parameter, b_g , can be estimated following the method based upon the factor analysis of the tetrachoric correlation matrix described earlier. The methods by which the maximum likelihood estimate of ability can be obtained numerically when each item follows the normal ogive model have been previously described by Samejima (1969; 1981). Although many researchers assume a particular model without checking how well it fits the data, there are ways to check the validity of the mathematical model being used (Samejima, 1981; 1984a).

Information functions and the error of estimation. A number of information functions have been defined in latent trait theory, the utility of which have been pointed out by other researchers (Birnbaum, 1968; Samejima, 1977). Among these information functions are the item response information function (Samejima, 1969), the item information function (Birnbaum, 1968), the response pattern information function, and the test information function. The reader is referred to Samejima

(1977; 1981, Ch. 3) for a review of these functions, including the mathematical definitions.

Of particular utility is the test information function, from which the standard error of estimation of ability can be obtained. The conditional distribution of the maximum likelihood estimate of ability, $\hat{\theta}$, given ability, θ , is asymptotically normally distributed, with mean θ and the inverse of the square root of the test information function as its standard deviation (Birnbaum, 1968). Thus the standard error of estimation in latent trait theory is expressed as a function of ability, and is equal to the inverse of the square root of the test information function.

Plausibility functions. The operating characteristic, or conditional probability, of selecting a particular distractor, given ability, is known as the plausibility function for this distractor (Samejima, 1981). If the distractor is well developed, and informative, the probability of selecting the distractor is likely to increase as ability increases up to a certain point of θ and then to decrease (Samejima, 1984a).

Samejima (1981) has developed various methods and approaches for estimating the operating characteristics of discrete item responses. A method is defined as a way of approximating the conditional density function of ability θ , given its maximum likelihood estimate $\hat{\theta}$, and an approach is a way of approximating the ability distributions of separate discrete response groups (Samejima, 1981). In order to estimate the plausibility functions, a combination of an approach and a method is used. These procedures are necessary

because the true distribution of ability for a group of examinees can not be directly observed.

To obtain the plausibility functions using one of Samejima's methods and approaches, the test information function of the Old Test must be constant for a meaningful interval of the latent trait. If this function is not constant, the latent trait must be transformed in such a way that the test information is constant. The methods that Samejima has developed for this transformation are well summarized in her 1981 Final Report (Samejima, 1981, Ch. 3).

Latent Trait Procedures

Although the identification of informative distractors was made from the content analysis described earlier in this chapter, the confirmation of their existence, or non-existence, required the analysis of actual response data. The rankings of distractors based solely upon their content was subjective and depended upon the evaluators' interpretations and expectations of examinee behavior. The mathematical identification of informative distractors necessitated the estimation of the functional relationship between item alternative selection and the underlying ability. This relationship, which is expressed by the plausibility function, was estimated under Samejima's supervision from response data using latent trait methods described earlier in this chapter.

Tetrachoric method item parameter estimation. The first step in the process of estimating the plausibility functions for the items of Levels 11, 12 and 13 of the Iowa Vocabulary subtest was to

dichotomize the item responses of all examinees and to produce the tetrachoric correlation matrix from the resultant inter-item contingency tables for these three levels. These tetrachoric correlation coefficients were obtained using a program written by Samejima, FSPC-84-22, that makes a polynomial approximation to the correlation coefficient. In dichotomizing the item responses, Samejima decided to treat items that were unanswered by the examinee as incorrect responses, which was consistent with the way the tests were scored in actual administrations. For the purposes of item analysis, unfortunately, this will make certain items appear to be somewhat more difficult than they would if all items were attempted by all examinees. On the other hand, the items would have appeared easier than they should have if these non-responses had been excluded. This would have been true if the examinees who excluded items tended to be of relatively low ability and would have been likely to respond incorrectly to the items.

For Level 11 Vocabulary subtest the tetrachoric correlation matrix for the 43 items of this subtest was based upon the dichotomized responses of the 2,364 examinees described earlier in this chapter. The tetrachoric correlation matrix for the 46 items of Level 12 Vocabulary subtest was based upon 2,413 examinees and for the 48 items of Level 13 Vocabulary subtest was obtained from 2,469 examinees.

As mentioned, use of the tetrachoric correlation coefficient is based on the assumption that a bivariate normal distribution exists for the underlying item variables, or response tendencies, for each

pair of items. In order to test whether or not this assumption appeared to be met by the items in each of these subtest levels, a computer program written by Samejima, FSPC-84-20, was used. This program generates a theoretical two-by-two contingency table for each pair of items, assuming a bivariate normal distribution for the response tendencies, from the estimated tetrachoric correlation coefficient and the estimated normal deviates. These normal deviate estimates were obtained from the proportion correct for each item using Samejima's computer program, FSPC-84-28. A chi-square test between the theoretical and the observed contingency tables was then performed with program FSPC-84-20.

The principal factor solution of factor analysis was applied to the tetrachoric correlation matrix for each of the three data sets, Vocabulary subtest Levels 11, 12 and 13. The factor analysis was conducted using Biomedical Computer Programs Multivariate Analysis Series 4, or BMDP4M. The squared multiple correlations of each item with all other items was used as the initial communality estimate. The number of factors taken for each data set was the maximum number of factors that could be extracted while yielding non-negative estimated eigenvalues corresponding to each of these factors.

There were two primary purposes of the factor analysis. One purpose was to test the assumption of unidimensionality of the latent space for each of the three data sets, or levels. The eigenvalues corresponding to the various factors were analyzed to determine whether or not there was a dominating factor that would justify the use of unidimensional latent trait theory. In addition, the factor

patterns were examined to see if more than one item loaded highly on the same secondary factor, which would suggest that certain items might need to be eliminated to avoid clusters. The second major purpose of the factor analysis of the tetrachoric correlation matrix was to obtain the factor loadings on the first or principal factor, which is necessary in estimating the tetrachoric method item parameters in the normal ogive model. If a dominating factor exists, this factor is operationally defined as the latent trait, θ , and the first factor loading expresses the correlation between the item variable, X_g , and the latent trait θ .

The estimation of the plausibility functions following the methods and approaches developed by Samejima requires a set of items whose item characteristic functions are known, which is called the Old Test. Selection of items to be used as the Old Test for each subtest level in this study, and for which the item parameters in the normal ogive model would be estimated, was based upon a review of the results of the factor analysis and of the chi-square tests of the assumption that the response tendencies for each pair of items followed a bivariate normal distribution. Consideration was given to excluding items from the Old Test that did not load highly on the principal factor or that had large chi-square values indicating a violation of the assumption of a bivariate normal distribution. One exception, however, was that the chi-squares generated by program FSPC-84-20 were not produced until after the selection of the Old Test items for Level 11 Vocabulary and thus were not a criterion for this one subtest level (the program had not been written at that time). The chi-square

results produced by this program were considered with regard to item selection for Levels 12 and 13 of the Vocabulary subtest.

The item discrimination and item difficulty parameters following the normal ogive model were estimated from the principal factor loading, λ_g , and the estimated cutting point of the response tendency, γ_g , using the computer program FSPC-84-19 written by Samejima. The item parameters were estimated for each item selected as part of the Old Test for each of the three levels.

Ability estimation and transformation of the latent trait.

Given the normal ogive item parameters obtained in the previous step and the dichotomized response patterns for each examinee to the n items of the Old Test, the maximum likelihood estimate of ability, $\hat{\theta}$, was computed for each examinee with Samejima's computer program MLE Theta. As was done previously in the tetrachoric correlation coefficient computation, missing item responses were treated as incorrect answers for purposes of dichotomizing the response patterns to obtain the maximum likelihood estimates.

For the 43 items of Level 11 Vocabulary subtest, a very rough approximation to the plausibility functions was made based upon the maximum likelihood estimates of ability and the item responses of the examinees. The examinees were arranged in order of ascending magnitude of their maximum likelihood estimates and divided into 30 subgroups of approximately equal size. The proportion of examinees within a subgroup who selected a particular alternative was used as a crude estimate of the conditional probability of selecting the alternative, given ability. The mean maximum likelihood estimate for

the subgroup was used as the estimate of ability. The use of 30 subgroups was an arbitrary decision that was felt to provide large enough subgroups (78-79 examinees) that sampling error would not be excessive and yet would allow for the estimation of enough points (30) on the ability scale that significant trends could be detected. As has been pointed out by Samejima (1981), this method, utilizing the frequency ratios obtained from the examinee subgroups, is based upon the maximum likelihood estimate of ability rather than ability itself, and thus is not a substitute for the plausibility functions. The utility of the frequency ratio method is considered to be as a screening procedure, as it may provide a relatively easy way to select items for further analysis.

As was described in an earlier section of this chapter, the test information function for the Old Test must be constant over a meaningful interval of ability in order to use Samejima's methods for estimating the plausibility functions. When this information is not constant, a polynomial transformation of the latent trait, θ , is made to a latent trait, τ , which provides constant test information (Samejima, 1979). This polynomial transformation was made for each of the three levels investigated in this study using the computer program Tau Transformation, written by Samejima. A different polynomial, of degree 7, was selected for the transformation involving each level.

This same polynomial of degree 7 was used to transform the finite maximum likelihood estimates of ability based upon the original latent trait to $\hat{\tau}$. The conditional mean of τ , given $\hat{\tau}$, and the second through the fourth conditional moments about the mean of τ ,

given $\hat{\tau}$, were computed for each finite $\hat{\tau}$, through the application of the computer program Conditional Moments MLE, that was written by Samejima for this purpose. Examinees not having finite maximum likelihood estimates or for whom the estimated probability density function of $\hat{\tau}$ or the conditional moments assume a negative value are excluded by this program.

Estimation of the plausibility functions. From the conditional mean and the conditional variance of τ , given $\hat{\tau}$, the conditional density function of the latent trait τ , given $\hat{\tau}$, was estimated for each subgroup selecting a specific item alternative and for the total group, using a normal density function (refer to Samejima, 1981, Ch. 5, for mathematics). This estimation process as well as the estimation of the operating characteristic of each item alternative was obtained using Samejima's computer program Simple/Weighted Sum NT. As was stated earlier, the operating characteristic for any of the item distractors is called a plausibility function while the operating characteristic for the correct alternative is an estimated item characteristic function that does not assume any particular mathematical model (Samejima, 1981). This estimated item characteristic function was plotted in the same graph as the plausibility functions for each item to be compared with the normal ogive item characteristic function obtained from the tetrachoric method item parameters. If the estimated non-parametric item characteristic function and the item characteristic function obtained by the tetrachoric method are very similar, then the normal ogive model is validated for a particular item. If the two item

characteristic functions are not close to each other, however, the results suggest that a different model might be more appropriate.

CHAPTER III

RESULTS

Content Analysis

The first step in the content analysis was for the author to rank the alternatives of each item in the Vocabulary subtest from 1 to 4, with 1 being the correct answer and 4 being the distractor that is expected to be least plausible to high-ability examinees. These predictions are given in Table III.1, with the letter of the distractor expected to attract the highest ability level (rank of 2), the distractor predicted to attract the next-highest ability level (rank of 3), and the distractor predicted to be selected by the lowest level examinee (rank of 4) presented for each item. It should be pointed out again that ability levels reflected by these rankings are not necessarily the same from item to item. As the item difficulties vary so do the ability distributions for examinees selecting distractors. An asterisk is placed beside each distractor that was not predicted to be informative. For example, for item 24 in Table III.1 we see that distractor D was ranked 2 (highest ability), C was ranked 3, and B was ranked 4 (lowest ability), but only distractor D was predicted to be an informative distractor. Alternative A for item 24 is not presented in the table since it is the correct answer. It should be noted that the identification of item alternatives appearing in this manuscript is different from the identification or labels of the alternatives as they are actually used in the Iowa Tests, in order to protect the confidentiality of the test. In the actual test forms,

Table III.1 Predictions by Author of Informative Distractors
for Each of 82 Items of Levels 11, 12 and 13 of
Vocabulary Subtest.

	Item	Highest	Second Highest	Lowest
1	24	D	C*	B*
2	25	B	D*	C*
3	26	C	D*	B*
4	27	D	C	B*
5	28	B	C	D*
6	29	C	B	D*
7	30	B	D*	C*
8	31	D	B	C*
9	32	C	B	D*
10	33	C	B	D*
11	34	D	C*	B*
12	35	C	B	D*
13	36	D	B*	C*
14	37	C	D	B*
15	38	B	C	D*
16	39	C	B	D*
17	40	C	B	D*
18	41	C	B*	D*
19	42	C	D*	B*
20	43	B	D*	C*
21	44	B	C	D*
22	45	D	B	C*
23	46	C	B	D*
24	47	D	B	C*
25	48	C	B	D*
26	49	B	D	C*
27	50	B	C*	D*
28	51	B	C	D*
29	52	B	C	D*
30	53	C	D	B*
31	54	B	D	C*
32	55	B	D	C*
33	56	B	C	D*
34	57	C	B*	D*
35	58	B	C	D
36	59	D	C	B*
37	60	C	B*	D*
38	61	C	D	B*
39	62	D	B*	C*
40	63	D	B*	C*

TABLE III.1 (Continued)

	Item	Highest	Second Highest	Lowest
41	64	B	C*	D*
42	65	C	B	D*
43	66	C	B	D*
44	67	B	C*	D*
45	68	B	C*	D*
46	69	B	D*	C*
47	70	D	B	C*
48	71	C	B*	D*
49	72	D	B*	C*
50	73	D	B*	C*
51	74	B	D*	C*
52	75	D*	C*	B*
53	76	B	D*	C*
54	77	C	D	B*
55	78	D*	C*	B*
56	79	B	C	D*
57	80	D	B	C*
58	81	B	D*	C*
59	82	B	D*	C*
60	83	C	B*	D*
61	84	D	B*	C*
62	85	C	B*	D*
63	86	D	B	C*
64	87	B*	C*	D*
65	88	C	B*	D*
66	89	D*	C*	B*
67	90	B	C*	D*
68	91	B*	C*	D*
69	92	C*	D*	B*
70	93	D	C*	B*
71	94	C	B	D*
72	95	D	C*	B*
73	96	C	B*	D*
74	97	B	C*	D*
75	98	B	C	D*
76	99	D	B*	C*
77	100	C	B*	D*
78	101	C	B	D*
79	102	B	D	C*
80	103	D	B*	C*
81	104	B	C	D*
82	105	C	B	D*

the alternatives are numbered 1 through 4 for the Vocabulary subtest. In this manuscript, however, the alternatives are labeled A, B, C and D, with A always indicating the correct answer. Thus the three distractors for each item will always be identified as B, C and D. Since the correct answer is always A and since the correct answer was correctly identified by the author for each item when the rankings of alternatives was made, the alternative with the rank of 1 (correct) for each item is not given in Table III.1.

It can be seen in Table III.1 that 76 of the 82 items, or 92.6 percent, were predicted to have informative distractors, and 39 of these items, or 47.6 percent, were expected to have two or more informative distractors. The only items not predicted to be informative were items 75, 78, 87, 89, 91 and 92. For these items the distractors appeared to be either too implausible or too similar to be distinguishable, in terms of ability level to which they would be attractive, and therefore informative distractors were not anticipated.

The reasons for the author's predictions of informative distractors were reviewed after the rankings were completed and the similarity in these reasons indicated that they could be generalized. These generalizations can be expressed as six different reasons or rules for the selection of distractors of the Vocabulary subtest, and are as follow:

1. Similarity in meaning--The reason most often found to explain a choice of distractor was the similarity in meaning between the distractor and the word in heavy type in the phrase presented, or

the key word. While some alternatives had very little or no apparent relation to the key word, others were found to be somewhere along a continuum, with words synonymous to the key word on one extreme, and words opposite in meaning on the other. In some cases the similarity was fairly direct while in others some degree of inference was required. It was expected by the author that distractors that are close to the correct meaning would tend to be selected by examinees of relatively high ability, while words that are far from the correct meaning would most likely be chosen by lower ability examinees. As an example, alternative 1 of the following item, that is similar in format to items of the Iowa Vocabulary subtest, is considered to be closer in meaning to the key word (exceeded) than are the other distractors.

- a. Exceeded the limit
 - 1) reached
 - 2) surpassed
 - 3) knew
 - 4) under

2. Meaning taken out of context--The second major reason behind the predictions was also based upon the meanings of the key words and of the distractors, but in this case a word or a part of the word must be taken out of its context in order for there to be some similarity between the key word and the distractor. Examinees selecting a distractor falling into this category are fairly likely to have some idea of the meaning of the key word, at least in a particular context, although it may not be in the context in which the word is used in the item. Alternative 2 of item b below, for example, is plausible when the root "view," from "interview," is taken out of

context.

- b. A long interview
 - 1) session of questioning
 - 2) scene
 - 3) debate
 - 4) interval of time

3. Logical in context--The third reason identified was that the distractor seemed sensible or logical when used in the context or phrase presented, regardless of its similarity or dissimilarity in meaning to the key word. Examinees choosing such an alternative are likely to be using context clues, which are an important factor in vocabulary development. While the examinees may not know the correct meaning of the key word, there is evidence that they understand the meanings of other words in the phrase presented and of the words used in the distractors, in order to be able to pick up on context clues. Thus these examinees may be of relatively high ability with regard to Vocabulary development or achievement. Alternative 2 of item c below illustrates a distractor that is logical in the context presented, since it is the only alternative other than the correct answer that is something that can be heard.

- c. Heard the uproar
 - 1) unpopular
 - 2) tune
 - 3) uplift
 - 4) noise

4. Readability of the phrase--The fourth factor or reason behind alternative selection was the "readability" of the phrase when the key word was replaced by the distractor. While this reason is somewhat hard to explain, some distractors seem to fit well into the phrase while others make the phrase seem too wordy or too choppy.

Distractors that consist of more than one word, or of relatively long words, are more likely to be eliminated for this reason. Examinees who select a particular distractor following this criterion of readability may not be very familiar with the meanings of the words presented, but make their choice because the words seem to read more smoothly when a particular alternative is chosen. Examinees in this group may not be particularly good readers and therefore may have additional difficulty determining the meaning when the context or the phrase is lengthy. Alternative 4 of example item d is a distractor likely to be considered more readable than the other alternatives, in the author's opinion, when substituted for the key word.

- d. A random selection
 - 1) by chance
 - 2) rarely occurring
 - 3) methodical
 - 4) careful

5. Similarity in spelling--The similarity in spelling between the key word and the distractor was considered a likely reason that examinees who did not know the meanings of the alternatives or the key word would select a particular alternative. Some distractors begin with the same letter or letters as the key word, while others have letters in common in the root or at the end of the word. This reason was expected to attract examinees who have little understanding of the words presented and are thus low in vocabulary achievement, and who likely do not have a better reason upon which to base their selections. Alternative 4 of example item b (begins with "inter") and alternative 3 (begins with "up") or alternative 1 (begins with "u") of

example item c given previously illustrate distractors with some similarity in spelling to the key word.

6. Similarity in length--The similarity in the length of words or in the number of syllables or words, between the key word and the distractor, was the sixth reason identified. For some items, one distractor is a single word while the other distractors consist of two or more words. If two distractors each consist of single words, an examinee might choose one because it has the same number of letters or syllables or is closest in length to the key word. As with reason 5 (similarity in spelling), this reason for alternative selection is likely to be followed by low level examinees who can not base their choice on word meanings. Alternatives 3 and 4 of example item d are both single words, for example, and alternative 4 is close in length to the key word.

The alternatives of 16 Vocabulary subtest items were evaluated by university faculty or public school teachers and were ranked from 1 to 4, as was done by the author for all of the Vocabulary subtest items. The rankings by the author were combined with the rankings by the seven faculty/teacher evaluators and the mean of these eight rankings for each alternative of the 16 selected items is given in Table III.2. In addition to the mean, the standard deviation of the rankings by the eight evaluators is presented in parenthesis after each mean, as an indication of the agreement between evaluators for each alternative. Kendall's coefficient of concordance is presented as a measure of agreement for each item.

It is observed in Table III.2 that the standard deviation of

TABLE III.2 Mean and Standard Deviation of Evaluator Rankings, by Item Alternative, and Kendall's Coefficient of Concordance, W, for 16 Selected Items of Vocabulary Subtest.

Item		Alternative				Kendall's W
		A	B	C	D	
		Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	
1	27	1.375 (0.518)	4.000 (0.0)	2.500 (0.756)	2.125 (0.835)	0.731
2	37	1.000 (0.0)	2.625 (0.916)	2.750 (0.707)	3.625 (0.518)	0.719
3	40	1.000 (0.0)	2.375 (0.518)	2.750 (0.250)	3.875 (0.354)	0.844
4	53	1.000 (0.0)	3.125 (0.641)	2.500 (0.756)	3.375 (0.916)	0.681
5	56	1.000 (0.0)	2.625 (0.916)	2.750 (0.707)	3.625 (0.518)	0.719
6	59	1.000 (0.0)	4.000 (0.0)	3.000 (0.0)	2.000 (0.0)	1.000
7	60	1.000 (0.0)	3.125 (0.641)	2.125 (0.354)	3.750 (0.463)	0.716
8	64	1.000 (0.0)	2.500 (0.926)	2.875 (0.354)	3.625 (0.744)	0.731
9	65	1.000 (0.0)	2.875 (0.354)	3.375 (0.916)	2.750 (0.886)	0.656
10	67	1.000 (0.0)	2.500 (0.756)	3.375 (0.744)	3.125 (0.835)	0.681
11	77	1.000 (0.0)	2.750 (0.886)	3.375 (0.744)	2.875 (0.835)	0.644
12	84	1.250 (0.707)	3.250 (0.250)	3.500 (0.756)	2.000 (0.535)	0.675
13	86	1.000 (0.0)	2.375 (0.518)	4.000 (0.0)	2.625 (0.518)	0.906
14	94	1.000 (0.0)	2.875 (0.835)	2.625 (0.518)	3.500 (0.926)	0.681
15	100	1.125 (0.354)	3.125 (0.835)	2.500 (0.926)	3.250 (0.886)	0.569
16	104	1.000 (0.0)	2.500 (0.535)	2.500 (0.535)	4.000 (0.0)	0.900

the rankings for the correct answer, alternative A, is equal to 0.0 for 13 of the 16 items, but that at least one evaluator misidentified the correct answer for each of items 27, 84 and 100. Interestingly, three of the eight evaluators disagreed with the choice of the correct answer for item 27, two selecting distractor D and one selecting distractor C, yet all eight evaluators agreed that alternative B would attract the lowest level examinees. All three evaluators who did not select alternative A as the correct answer for item 27 gave alternative A a ranking of 2, indicating that they thought it was the distractor that would appeal to examinees of relatively high ability. One evaluator misidentified the correct answer for item 84, selecting alternative D as the correct answer and alternative C as the highest distractor. Alternative A only received a ranking of 3 from this evaluator. One evaluator identified distractor C as the correct answer for item 100 and gave a ranking of 2 to the actual correct answer. Item 59 is interesting in that all eight evaluators ranked the four alternatives in the same exact order. In addition to the three distractors for item 59 and B of item 27, no variation was seen in the rankings for distractor C of item 86 and D of item 104. The overall reliability of the rankings by the evaluators was quite high, yielding a Horst reliability coefficient of 0.816 .

The reasons given by the faculty/staff evaluators for their selections were generally quite consistent with the reasons given by the author earlier. All six of the major reasons presented earlier in this chapter were listed by at least one evaluator, with the similarity in meaning between the key word and the distractor being

the most common reason for selecting a distractor as informative. The idea that a distractor is similar in meaning to the key word is defined quite broadly by the author and in some cases refers to two words that are relatively similar concepts or share common connotations or associations. For example, the words "evil" and "selfish" have distinctly different meanings, but both have strong negative connotations and it would not be unusual to associate these two characteristics as being likely to occur together.

In addition to the six general reasons presented earlier, two other major reasons for selecting a distractor as being informative were identified by the faculty/teacher evaluators. These two reasons are in a sense related to some of the reasons or rules identified earlier, but are still distinct enough to be considered as separate categories. They will be numbered here as 7 and 8, as a continuation of the rules listed previously:

7. Opposite in meaning--Although it was stated earlier that a distractor which was far from the correct meaning was expected by the author to attract low level examinees, a case can be made that a distractor which is opposite in meaning to the key word may often be selected by examinees of relatively high ability. These two words may involve a similar concept, or may be considered to be at different ends of the same dimension, as was discussed in Chapter I in reference to the study by Smith (1982). Alternatives that imply a similar concept might cue the examinee if he has enough ability to see the underlying relation or concept. If the examinee does not know the meaning of the key word but does know the meanings of the distractors,

it would probably make sense to choose one of two alternatives that were known to be opposite to each other. On the other hand, as one faculty/teacher evaluator noted, it is unlikely that either of two alternatives that are extremely close in meaning will be the correct answer, since there should only be one correct answer to an item. Alternative 3 of example item d illustrates a distractor that is opposite to the key word.

8. Familiarity--There was agreement among some of the evaluators that certain distractors might tend to be selected because they were more familiar words to the examinee than were the other alternatives. This is consistent with the findings of Strang (1977), cited in Chapter I, in which less familiar alternatives were found to be less frequently selected. Words that are not familiar are thus not well understood and probably will be only be selected as the result of guessing. Alternatives 2 and 3 of example item d given earlier in this section are less likely to be familiar to lower ability examinees in elementary school or junior high than is distractor 4, which might be chosen simply because it is a more familiar word.

Reasons 2 and 3 for selection of certain distractors presented earlier in this section deal with the context of the key word, such as how context clues may be used by the examinee or how a word may be taken out of context. One evaluator emphasized the importance of the context in his own evaluations, and noted that when a short phrase was presented with the key word, he found that he tended to expand the phrase into a complete sentence in order answer the item. By making the key word more meaningful in this way, however, he realized that he

might actually be making the item more complex than it should be, or might be changing the meaning somewhat by altering the context. In this way, bright or creative examinees might actually be penalized for reading more into an item than was intended, or because there may be a variety of associations that they can make to the key word that a lower ability examinee could not. This type of situation would appear less likely for items in which the key word actually occurred as part of a complete sentence or a nearly complete sentence.

Latent Trait Analysis

Tetrachoric correlation. The tetrachoric correlation coefficients generated for each of Levels 11, 12 and 13 were, with very few exceptions, non-negative, with the vast majority being in the range from 0.100 to 0.400 . There were several correlations of 0.0 observed within each subtest, and it should be noted that the computer program used for the tetrachoric correlation coefficient estimation, FSPC-84-22, automatically sets the estimated coefficient equal to 0.0 when the initial estimate is very close to zero.

For Level 11 Vocabulary there were no negative tetrachoric correlations observed, and the correlations ranged in value from 0.0 to 0.491 . The chi-square values obtained from computer program FSPC-84-20, which tests the assumption that the response tendencies for each pair of items follow a bivariate normal distribution, were generally quite small. Exceptions occurred for items 24 and 44, however. The total sum of the chi-squares obtained for these items against the other 42 items of the subtest were 4.364 and 5.186 ,

respectively, while all other items had total chi-squares less than 1.5 . For most of the other 41 items in Level 11 Vocabulary, the largest chi-square value observed was paired either with item 24 or item 44. These results indicate that the assumption of bivariate normality was likely violated with these two items. As was stated in Chapter II, completion of program FSPC-84-20 and the chi-square computation for Level 11 Vocabulary came late in the plausibility function estimation process and was not used to eliminate certain items from the Old Test for this subtest level.

The tetrachoric correlation coefficients for Level 12 Vocabulary subtest were all non-negative, and varied from 0.0 to 0.549 . The chi-squares obtained from FSPC-84-20 were generally quite small, with the sum of the chi-squares for each item against the other 45 items of this subtest mostly under 1.0 . The total chi-squares for items 72 and 74, on the other hand, were 7.760 and 4.654 , respectively, and these two items were not used as part of the Old Test due to these large values. Except for items 72 and 74, the largest sum of chi-squares observed was 3.174 , for item 44, which also produced a relatively large chi-square total for Level 11.

There were two negative tetrachoric correlations obtained for Level 13 Vocabulary, and both involved item 102. Both negative correlations were very close to zero, having values of -0.014 , for item 102 with item 67, and -0.030 , for item 102 with item 70. Overall, the range of estimated correlations was -0.030 to 0.533 . The sums of the chi-squares from program FSPC-84-20 were quite large for items 74, 100 and 102, with the values of 13.847 , 10.114 and

6.693 , respectively. Due to this relatively strong indication of the violation of the assumption of a bivariate normal distribution for the response tendencies of these items against the other items in the subtest, these three items were excluded from the Old Test for Level 13. The sums of the chi-squares for the remaining 47 items of the subtest were all less than 4.0 .

Factor analysis. The principal factor solution of factor analysis applied to the tetrachoric correlation matrix for the 43 items of Level 11 Vocabulary subtest gave strong support for the unidimensionality of the latent space. Forty factors were extracted in the factor analysis. The eigenvalue corresponding to the first factor was much larger than those of the succeeding factors, and there was a steady decline in the eigenvalues after the first factor. The eigenvalues dropped from 11.417 for the first factor to 1.040 for the second factor. The factor loadings for the two items having the largest chi-square totals when the assumption of bivariate normality was tested, items 24 and 44, had by far the smallest factor loadings on the first factor. The factor loading on the first factor, operationally defined as the latent trait, θ , was 0.192 for item 24 and was 0.142 for item 44. The factor loadings for the other 41 items in this subtest ranged from 0.327 for item 50 to 0.691 for item 30.

After exclusion of items 72 and 74 of Level 12 Vocabulary subtest due to their apparent violation of the assumption of bivariate normality for the joint distribution of the response tendencies, factor analysis based upon the remaining 44 items showed a dominating

factor. Eigenvalues dropped from 11.138 to 1.092 between the first factor and the second factor, and then dropped slowly thereafter. The first factor loadings for Level 12 ranged from 0.223 for item 44 to 0.745 for item 48.

The eigenvalues for Level 13 Vocabulary subtest did not drop quite as sharply from factors one to two as did those of the earlier subtest levels, but still indicate a strong first factor. There was also a slightly larger drop between the second and third factors than was observed with Levels 11 and 12. The eigenvalues for the first three factors were 11.035 , 1.380 , and 0.847 , respectively. The smallest principal factor loading observed for this level was 0.210 , for item 96, and the largest was 0.705 , for item 90.

Item parameter computation. Based upon the factor loading on the principal factor and the proportion of examinees correctly answering the item, the item discrimination parameter, a_g , and the item difficulty parameter, b_g , in the normal ogive model were estimated for each of the three levels. Theoretically a_g and b_g can range from negative to positive infinity, but with multiple-choice items typically $a_g < 2.0$ and b_g ranges from approximately -3.0 to 3.0 , when ability is scaled to have a mean and standard deviation of 0 and 1 , respectively (Harris, 1989). Since θ was standardized during the estimation of the item parameters, we should expect item parameters in those ranges.

In the normal ogive model, defined by equation (2.7), b_g is the value of ability θ for which the item characteristic function is equal to 0.5 . For example, when θ is standardized and $b_g = -2.0$,

an examinee whose ability is two standard deviations below the mean has a 50-50 chance of correctly answering the item. This would be a relatively easy item whereas an item with a positive b_g would be relatively difficult for the group as a whole.

The item discrimination parameter a_g is proportional to the slope of the normal ogive item characteristic function at its point of inflexion, such that the slope is equal to $a_g(2\pi)^{-1/2}$. The larger the discrimination parameter, the steeper the item characteristic function and the greater the change in probability of success as one goes up or down the ability scale. To illustrate, consider an example in which three items differ in item discrimination but all have an item difficulty parameter equal to 0.0. For the first item, $a_g=0.25$ (a rather flat curve), $P_g(-1.0)=0.401$ and $P_g(1.0)=0.599$. In other words, the probability of correctly answering item g increases by only 0.198 as ability increases from one standard deviation below the mean to one standard deviation above. For the second item in the example, $a_g=0.5$, $P_g(-1.0)=0.309$ and $P_g(1.0)=0.691$, an increase in probability of 0.382. For the third item, $a_g=1.0$ (a moderately steep curve), the increase in probability is 0.682 as ability increases from -1.0 ($P_g(-1.0)=0.159$) to 1.0 ($P_g(1.0)=0.841$).

Table III.3 presents the estimated item parameters for the 43 items of Level 11 Vocabulary subtest. The estimated item discrimination parameter, a_g , is proportional to the principal factor loading, as can be seen from equation (2.9), and therefore items 24 and 44, which had the lowest factor loadings of the 43 items

TABLE III.3 Estimated Item Parameters for Each of 43 Items
of Level 11 Vocabulary Subtest.

Item	Discrimination	Difficulty
24	0.196	-4.257
25	0.829	-1.000
26	0.614	-0.821
27	0.594	-0.340
28	0.669	-0.900
29	0.867	-1.077
30	0.956	-0.557
31	0.938	-0.179
32	0.940	-0.803
33	0.434	-2.331
34	0.598	-1.210
35	0.489	-0.569
36	0.657	-0.987
37	0.351	0.577
38	0.665	-0.468
39	0.333	-0.676
40	0.683	0.402
41	0.531	-0.948
42	0.436	0.258
43	0.672	-0.867
44	0.143	4.175
45	0.898	-0.357
46	0.612	-0.318
47	0.494	-0.781
48	0.849	0.054
49	0.421	-0.626
50	0.346	-0.250
51	0.664	-0.420
52	0.640	0.217
53	0.402	0.526
54	0.573	0.126
55	0.667	-0.342
56	0.593	1.007
57	0.370	0.398
58	0.416	0.782
59	0.491	-0.731
60	0.678	-0.170
61	0.519	0.748
62	0.938	-0.485
63	0.637	-0.398
64	0.818	-0.042
65	0.606	0.595
66	0.604	-0.376

in the subtest, also had the smallest item discrimination parameters. The estimated discrimination parameters for items 24 and 44 were 0.196 and 0.143 , respectively, while the next lowest estimate was 0.333 for item 39. The largest discrimination parameter estimate was 0.956 for item 30, that was earlier seen to have the largest factor loading. The estimated difficulty parameter estimates ranged in value from -4.257 for item 24 to 4.175 for item 44. Thus we see that the two items with the smallest discrimination estimates, items 24 and 44, have the largest difficulty parameters in terms of absolute value.

For Level 12 Vocabulary subtest the item discrimination estimates ranged from 0.273 for item 67 to 1.117 for item 48. The difficulty parameter estimates varied from a low of -2.034 for item 41 to a high value of 2.187 for item 77. The estimated item parameters for all 44 Old Test items are given in Table III.4.

Table III.5 presents the estimated item parameters for Level 13 Vocabulary subtest, for all 45 items used as the Old Test. The discrimination parameter estimates range from 0.215 for item 96 to 0.994 for item 90. Difficulty parameter estimates vary from -2.728 for item 59 to 1.710 for item 72.

Ability estimation. Out of the 2,364 examinees in the revised data for Level 11, three correctly answered all 43 items of the Vocabulary subtest and therefore obtained maximum likelihood estimates of positive infinity. None of the examinees for Level 11 Vocabulary subtest, or for any of the other three subtest levels examined in this study, incorrectly answered all of the Old Test items and obtained negative infinity as the maximum likelihood estimate of ability. The

TABLE III.4 Estimated Item Parameters for Each of 44 Old
Test Items of Level 12 Vocabulary Subtest.

Item	Discrimination	Difficulty
40	0.803	-0.261
41	0.511	-2.034
42	0.376	-0.140
43	0.712	-1.397
44	0.229	1.571
45	1.000	-0.995
46	0.690	-1.065
47	0.576	-1.440
48	1.117	-0.777
49	0.351	-1.227
50	0.483	-0.823
51	0.637	-0.959
52	0.865	-0.284
53	0.479	-0.013
54	0.714	-0.599
55	0.633	-0.888
56	0.683	0.505
57	0.456	-0.842
58	0.505	-0.001
59	0.312	-1.844
60	0.662	-0.791
61	0.514	0.142
62	0.886	-1.358
63	0.579	-1.342
64	0.794	-0.583
65	0.738	-0.116
66	0.766	-1.222
67	0.274	1.834
68	0.444	-0.472
69	0.623	-0.571
70	0.365	-0.421
71	0.497	-0.800
73	0.678	-0.363
75	0.874	-0.963
76	0.328	-1.144
77	0.331	2.187
78	0.701	-0.475
79	0.620	0.133
80	0.535	0.010
81	0.464	0.587
82	0.333	-0.748
83	0.363	0.797
84	0.564	0.756
85	0.570	0.610

TABLE III.5 Estimated Item Parameters for Each of 45 Old
Test Items of Level 13 Vocabulary Subtest.

Item	Discrimination	Difficulty
58	0.521	-0.572
59	0.241	-2.728
60	0.708	-0.978
61	0.599	-0.405
62	0.888	-1.770
63	0.567	-2.034
64	0.780	-0.854
65	0.766	-0.496
66	0.842	-1.722
67	0.356	0.596
68	0.518	-0.922
69	0.587	-1.349
70	0.376	-1.300
71	0.460	-1.390
72	0.224	1.710
73	0.770	-1.002
75	0.796	-1.599
76	0.298	-2.065
77	0.547	1.200
78	0.676	-1.046
79	0.573	-0.367
80	0.653	-0.468
81	0.606	0.107
82	0.301	-1.442
83	0.266	0.810
84	0.455	0.448
85	0.493	-0.204
86	0.302	1.599
87	0.784	-0.309
88	0.655	-0.768
89	0.637	-0.393
90	0.994	-0.157
91	0.650	-0.468
92	0.967	-1.180
93	0.553	-0.016
94	0.493	0.625
95	0.735	0.109
96	0.215	-0.696
97	0.736	-0.138
98	0.683	0.618
99	0.341	-0.055
101	0.514	0.462
103	0.459	-0.632
104	0.491	0.936
105	0.455	-0.608

range of finite maximum likelihood estimates for Level 11 Vocabulary was -2.674 to 5.948 .

For Level 12 Vocabulary subtest six examinees out of 2,413 obtained maximum likelihood estimates of positive infinity, and the range of finite estimates was -2.802 to 4.903 .

Two of the 2,469 examinees for Level 13 correctly answered all 45 Old Test items for Level 13 Vocabulary subtest. The finite maximum likelihood estimates ranged from -3.119 to 5.206 .

Information functions and transformation of the latent trait.

The square root of the test information function, $[I(\theta)]^{1/2}$, and the polynomial that was judged to be the best approximation over a meaningful range of ability θ , is presented in Figure III.1 for each of the three levels used in this study. In each graph the square root of the test information function is plotted as a solid curve and its approximation as a dotted curve. The polynomial approximation in each case is of degree 7.

The latent trait θ for each level was transformed to the latent trait τ having an approximately constant test information function over a meaningful interval. For each level of the Vocabulary subtest the target constant, C , of the square root of the test information function, $[I^*(\tau)]^{1/2}$, was 4.5 .

Plausibility functions. Within each of the three levels, the maximum likelihood estimate of ability on the original latent trait, denoted by $\hat{\theta}$, was transformed to $\hat{\tau}$, and the conditional mean and the conditional variance of τ , given $\hat{\tau}$, were computed for each

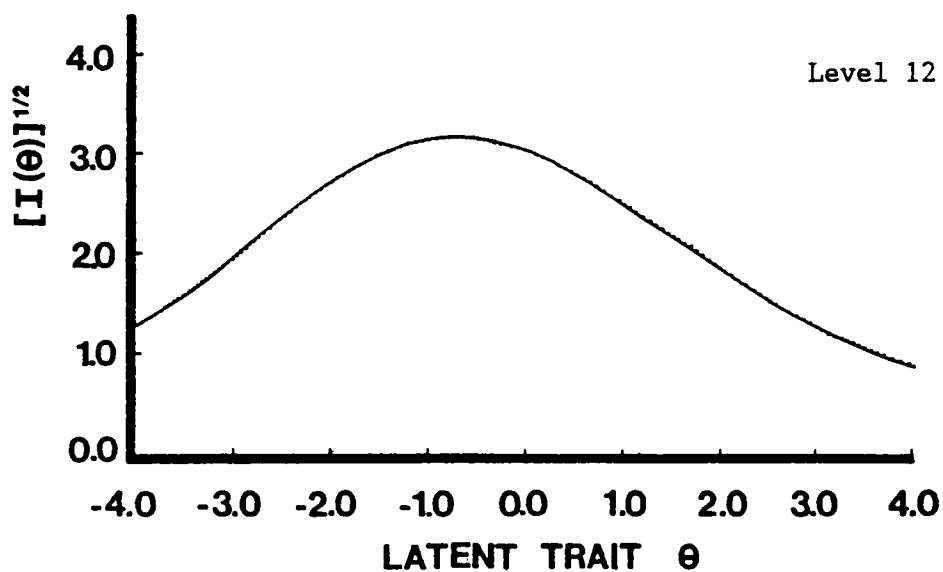
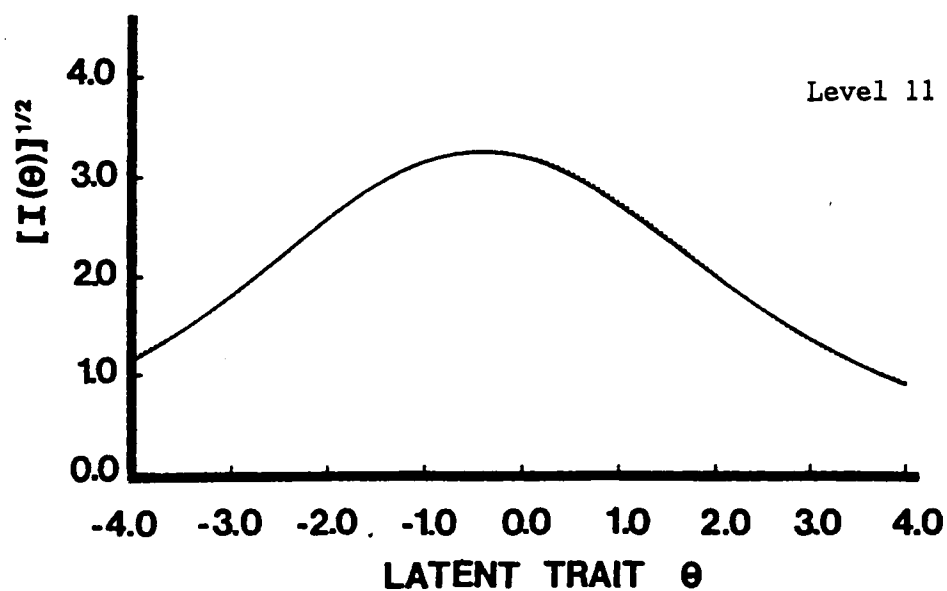


FIGURE III.1 Square Root of the Test Information Function, $[I(\theta)]^{1/2}$, and Its Approximation by the Polynomial of Degree 7 Obtained by the Method of Moments for Each of Levels 11, 12 and 13 of the Vocabulary Subtest.

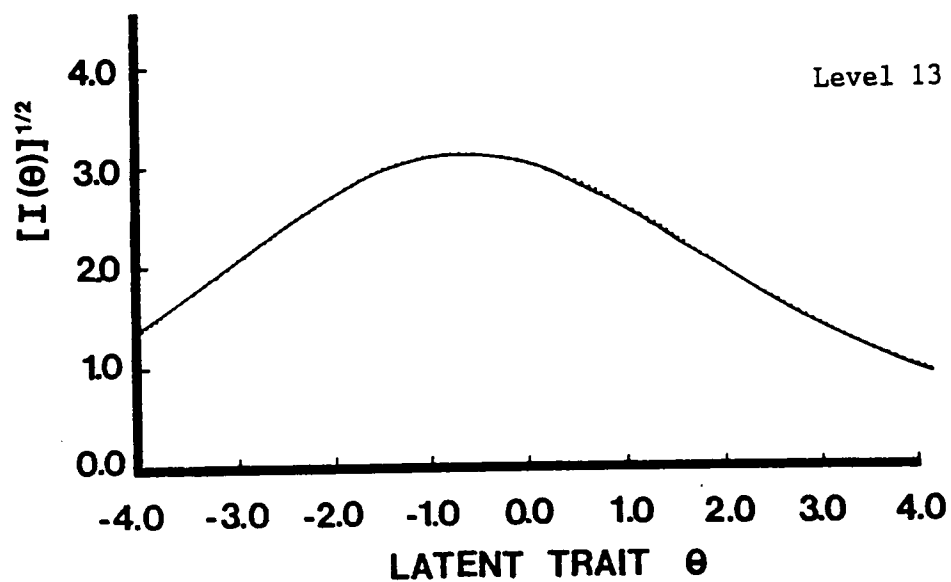


FIGURE III.1 (Continued)

examinee, as was discussed in Chapter II. Examinees who had positive infinity as their maximum likelihood estimate of ability or for whom the probability density function of $\hat{\tau}$, as estimated by a polynomial of degree 4, was negative, or for whom the second or fourth conditional moments about the mean of τ , given $\hat{\tau}$, were negative were excluded from the group of examinees upon which the operating characteristics (plausibility functions and non-parametric item characteristic function) were estimated. The estimation is based upon well over 2,000 examinees in each case after the exclusion process.

Figure III.2 presents the operating characteristics for each of the 82 items of Levels 11, 12 and 13 of the Vocabulary subtest, together with the item characteristic function in the normal ogive model, which is shown by a solid curve. The non-parametric item characteristic function, or the operating characteristic for the correct answer, alternative A, is shown as a dotted curve in each graph. Alternative B for each item is shown as a short dashed curve, alternative C as a dashed curve, and alternative D as a long dashed curve. For items that were part of the Old Test, the normal ogive item parameters were those obtained by the tetrachoric method and which were presented earlier in this chapter. For those few items in Level 12 and 13 that were not Old Test items, the normal ogive item parameters were estimated by finding the normal ogive item characteristic function that best fit, in the least squares sense, the weighted discrete points of the non-parametric estimated item characteristic function. These discrete points were weighted by the theoretical probability density function of θ , which is a standard

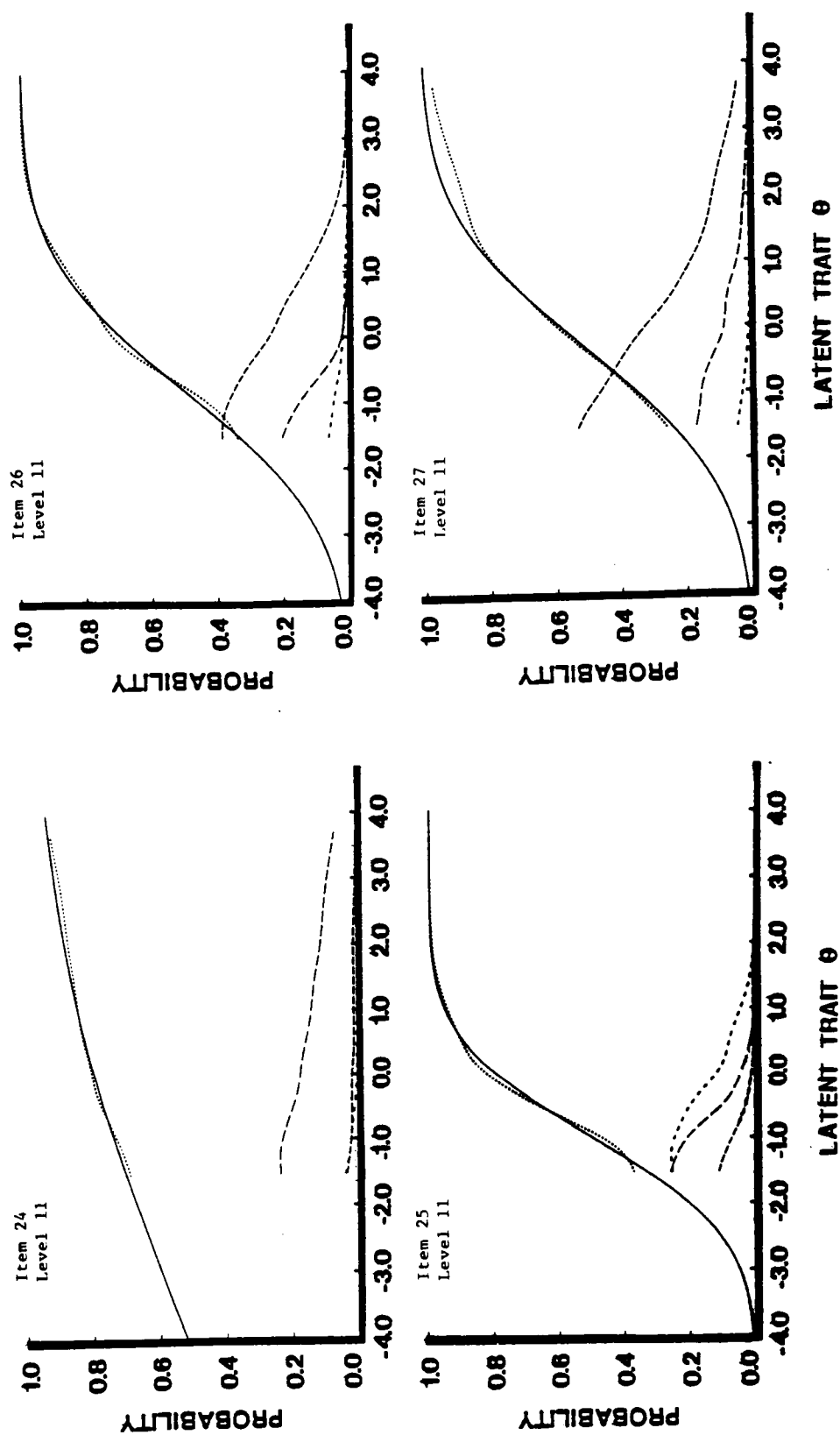


FIGURE III.2 Estimated Item Characteristic Function Following the Normal Ogive Model (Solid Curve) and the Non-parametric Estimated Item Characteristic Function (Dotted Curve) and Plausibility Functions (Dashed Curves) by the Simple Sum Procedure/Normal Approach Method, for Each of 82 Items of Vocabulary Subtest.

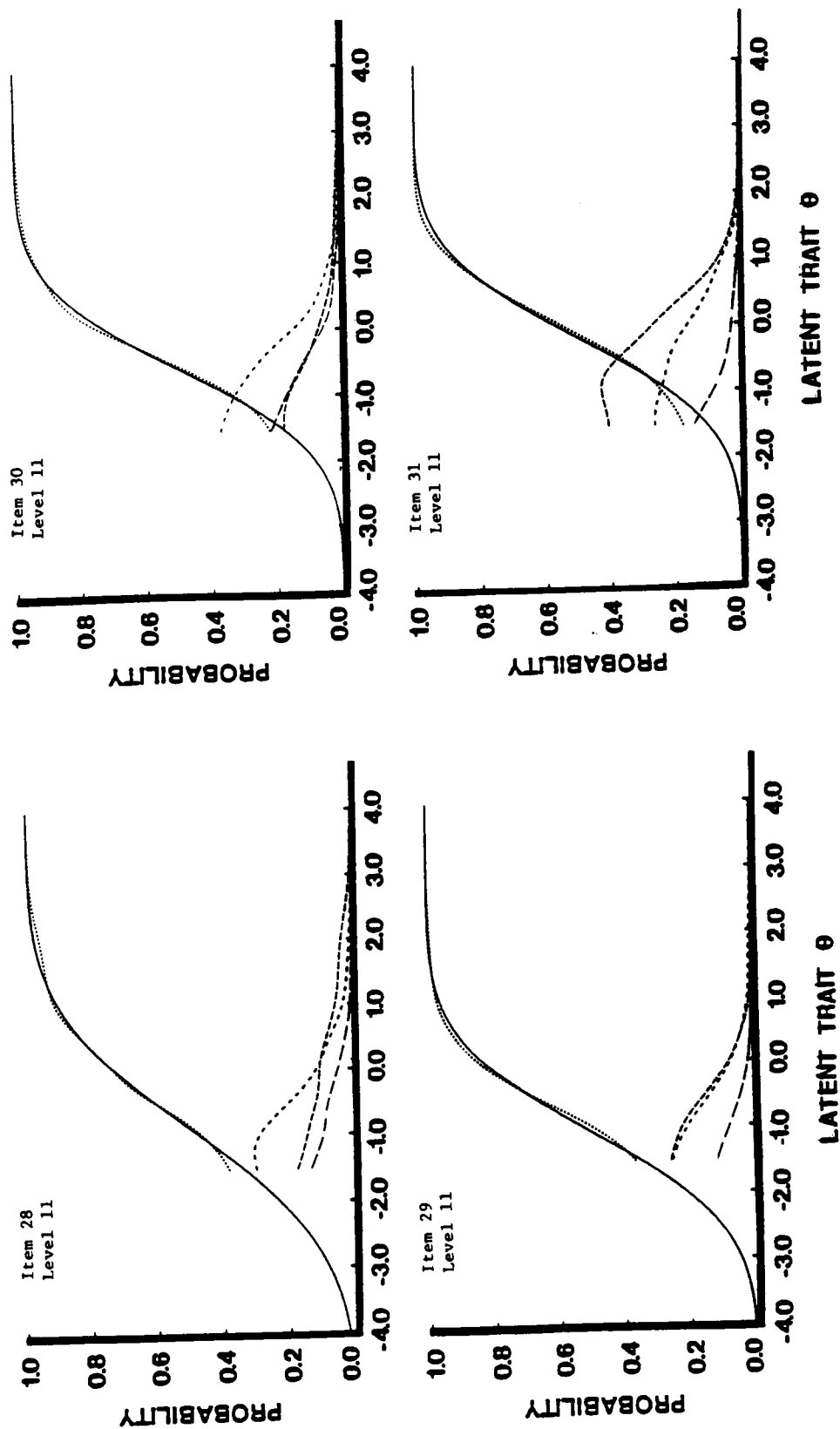


FIGURE III.2 (Continued)

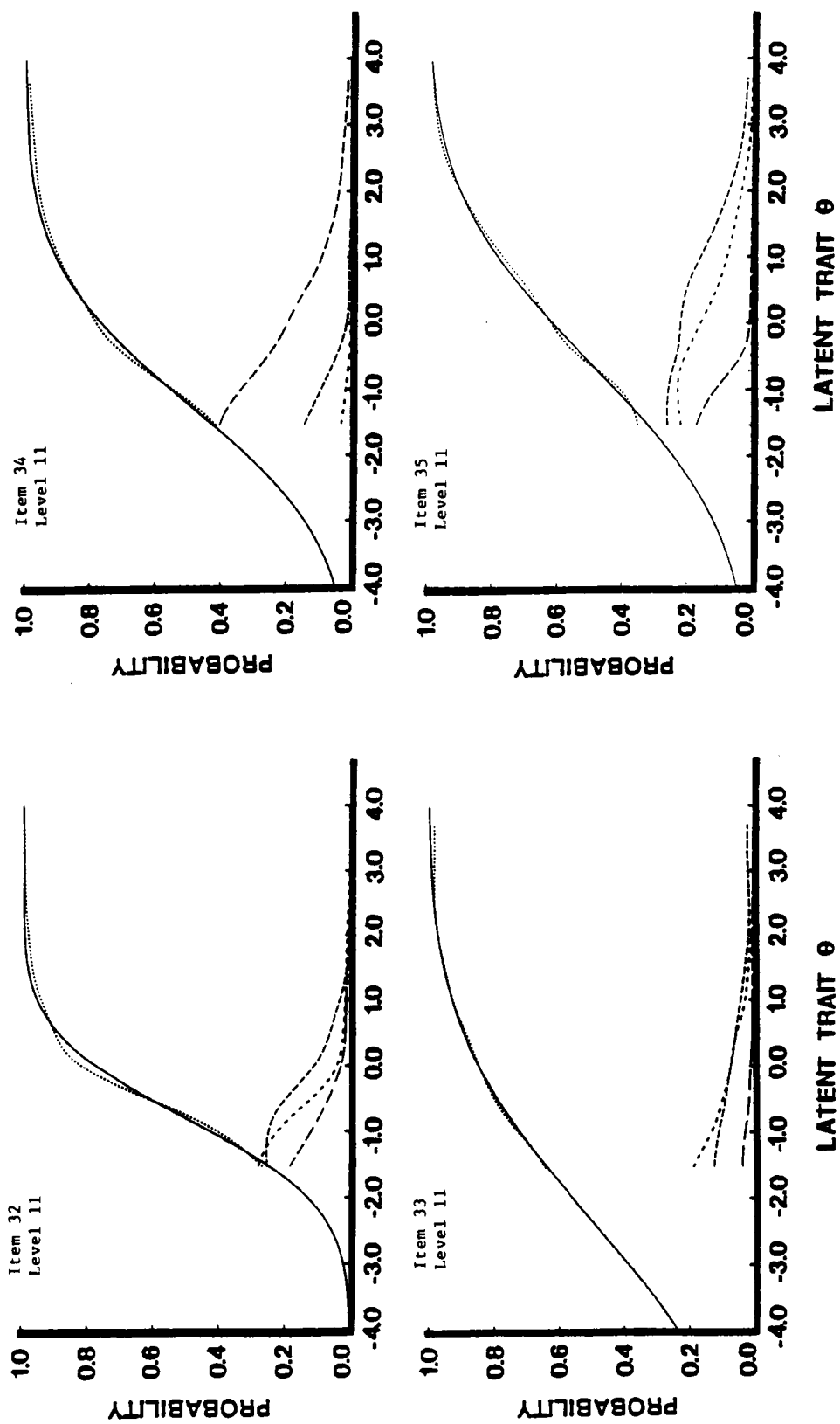


FIGURE III.2 (Continued)

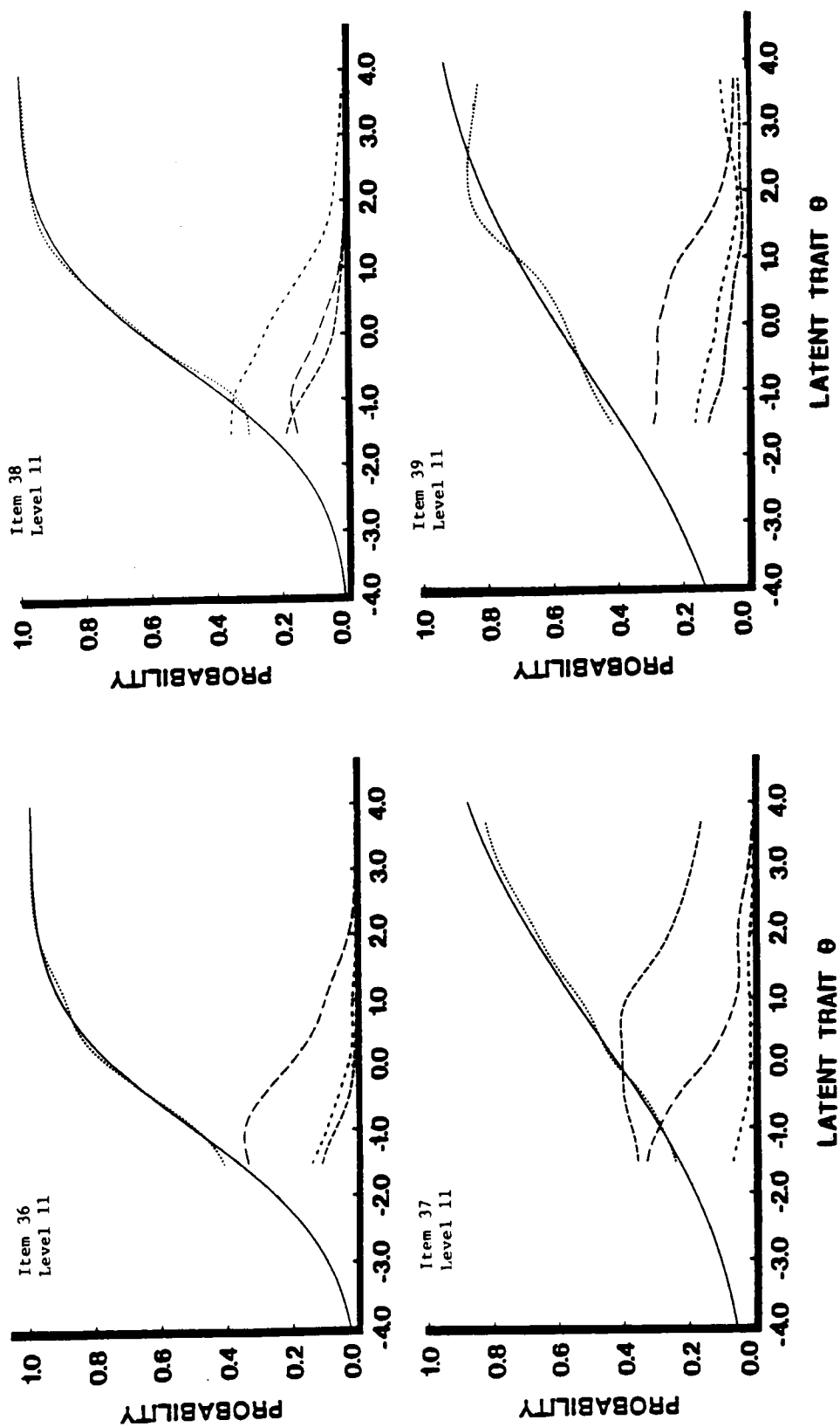


FIGURE III.2 (Continued)

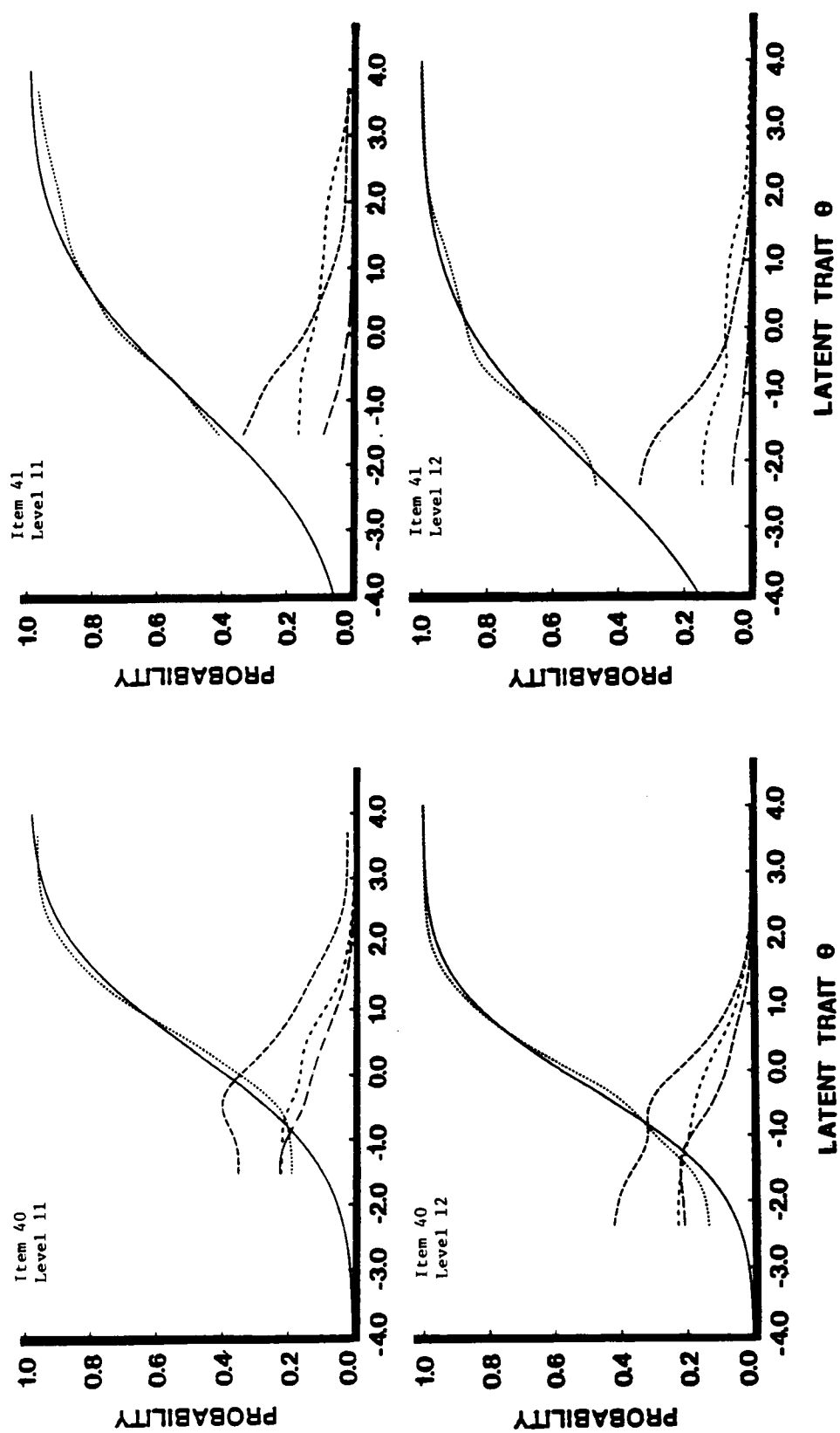


FIGURE III.2 (Continued)

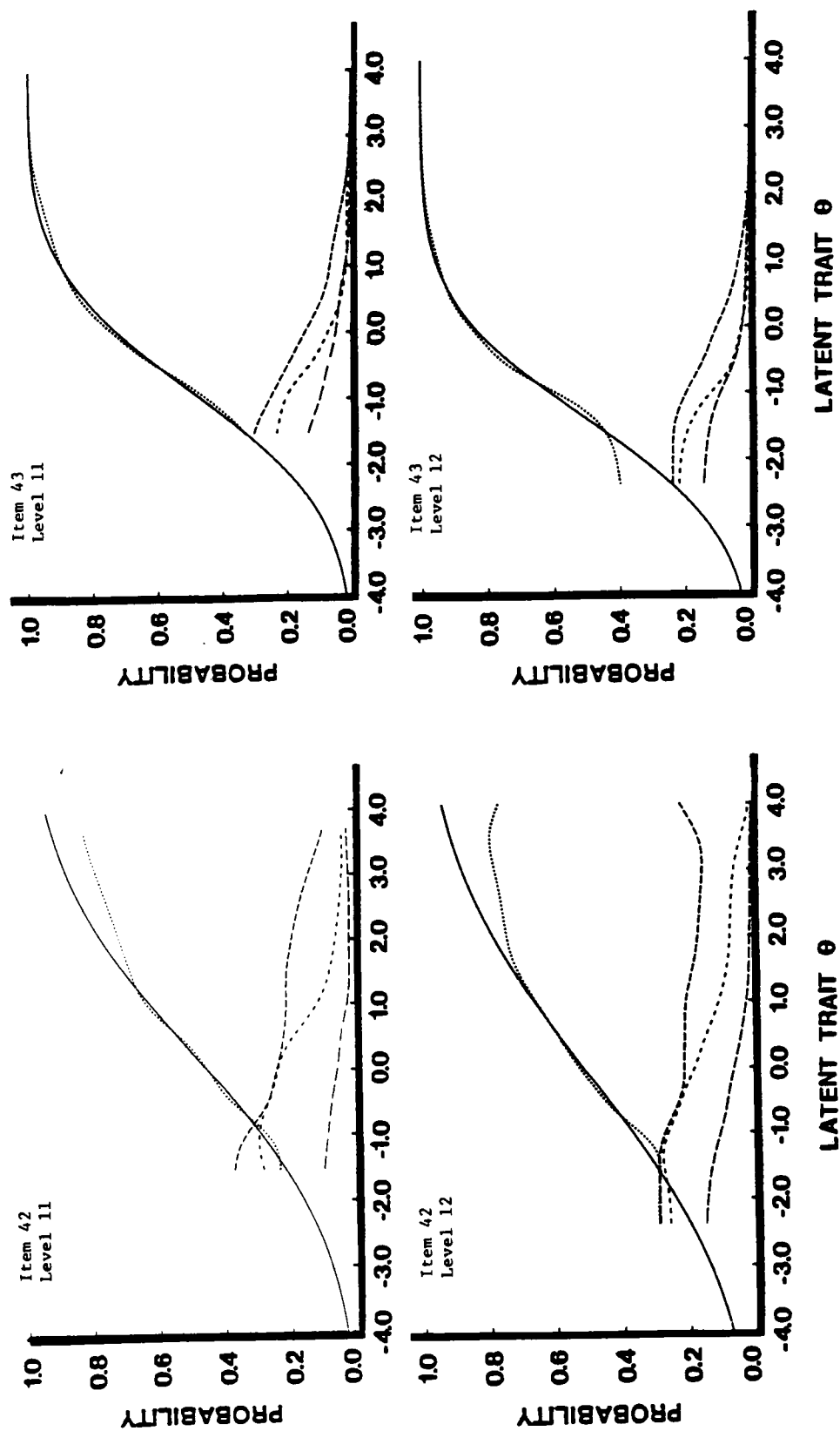


FIGURE III.2 (Continued)

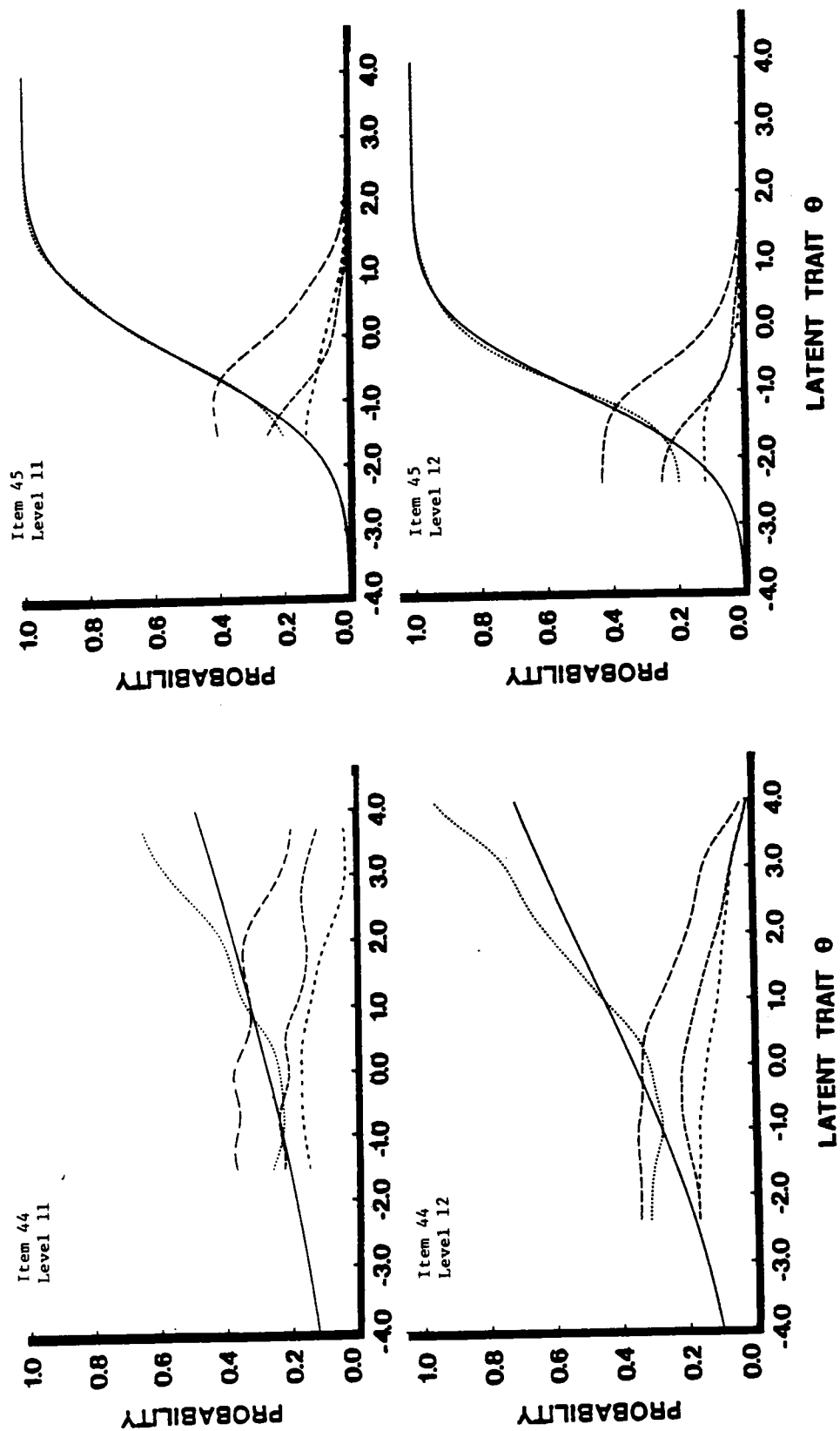


FIGURE III.2 (Continued)

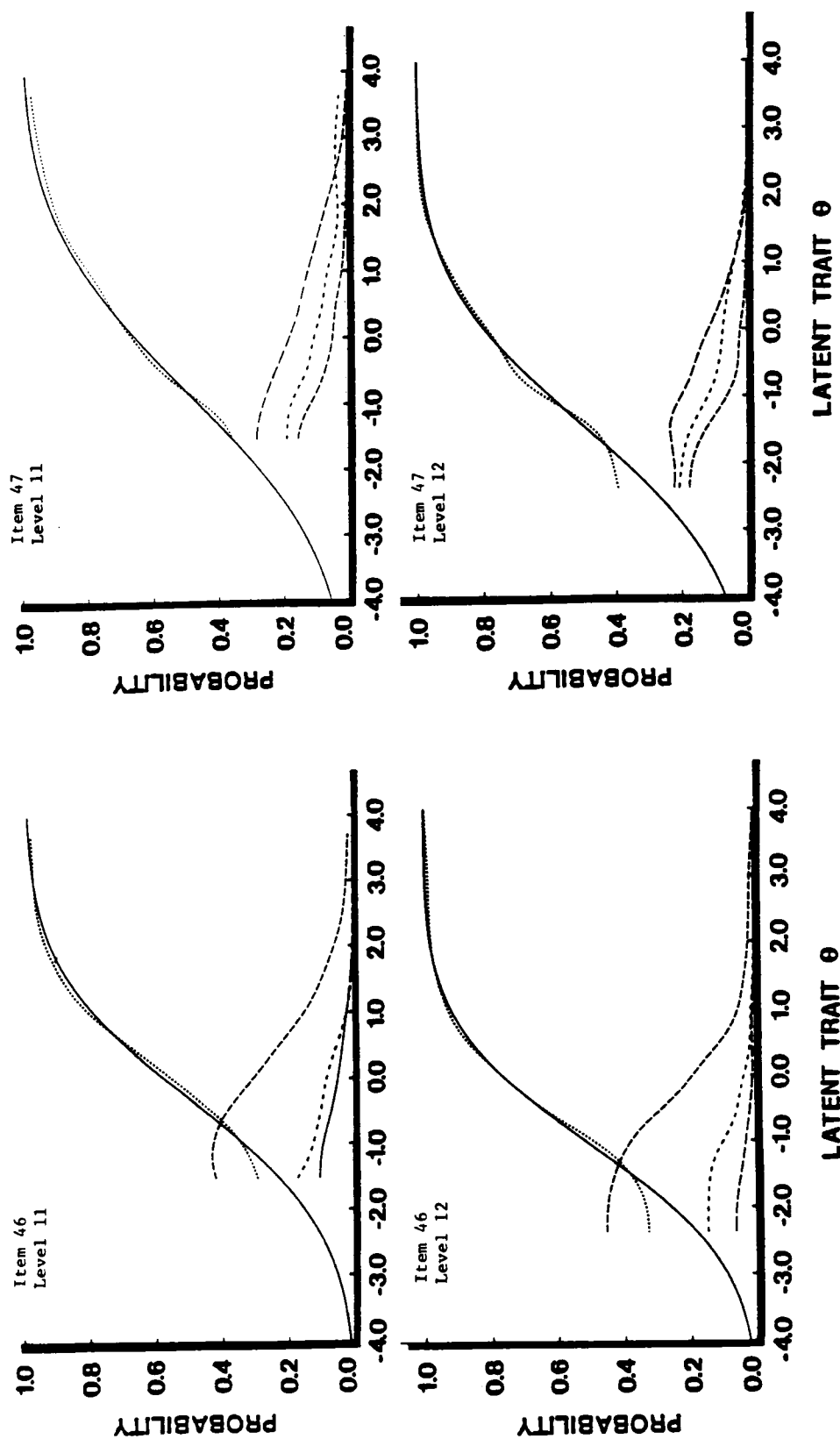


FIGURE III.2 (Continued)

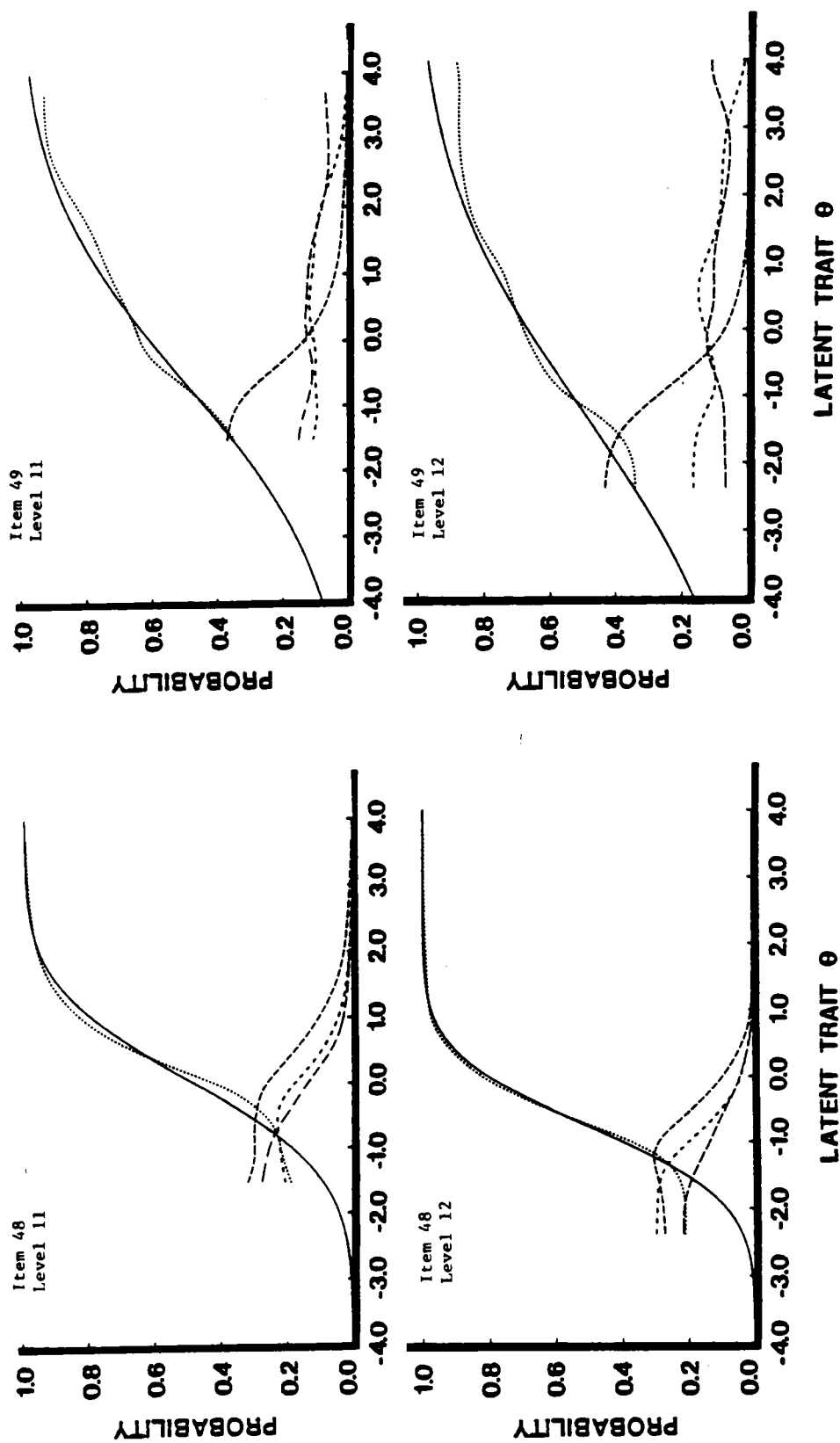


FIGURE III.2 (Continued)

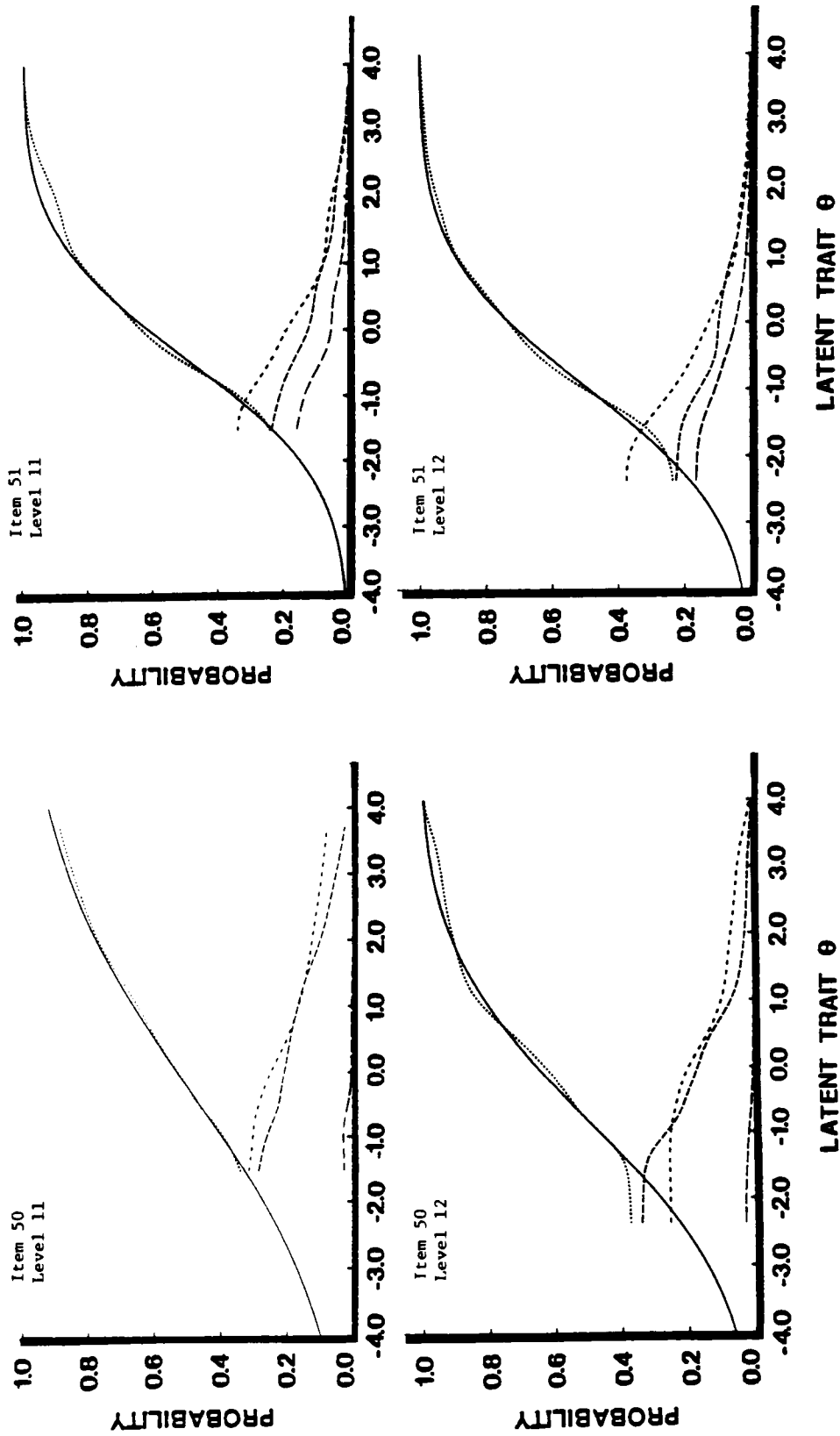


FIGURE III.2 (Continued)

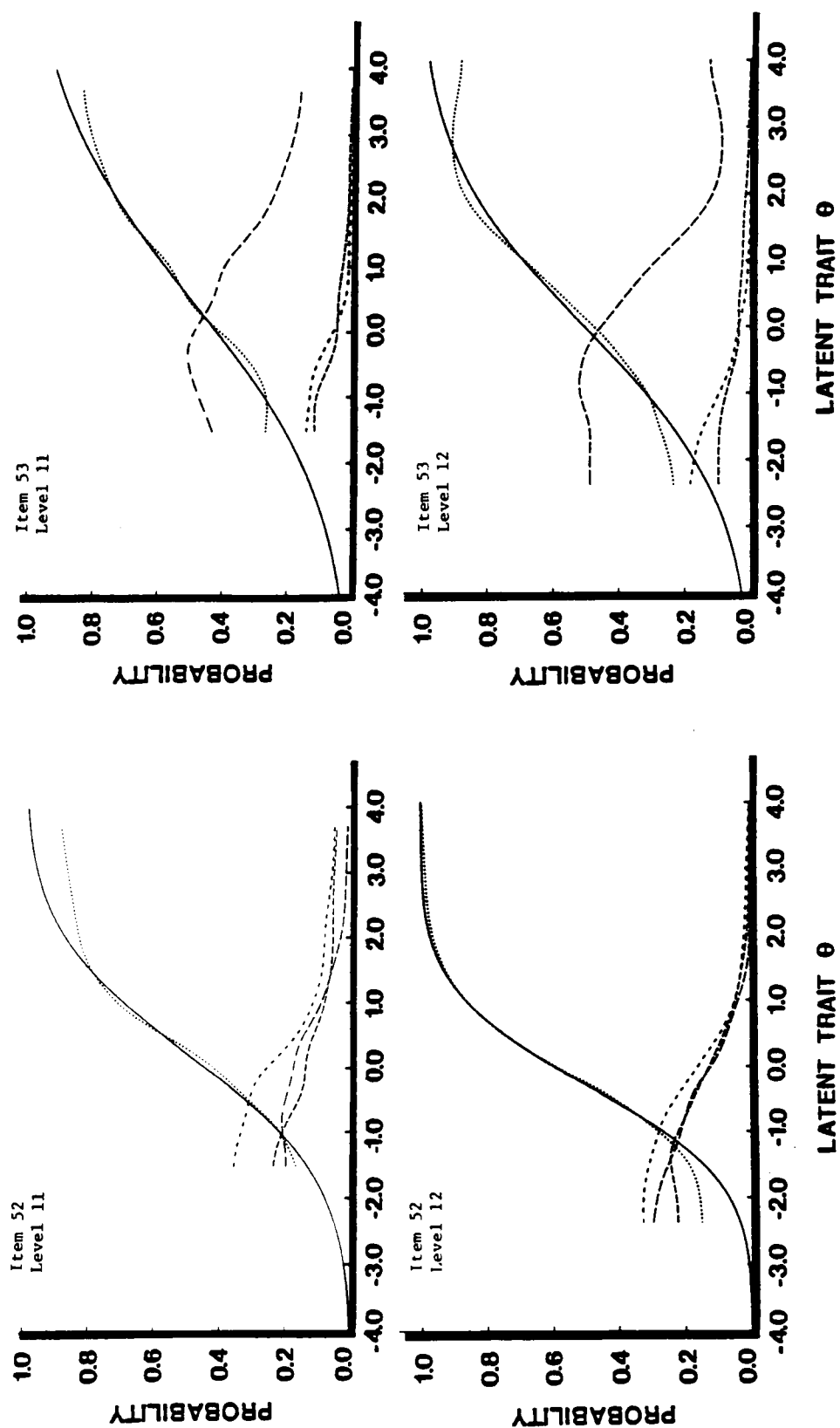


FIGURE III.2 (Continued)

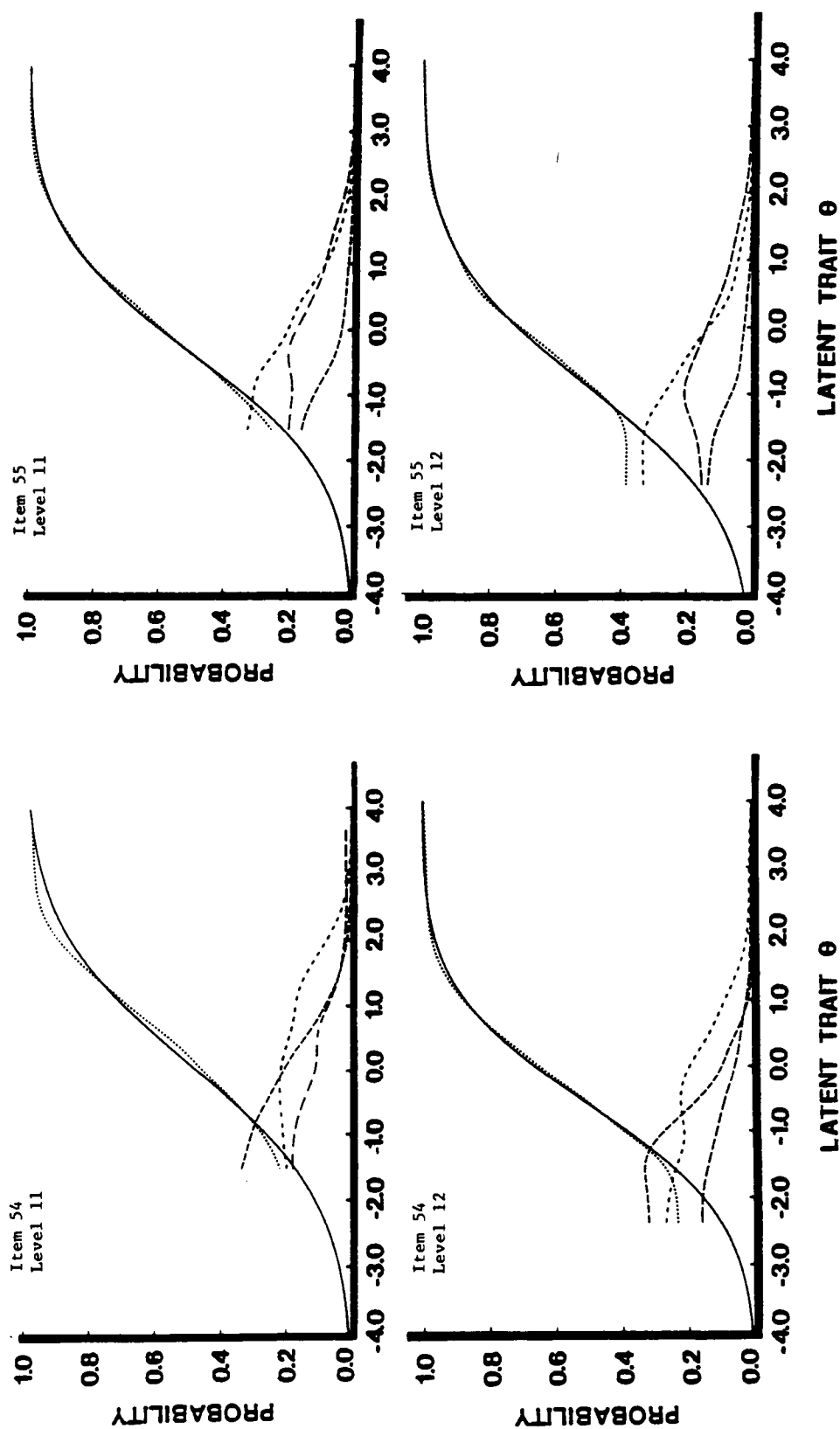


FIGURE III.2 (Continued)

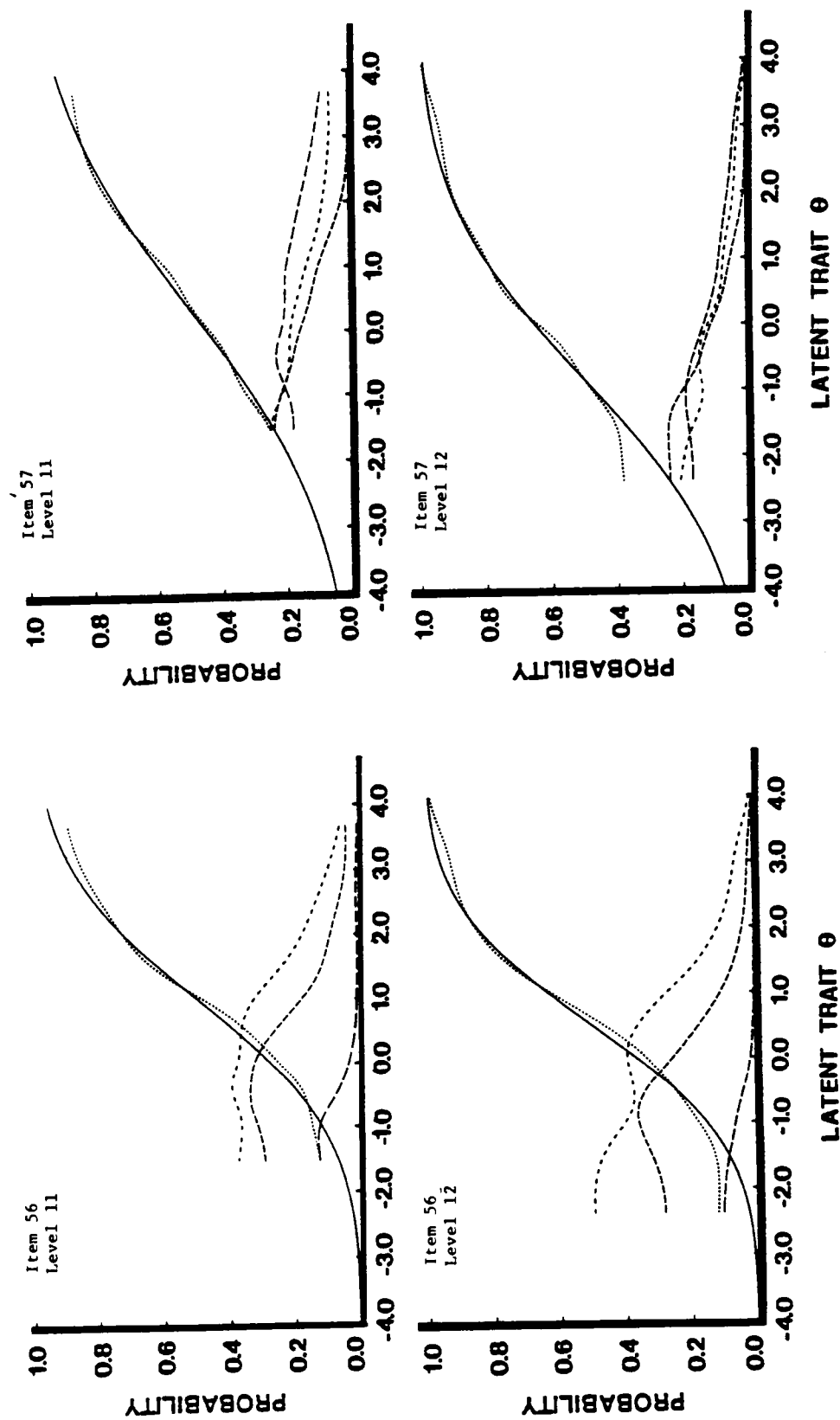


FIGURE III.2 (Continued)

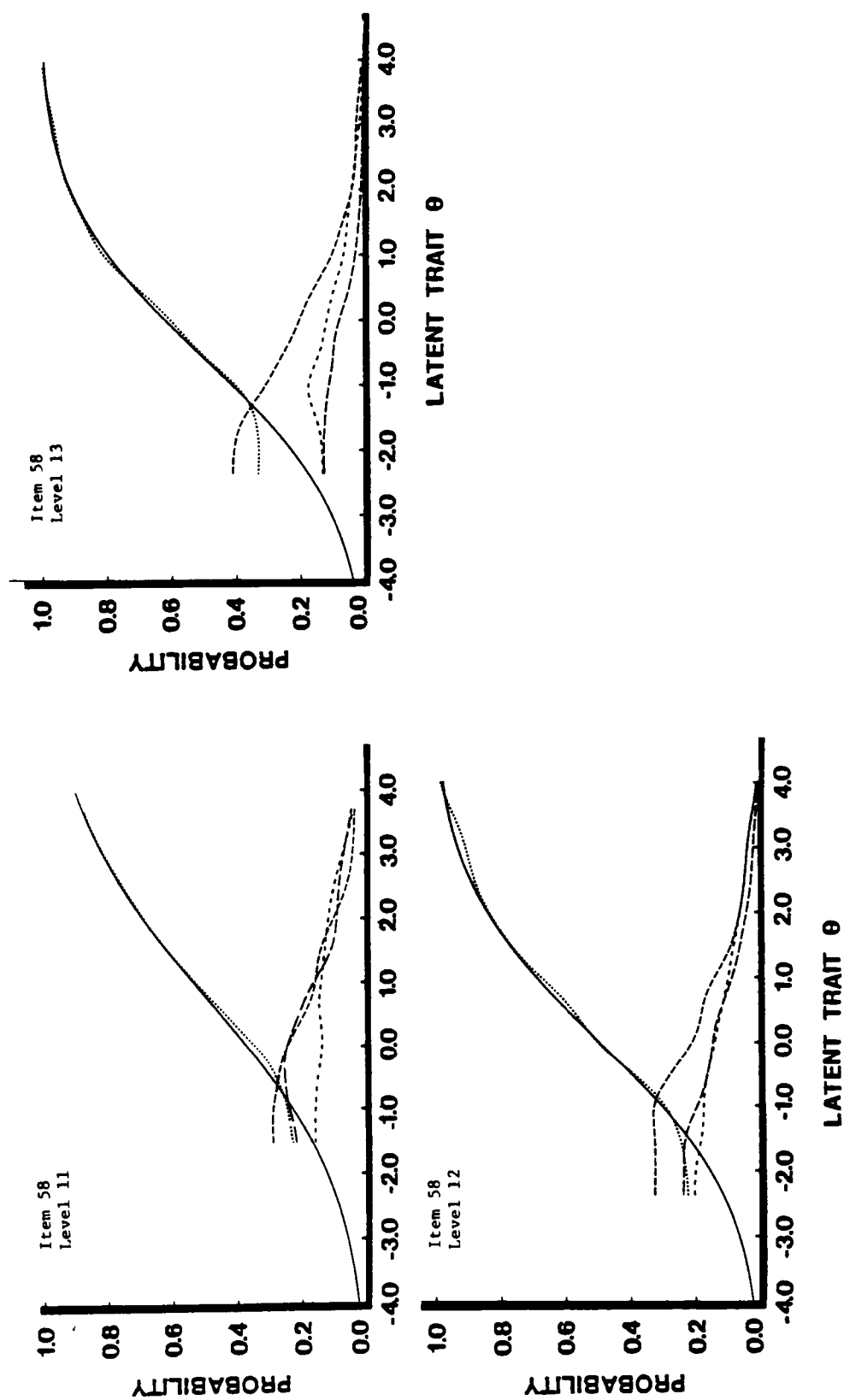


FIGURE III.2 (Continued)

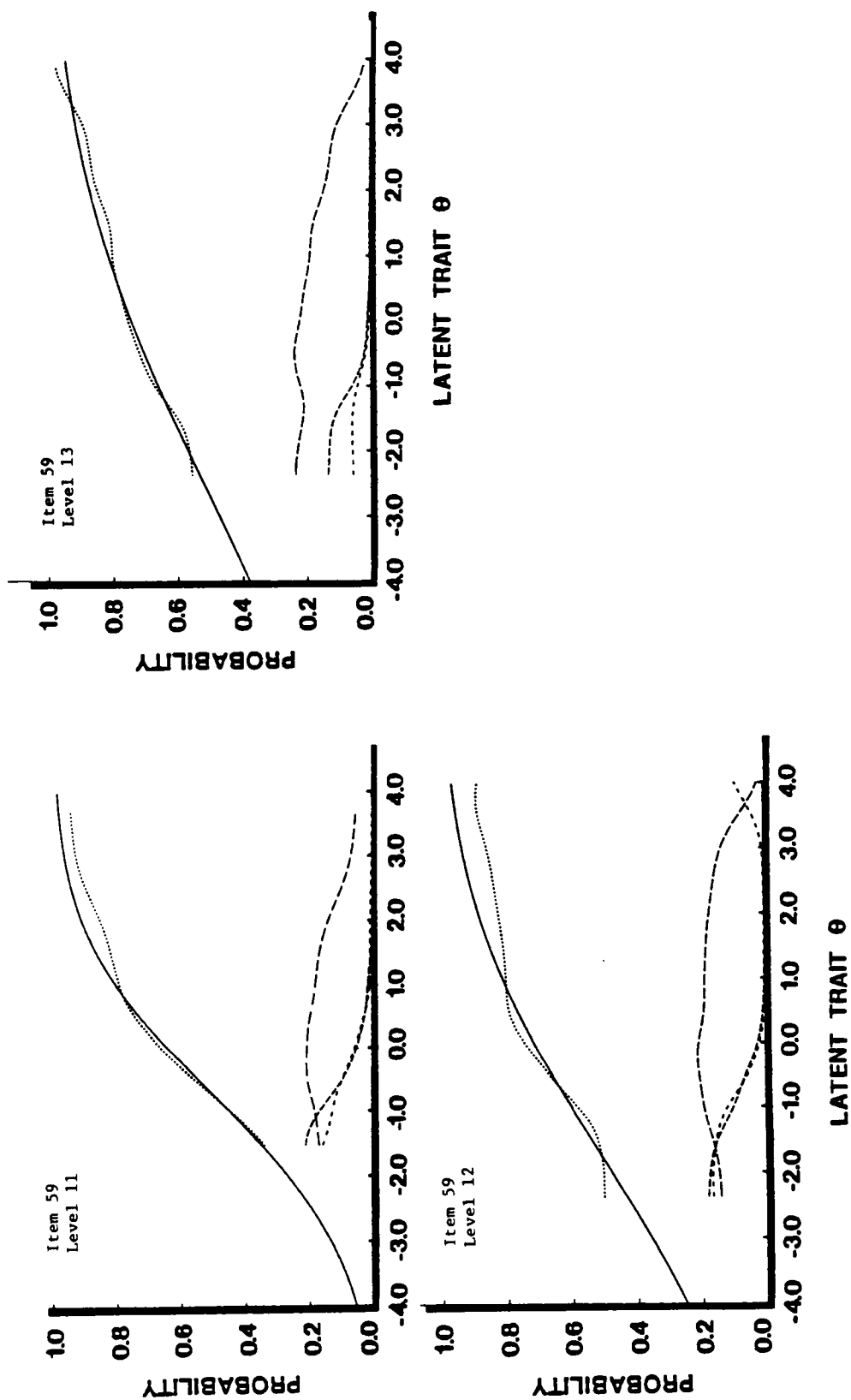


FIGURE III.2 (Continued)

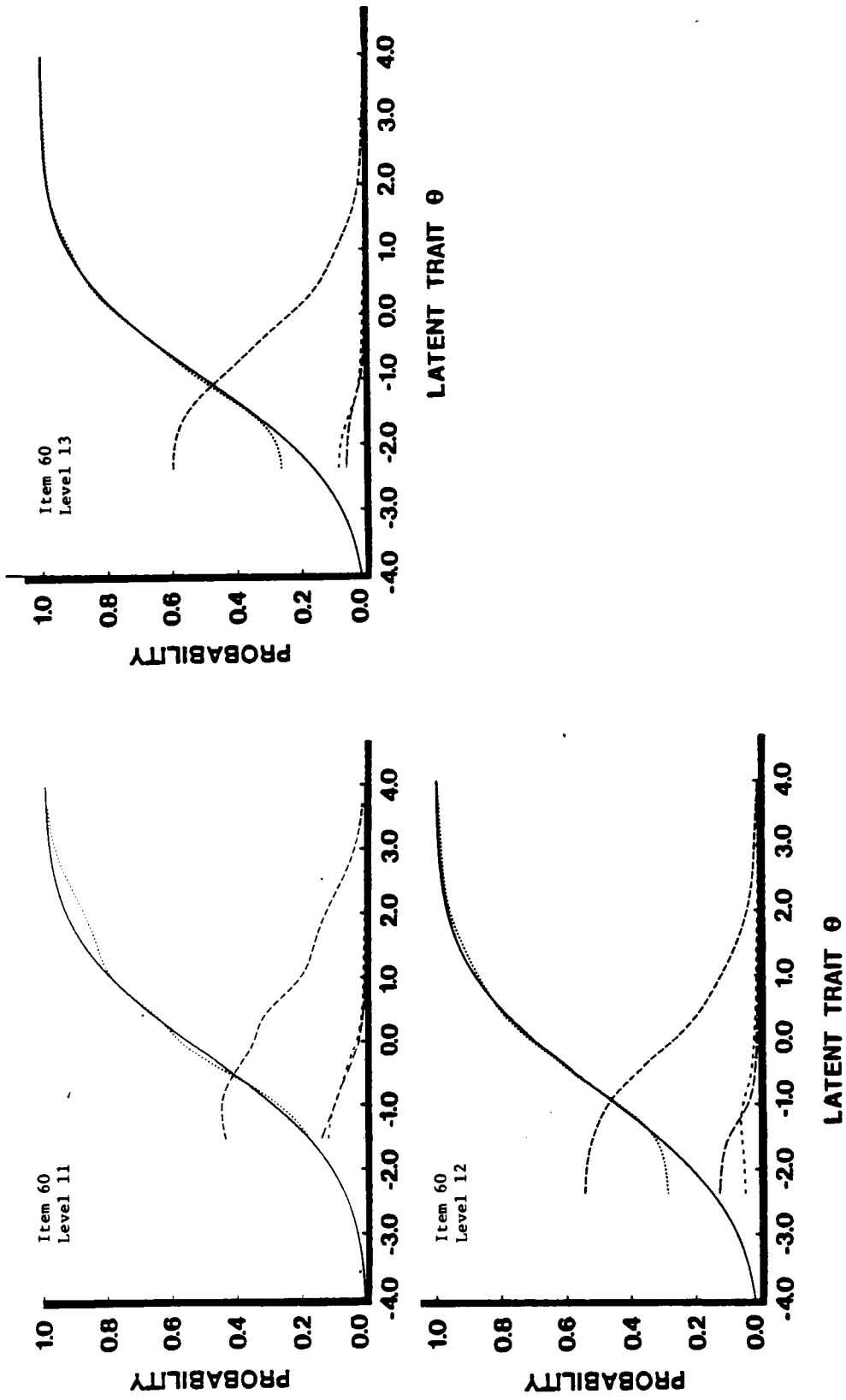


FIGURE III.2 (Continued)

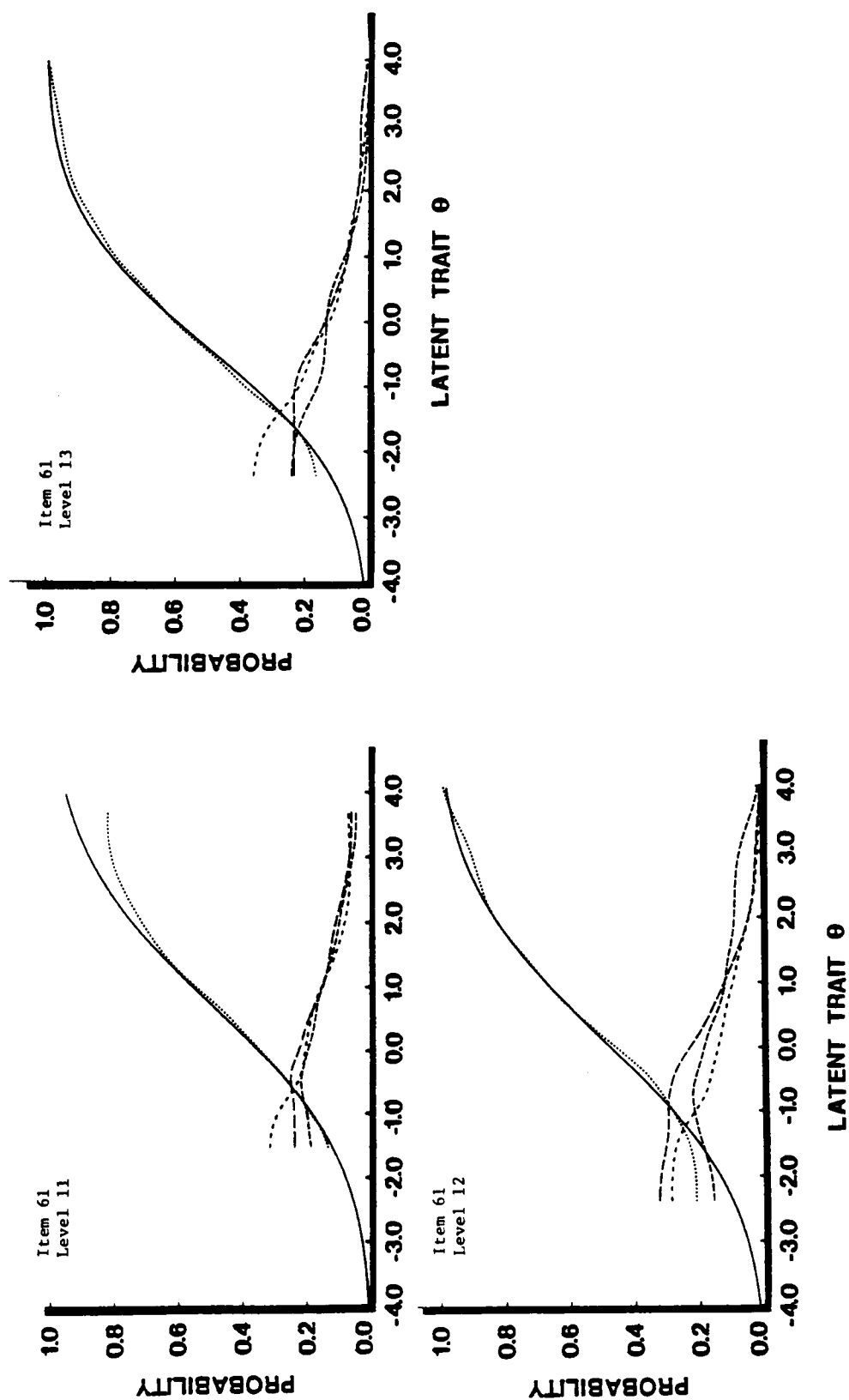


FIGURE III.2 (Continued)

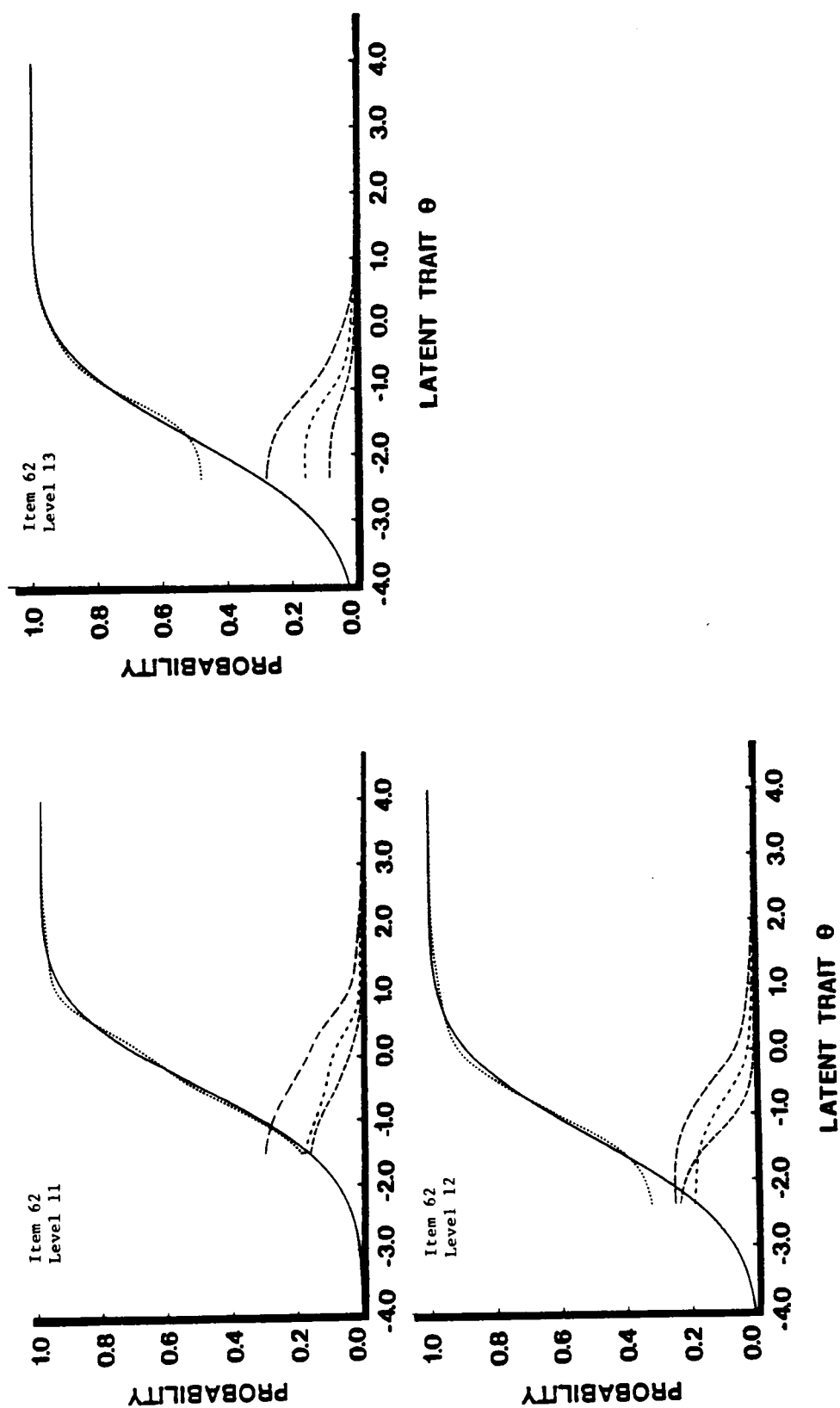


FIGURE III.2 (Continued)

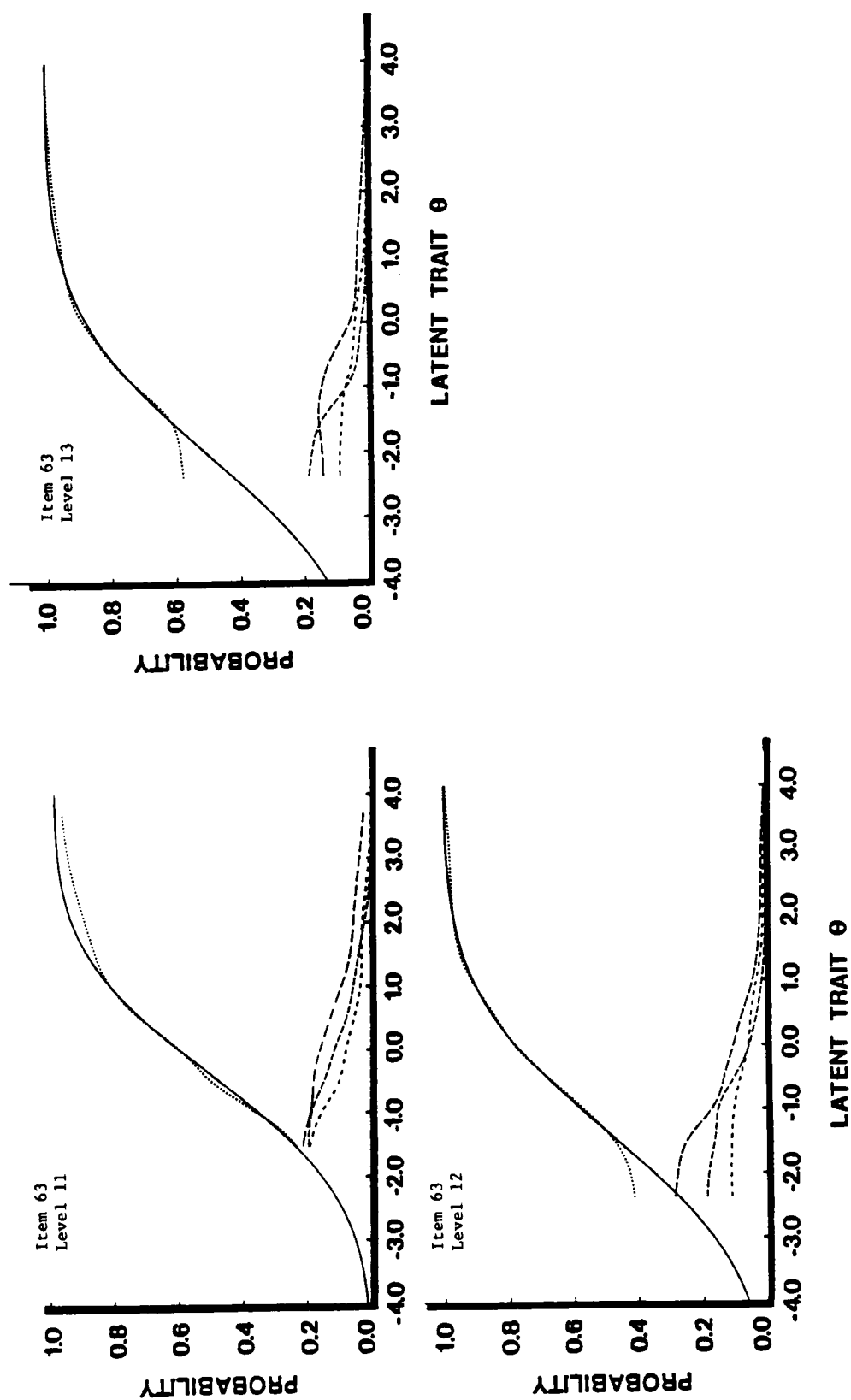


FIGURE III.2 (Continued)

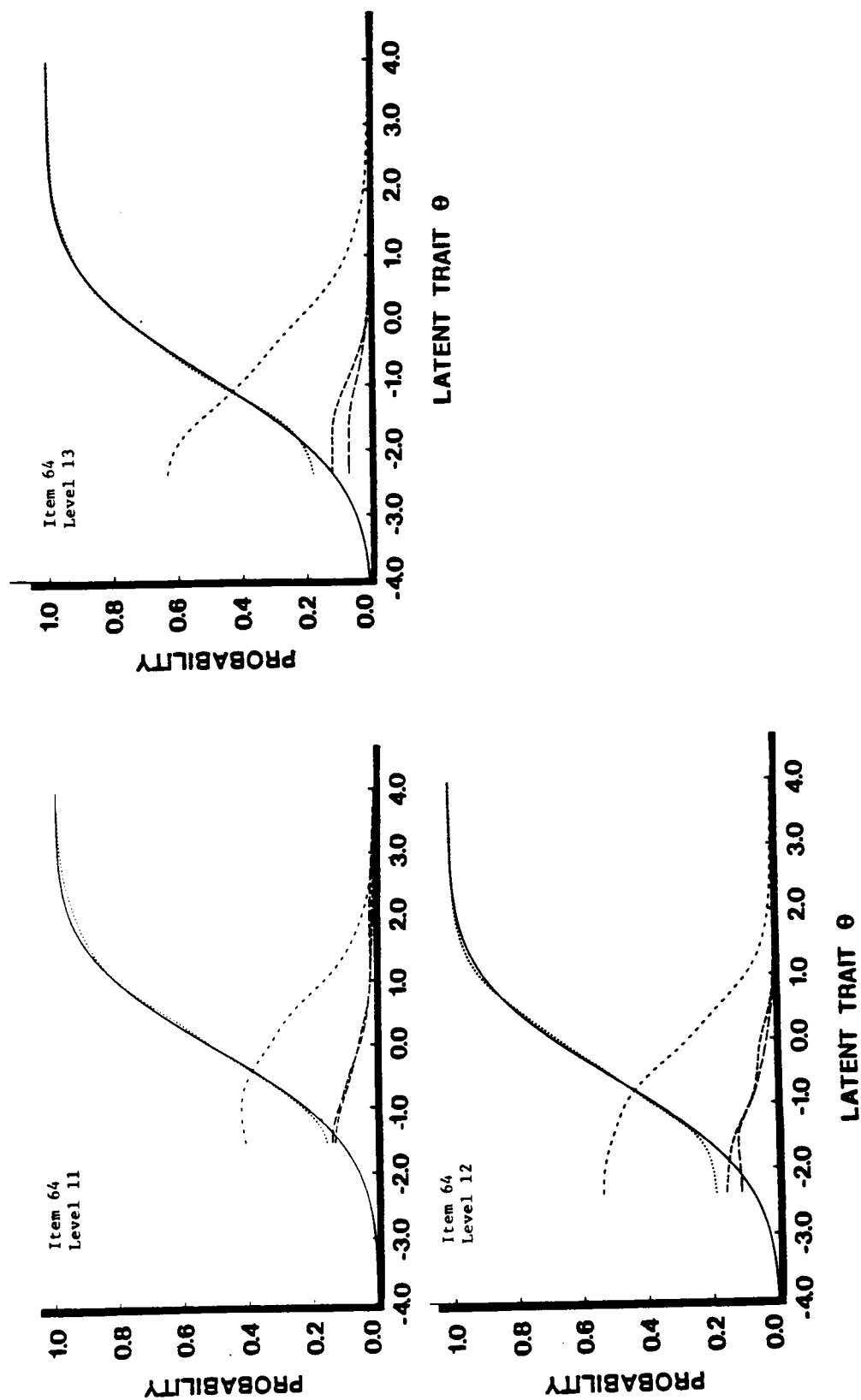


FIGURE III.2 (Continued)

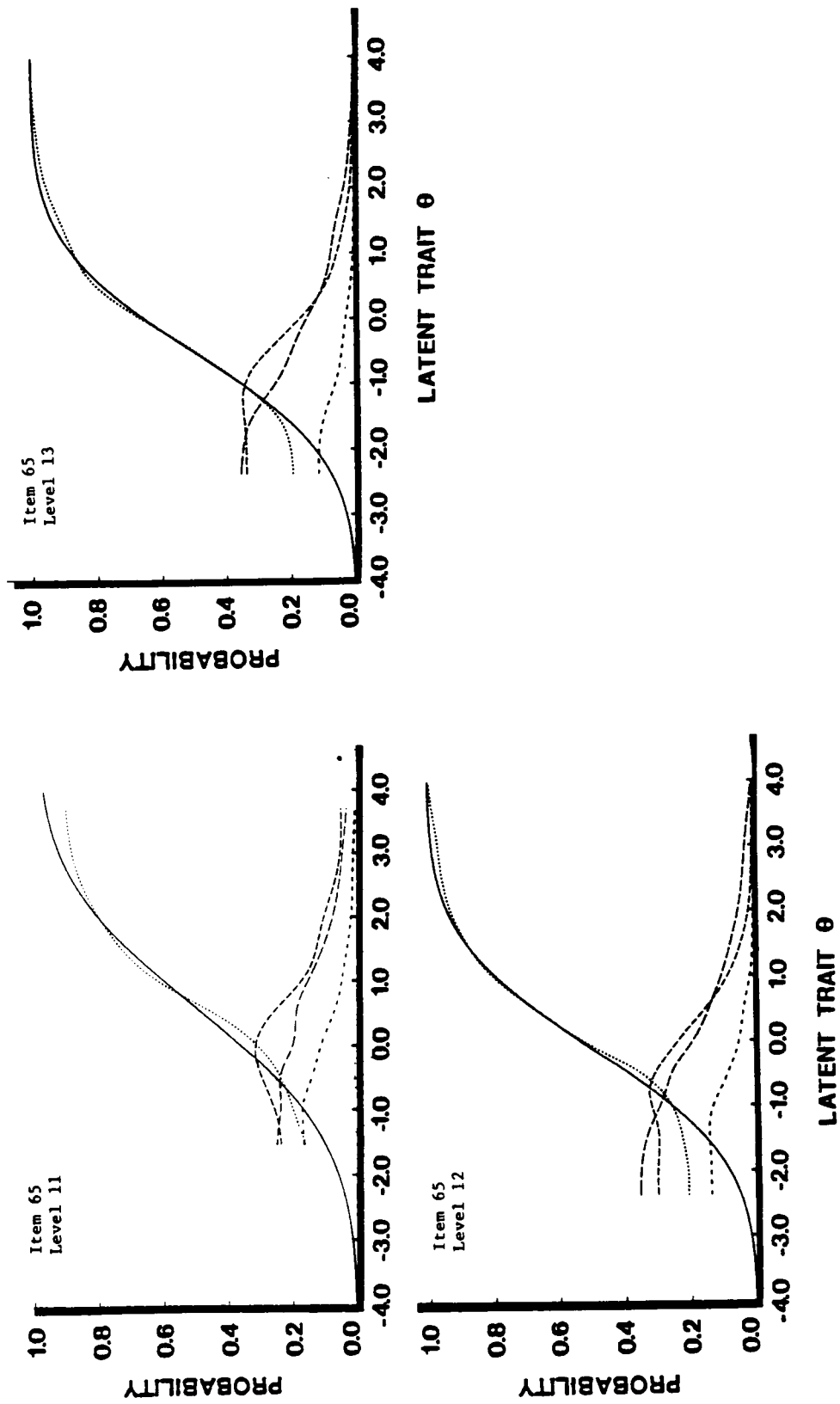


FIGURE III.2 (Continued)

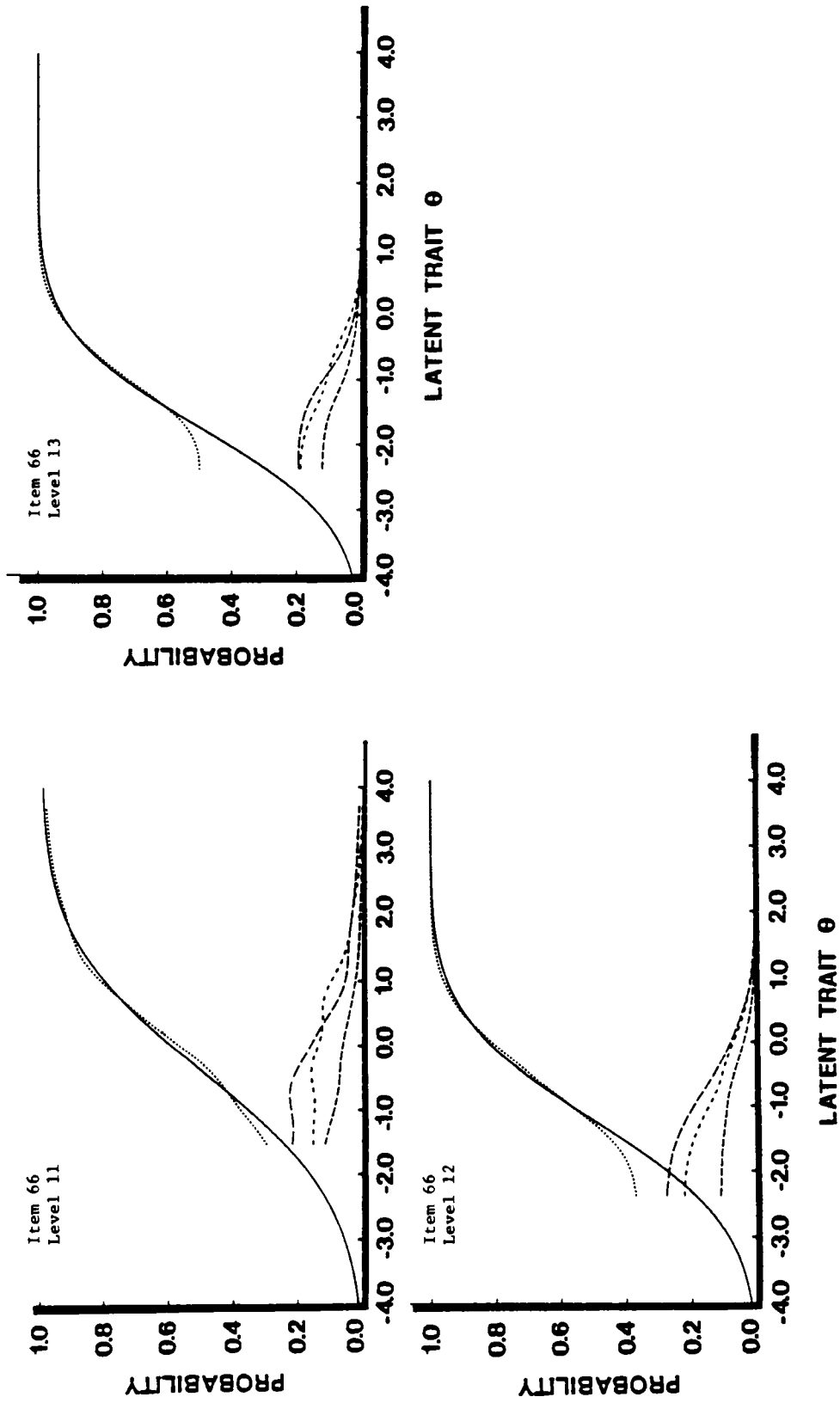


FIGURE III.2 (Continued)

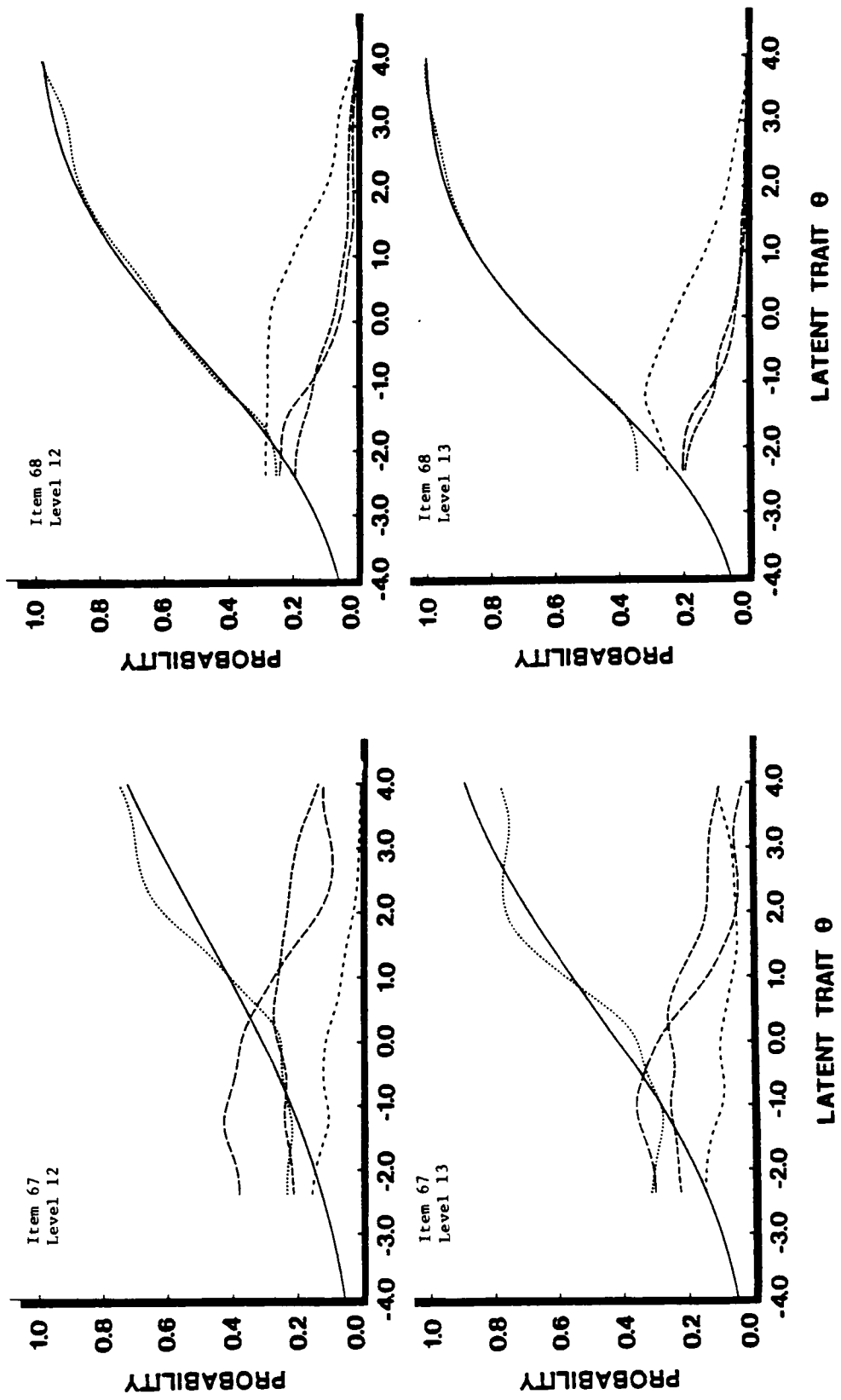


FIGURE III.2 (Continued)

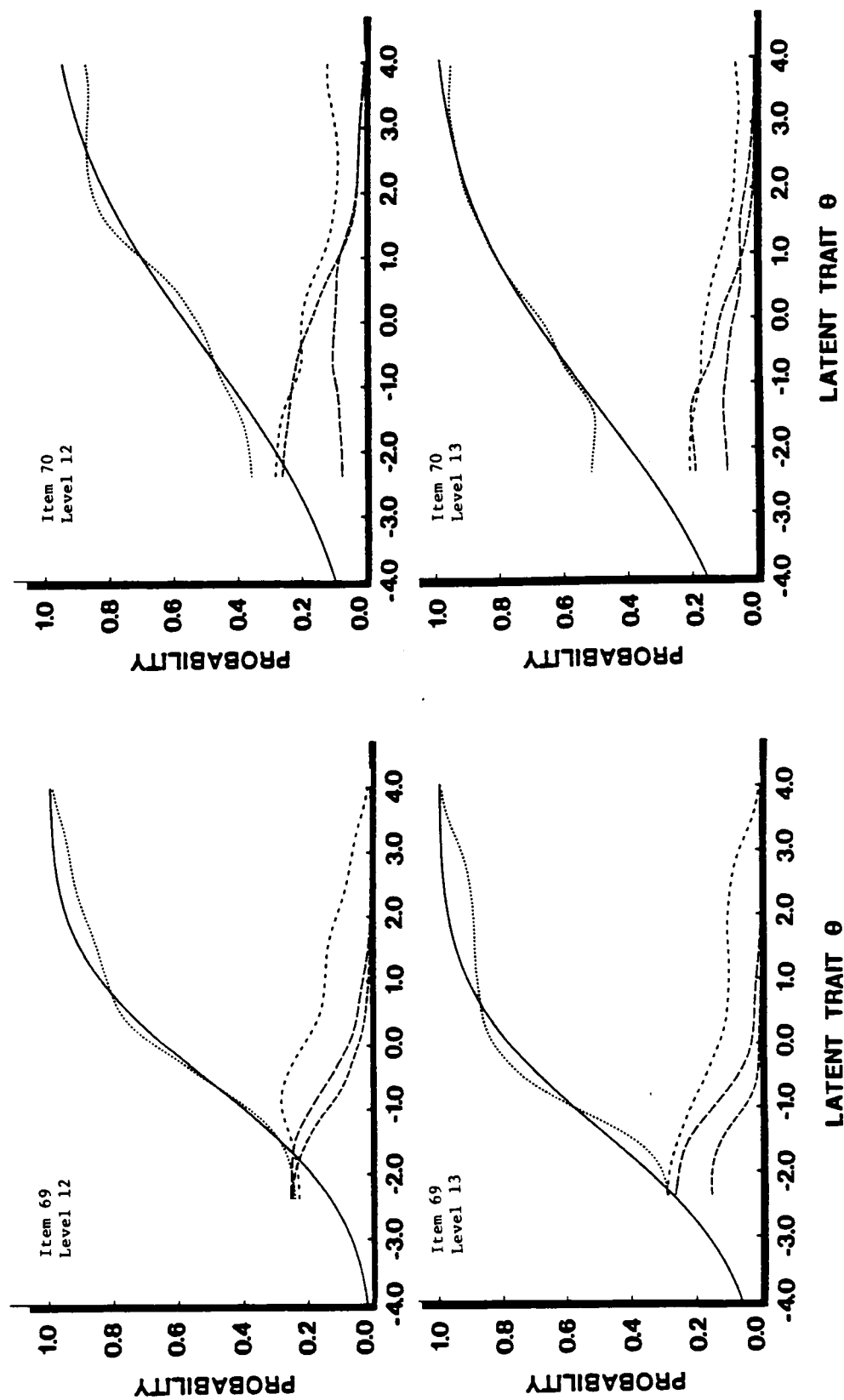


FIGURE III.2 (Continued)

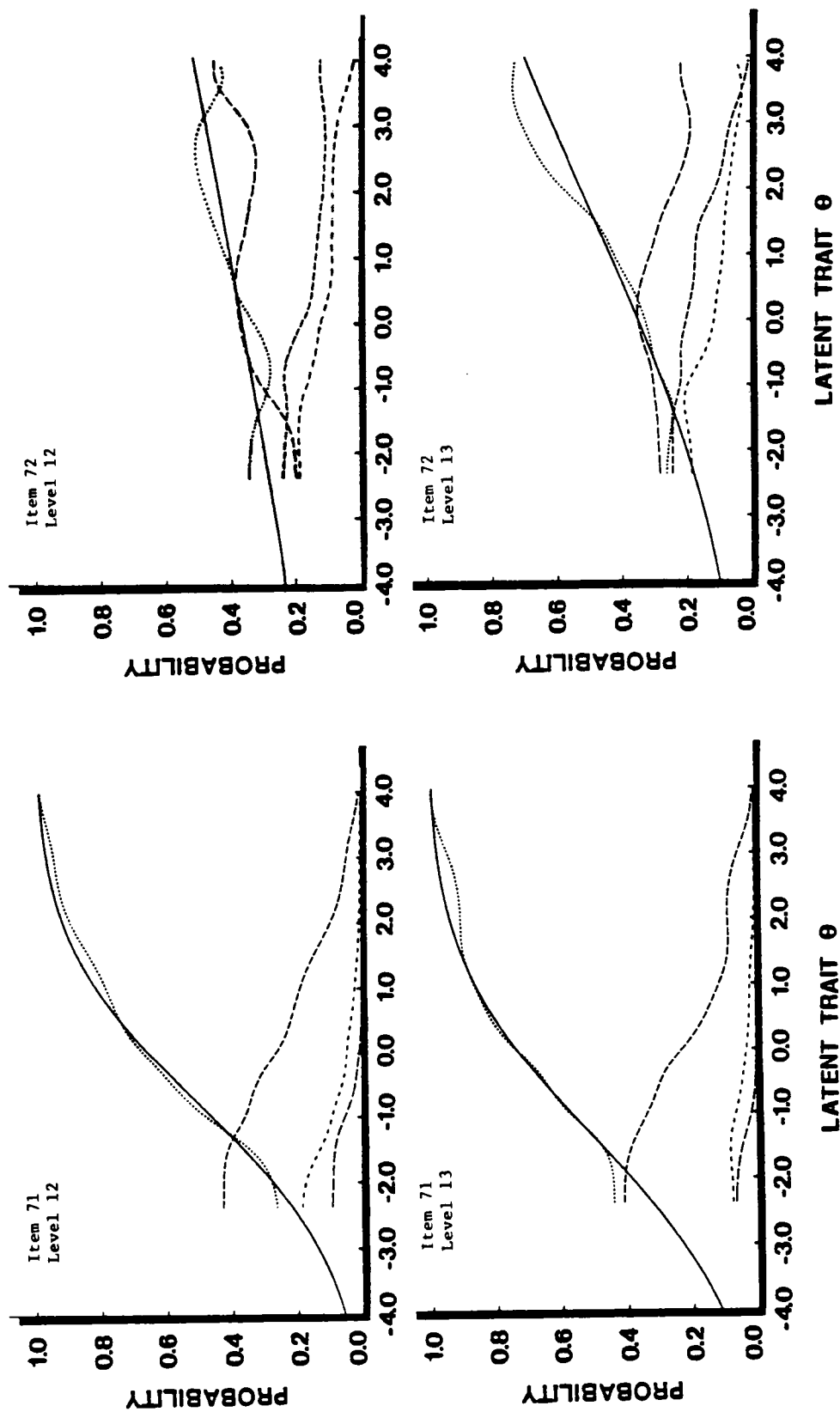


FIGURE III.2 (Continued)

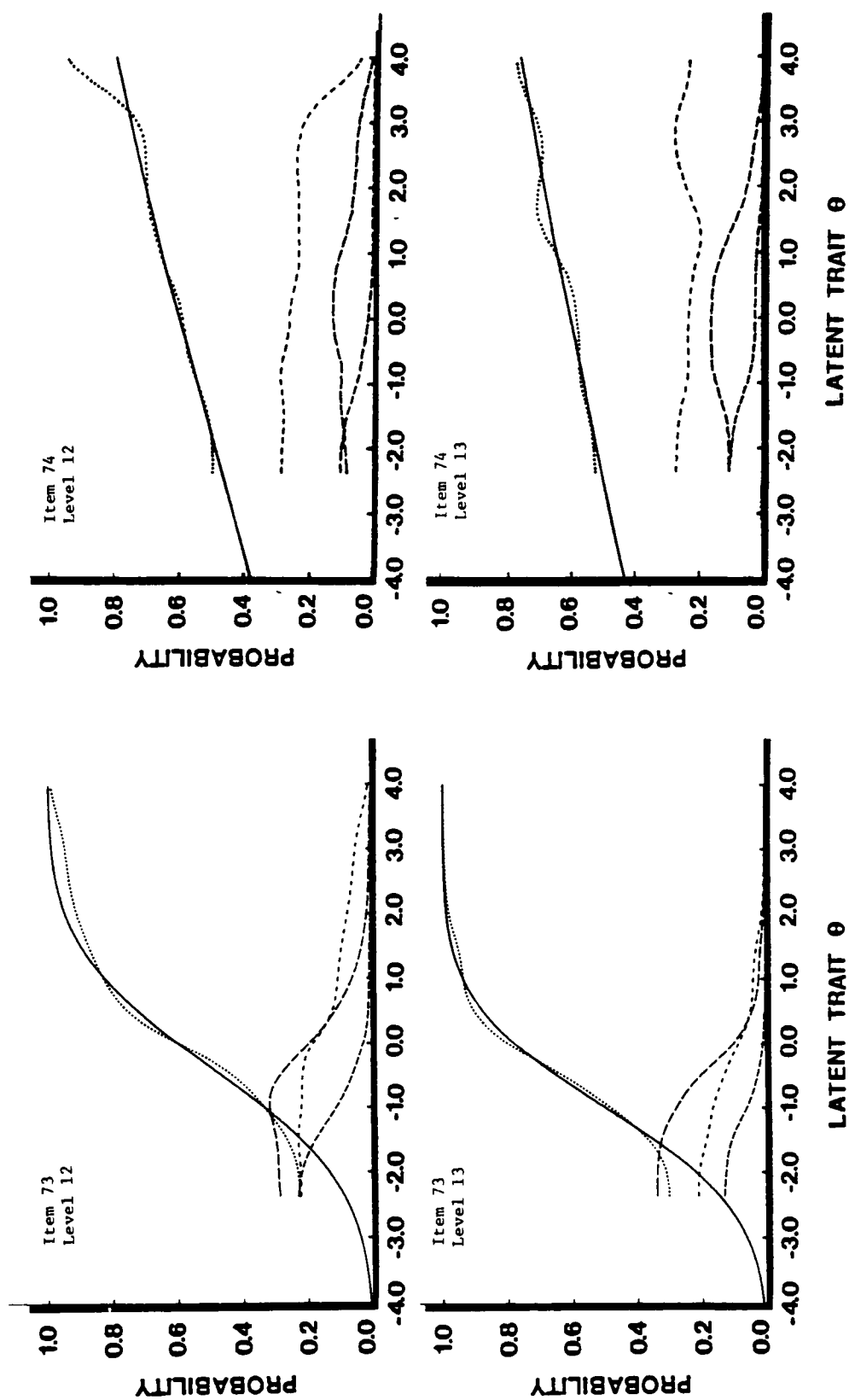


FIGURE III.2 (Continued)

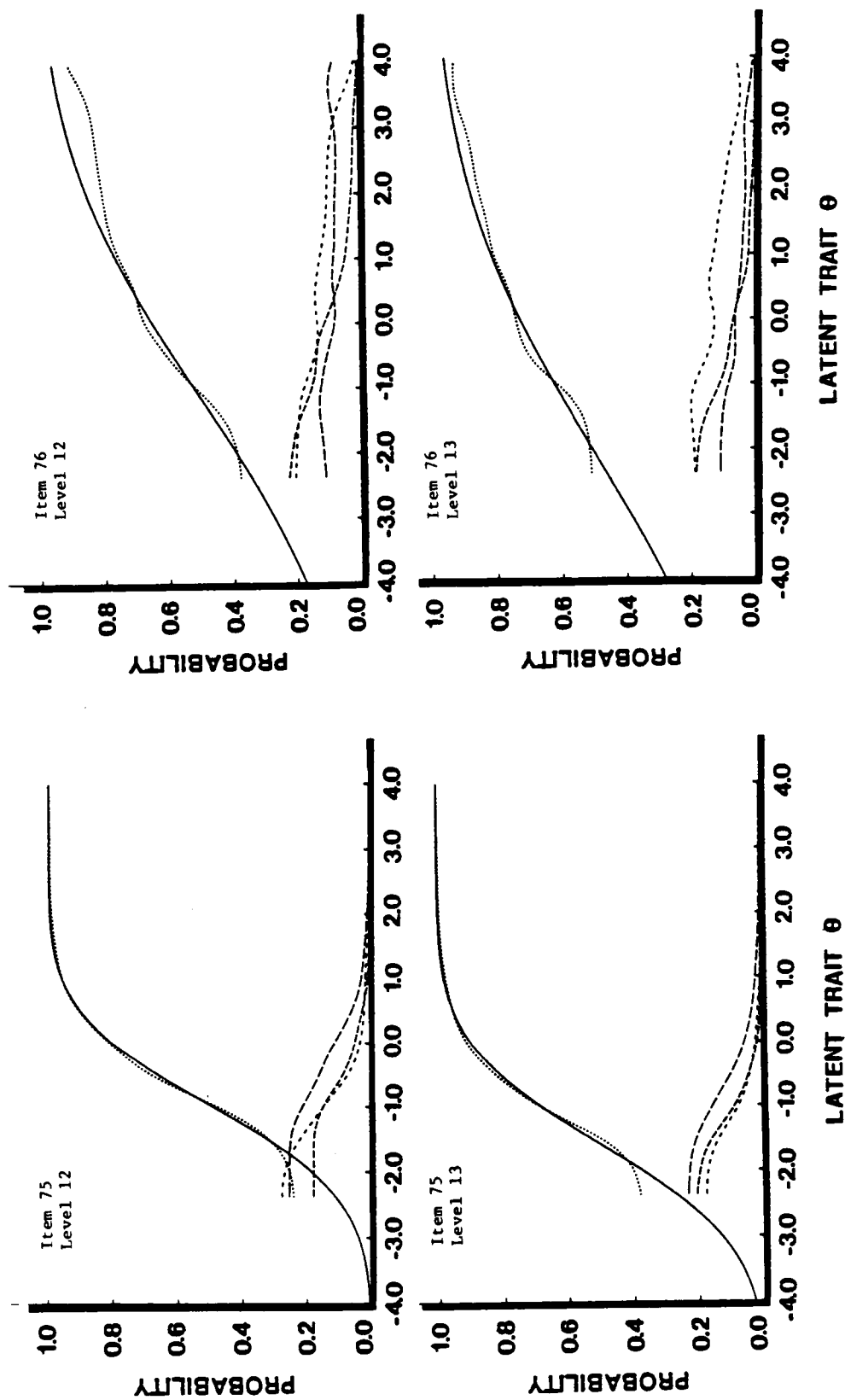


FIGURE III.2 (Continued)

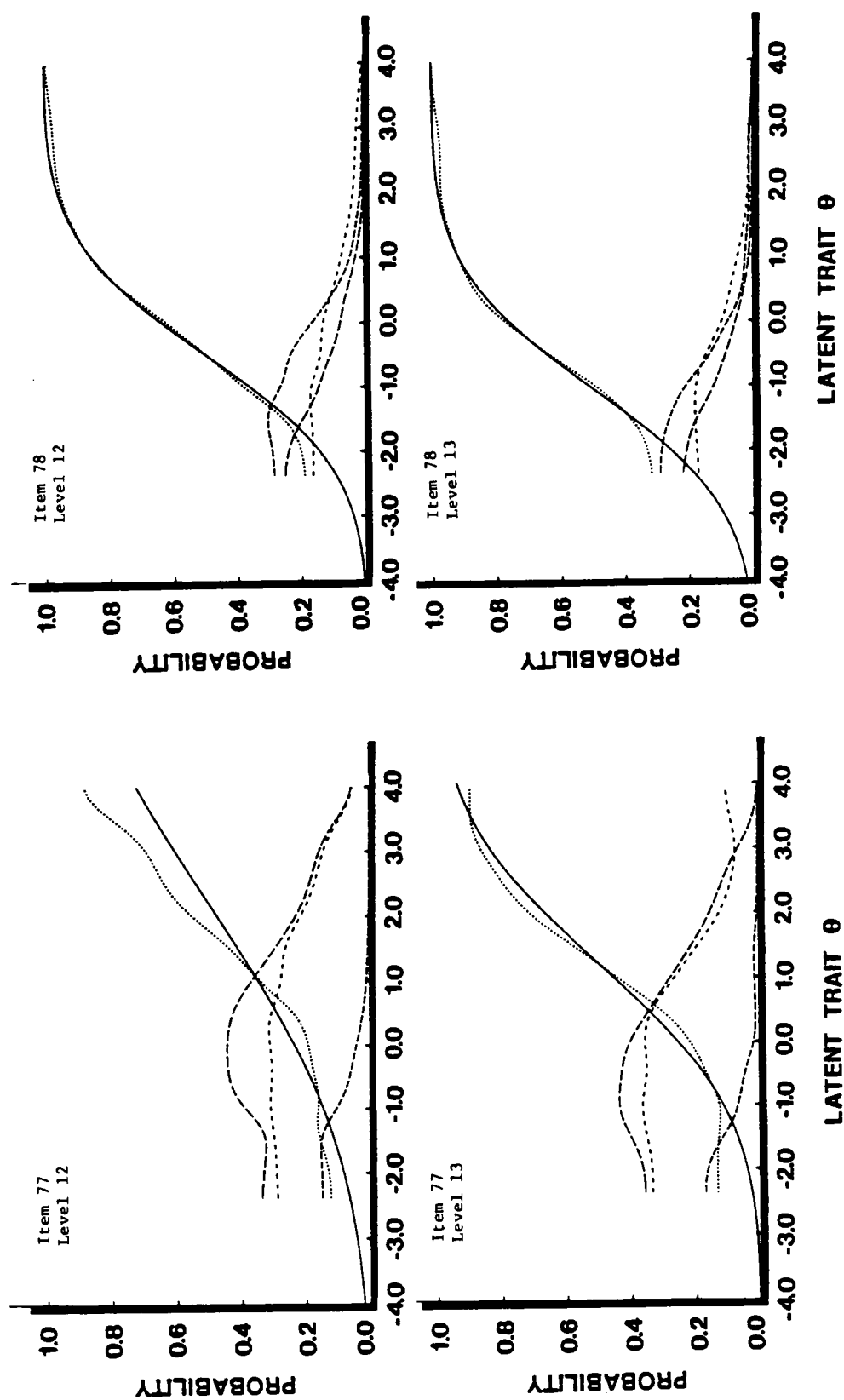


FIGURE III.2 (Continued)

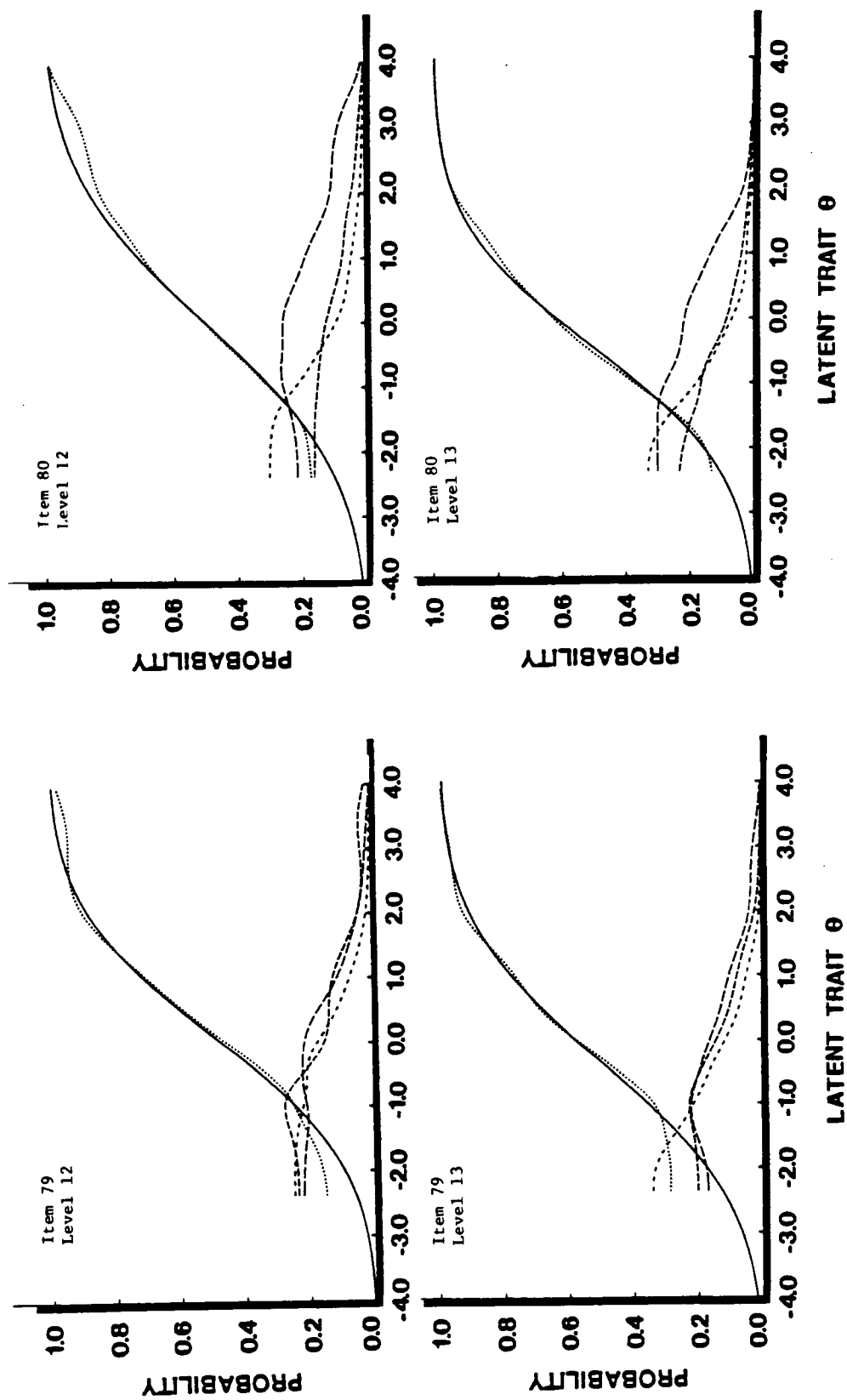


FIGURE III.2 (Continued)

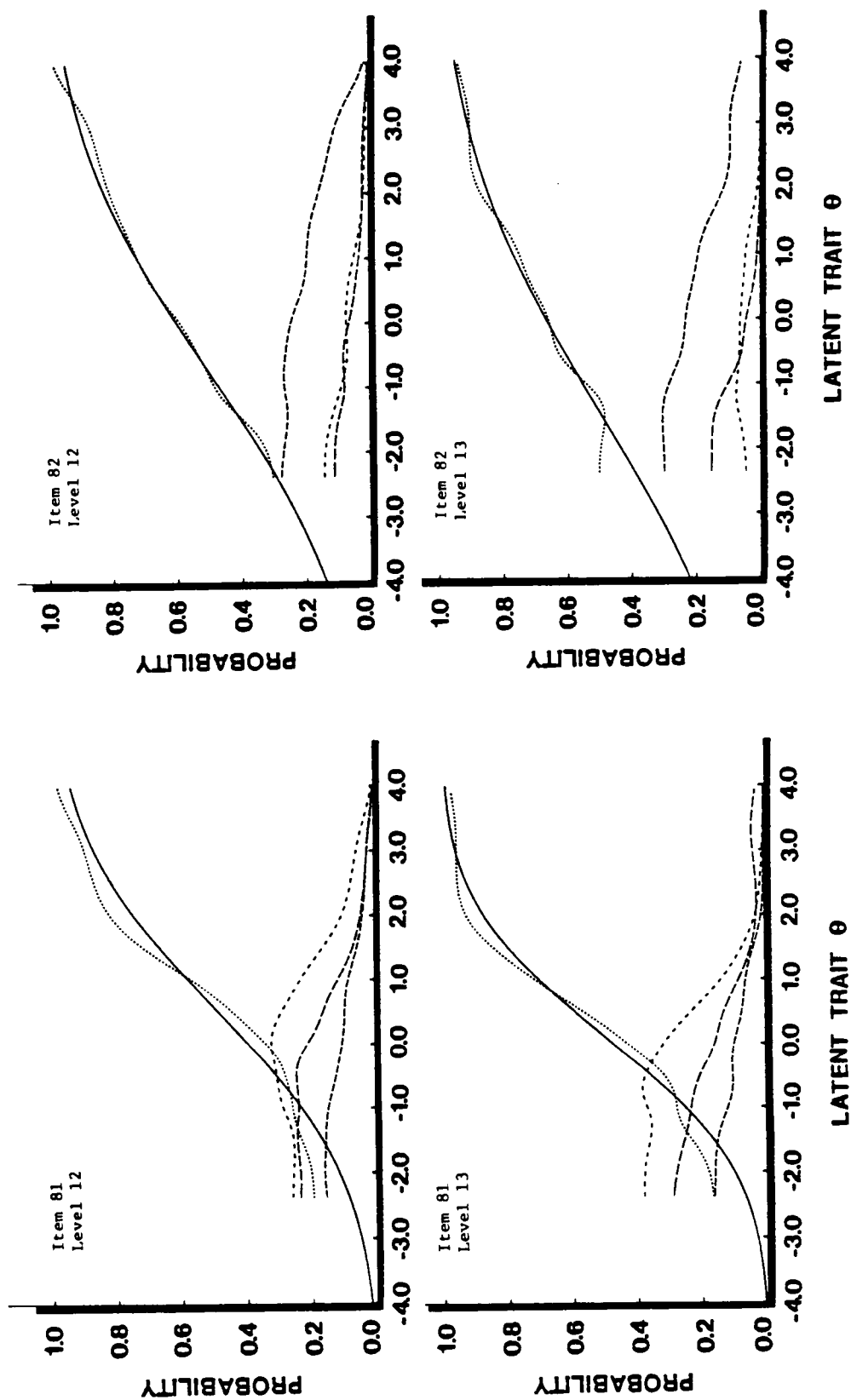


FIGURE III.2 (Continued)

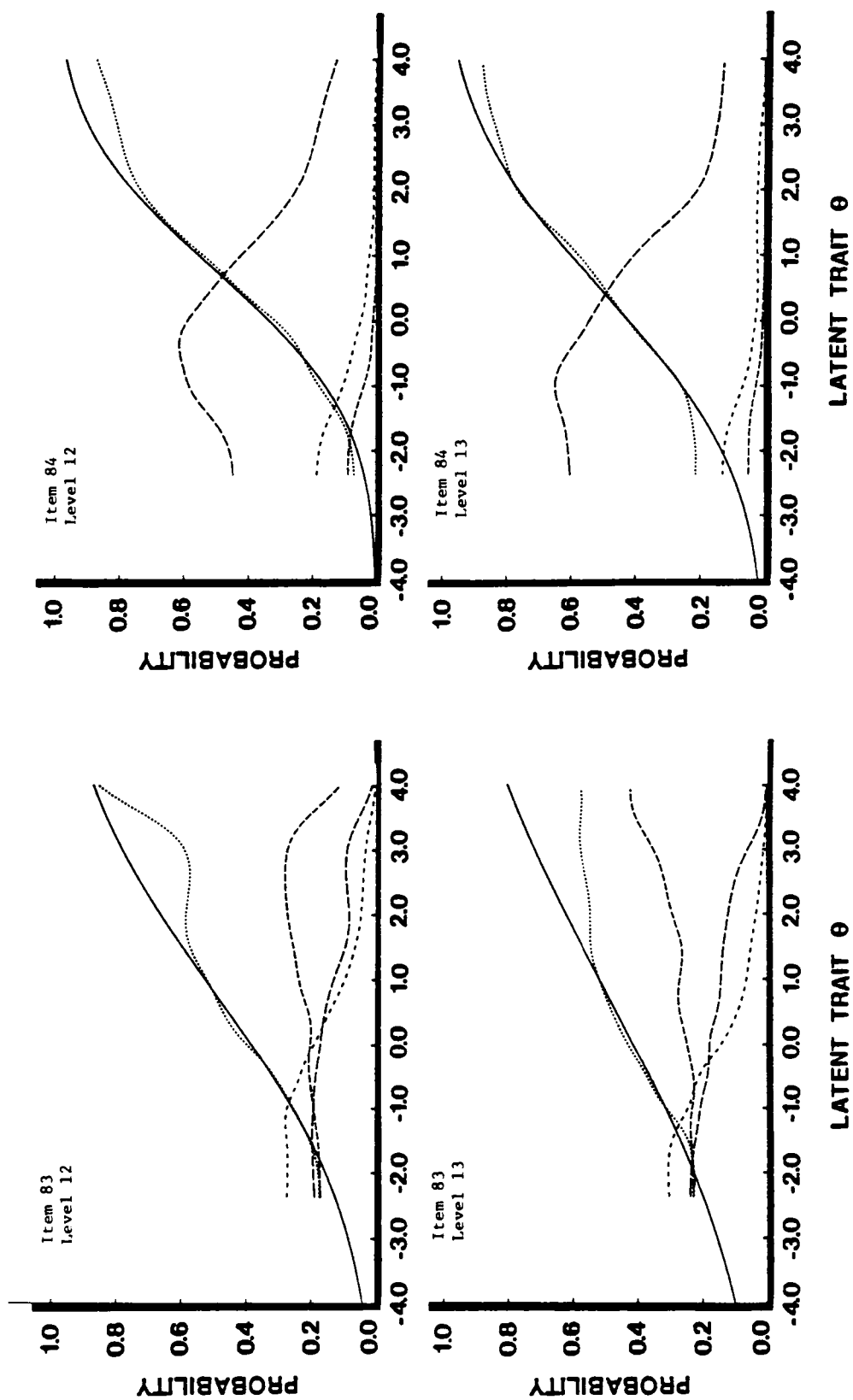


FIGURE III.2 (Continued)

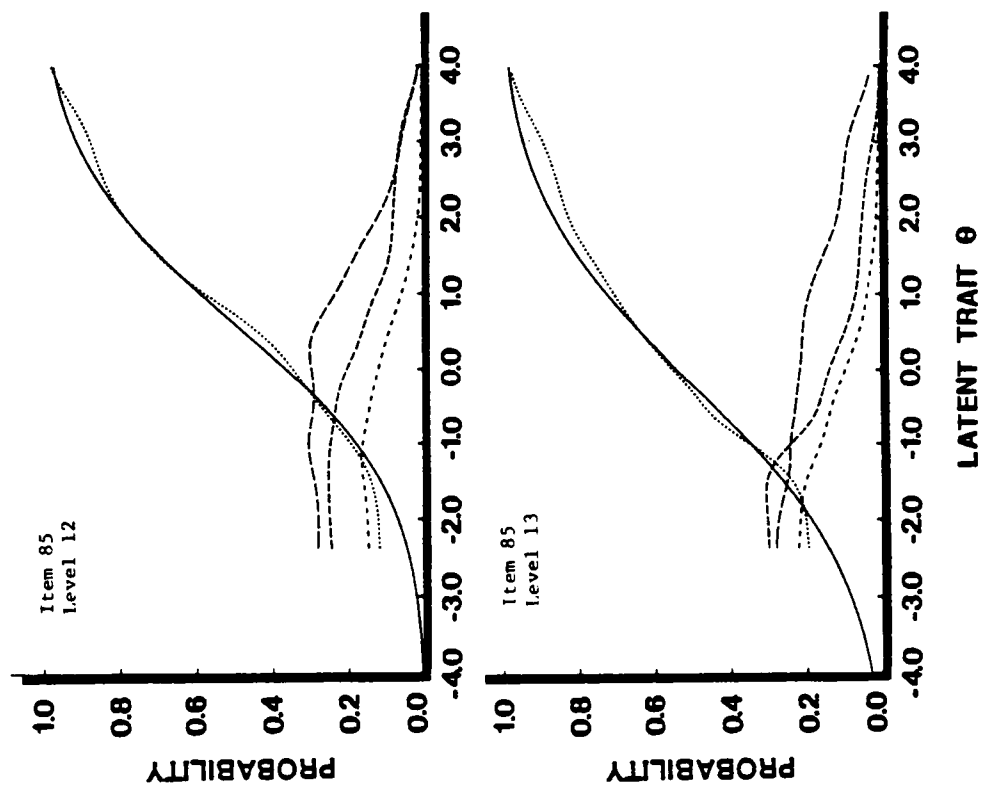


FIGURE III.2 (Continued)

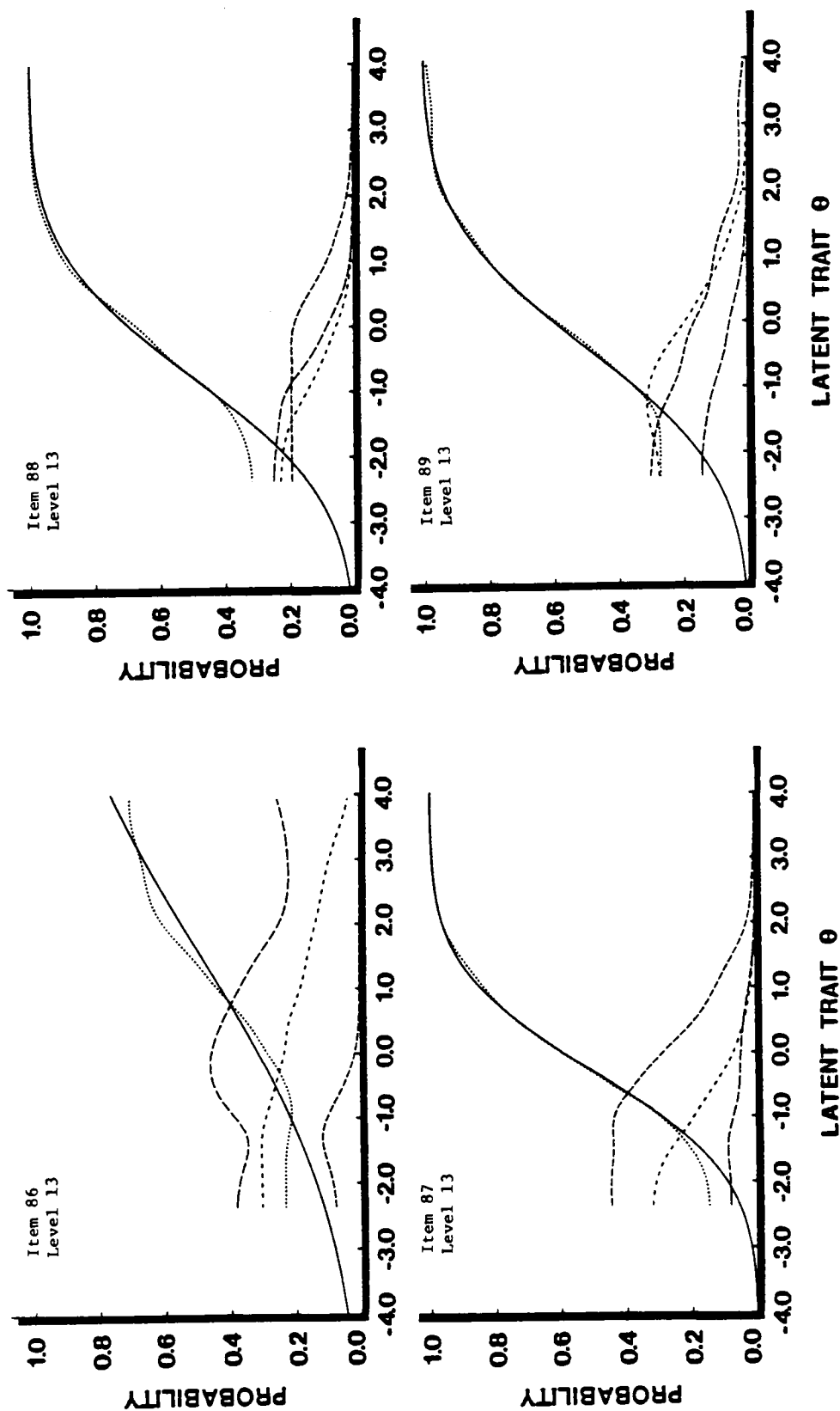


FIGURE III.2 (Continued)

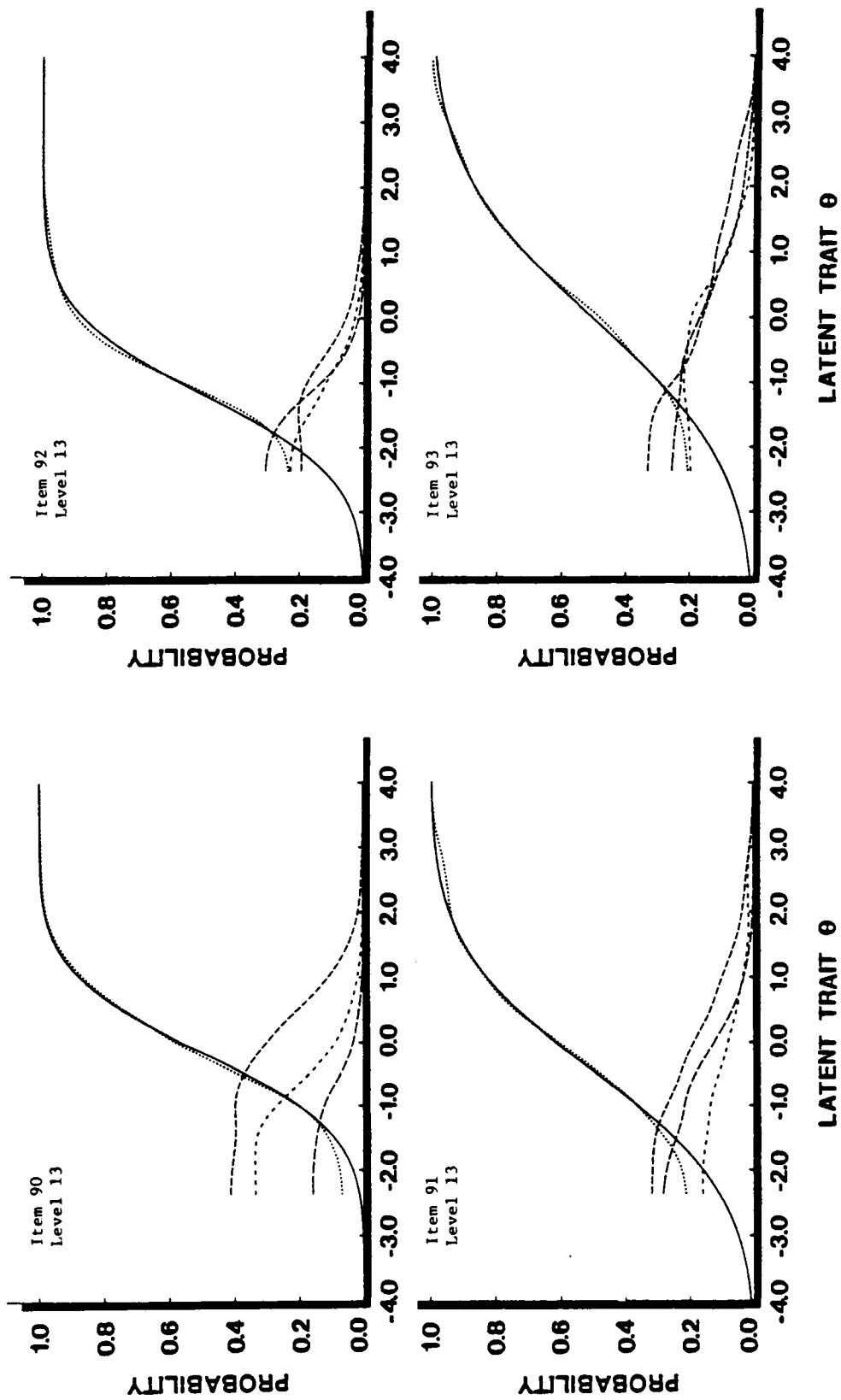


FIGURE III.2 (Continued)

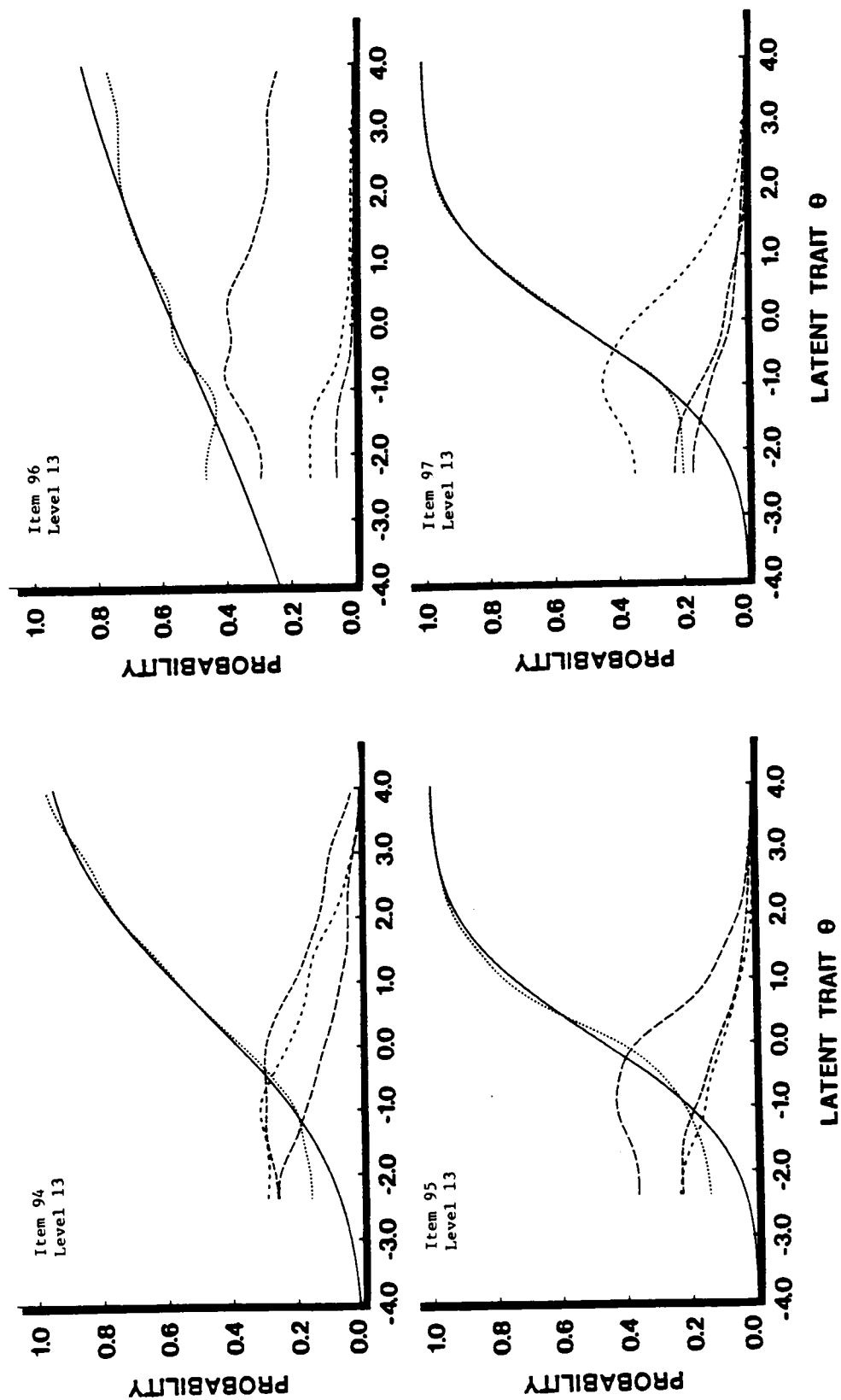


FIGURE III.2 (Continued)

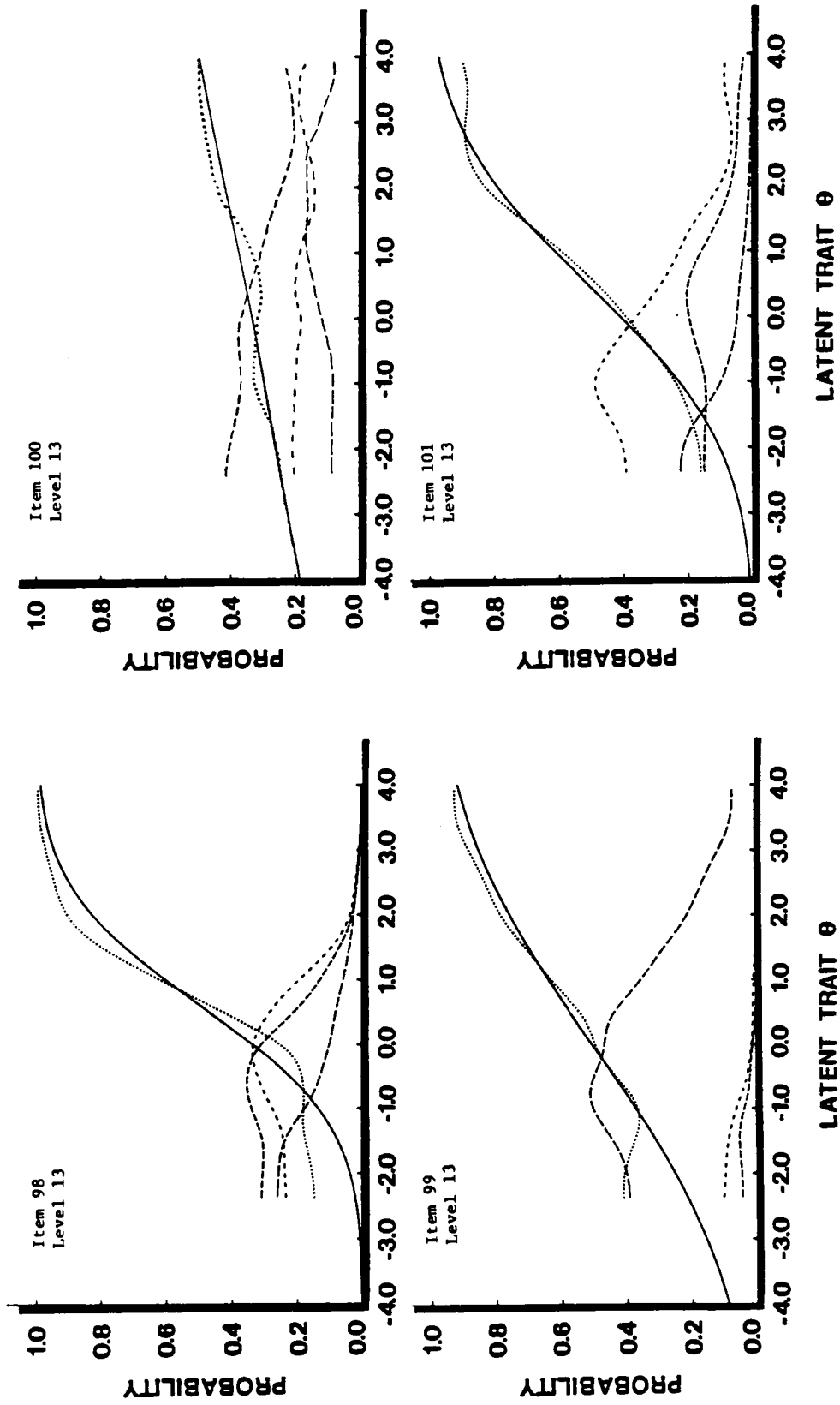


FIGURE III.2 (Continued)

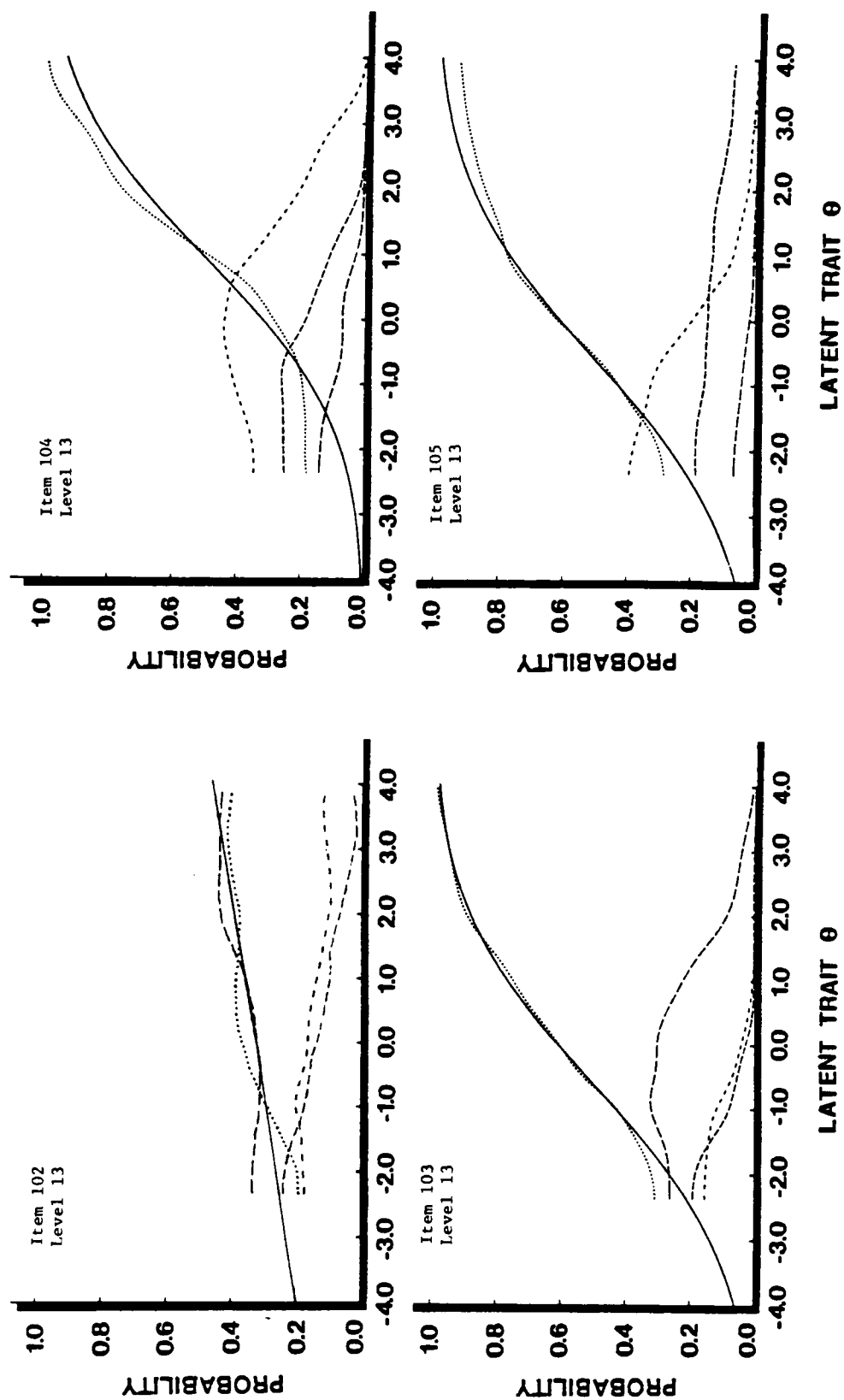


FIGURE III.2 (Continued)

normal density function, in order to provide better estimation corresponding to a meaningful interval of ability, θ . The estimation of these normal ogive item parameters was obtained through the use of Samejima's computer program FSPC-85-14.

For items that only appear in one level, such as items 24 through 39 (Level 11) and items 86 through 105 (Level 13), the graphs of the plausibility functions in Figure III.2 appear four items to a page, from top to bottom and left to right. For items that appear in two levels, such as items 40 through 57 (Levels 11 and 12) and items 67 through 85 (Levels 12 and 13), the plausibility functions for two items are presented on one page, with the results from the earlier level above the latter. For items 58 through 66, that appear in all three levels, one item is presented per page, with the three graphs obtained for these three levels shown together.

For most of the items that occur in two or more levels, the results with regard to informative distractors are consistent. Because the latent trait scale for each level is different, adjusted to each group of examinees, so that each ability distribution is standardized with respect to that particular scale, we can not combine the results from two or more levels directly. However, we can look at the shapes of the various plausibility functions and look for common trends between items. For example, for item 60 (p. 92) we see that distractor C, the dashed curve, is informative for each level, chosen by relatively low ability examinees at well above the chance, or 0.25, level in each case. Alternatives B and D are chosen much less frequently over the range of ability, and are generally quite close to

each other in the three graphs. On the other hand, there are a few items for which the results of the different levels are not as consistent. For instance, for item 85 (p. 108), the results for Level 12 suggest that distractor D, given by the long dashed curve is the most informative, while for Level 13 the curve for this distractor is flatter and gives less evidence of being informative than does distractor C. Although there is not a great difference between the curves for Levels 12 and 13, the Level 12 curve for D does show a modest increase in probability over some range of ability--the corresponding Level 13 curve does not.

Table III.6 presents the distractor or distractors for each of the 82 Vocabulary items that the author judged to be informative based upon the graphs of the plausibility functions. Informative distractors were generally identified by having plausibility functions that peaked clearly above the chance level of probability. Typically they increased with ability over a particular range of the ability scale and then decreased. In cases where there was contradictory evidence for different levels, a decision had to be made taking into account the fact that although some ability ranges between successive levels will overlap, the lower level is more likely to indicate informative distractors since it contains examinees of overall lower ability, or achievement (younger age group), and lower ability examinees are generally more likely to be attracted to distractors. As can be seen from Table III.6, 61 of the 82 items were observed to have informative distractors, and 9 of these had two informative distractors. When two informative distractors are observed for one

TABLE III.6 Observed Informative Distractors for 82 Items
of Levels 11, 12 and 13 of Vocabulary Subtest.

	Item	Observed
1	24	---
2	25	---
3	26	C
4	27	C
5	28	B
6	29	---
7	30	B
8	31	C
9	32	---
10	33	---
11	34	D
12	35	C
13	36	D
14	37	C-D
15	38	B
16	39	D
17	40	C
18	41	C
19	42	C
20	43	---
21	44	D
22	45	D
23	46	C
24	47	D
25	48	C
26	49	C
27	50	B-C
28	51	B
29	52	B
30	53	D
31	54	C
32	55	B
33	56	B-C
34	57	---
35	58	C
36	59	---
37	60	C
38	61	B
39	62	D
40	63	---

TABLE III.6 (Continued)

	Item	Predicted
41	64	B
42	65	C-D
43	66	---
44	67	D
45	68	B
46	69	---
47	70	---
48	71	C
49	72	D
50	73	D
51	74	---
52	75	---
53	76	---
54	77	D
55	78	---
56	79	B
57	80	D-B
58	81	B
59	82	---
60	83	C
61	84	D
62	85	D-C
63	86	D
64	87	C-B
65	88	---
66	89	B
67	90	C-B
68	91	---
69	92	D
70	93	C
71	94	---
72	95	D
73	96	C
74	97	B
75	98	B-C
76	99	D
77	100	---
78	101	B
79	102	D
80	103	D
81	104	B
82	105	B

item, the distractor generally appealing to the higher ability examinees is given first.

Of the 37 items of the Vocabulary subtest that were administered to exactly two levels, items 40 through 57 (Levels 11 and 12) and items 67 through 85 (Levels 12 and 13), 28 items were judged to have at least one informative distractor. As was stated earlier in this section, for most of these items the results obtained from the different levels are consistent. Of the 28 items administered in two levels, 23 items gave the same conclusion with regard to the identification of informative distractors when the plausibility functions obtained from each level were analyzed separately. Item 46 (p. 84) provides a good example of such an item, as the results obtained separately for each of Levels 11 and 12 clearly indicate that alternative C is the only informative distractor. For three of the items, items 42, 47 and 79, an informative distractor was indicated from the results for one of the two levels, but not for the other level. For item 42 (p. 82), for example, alternative C is identified as an informative distractor based upon the plausibility functions for Level 11, as the probability of it being selected is close to 0.40 for lower ability examinees (latent trait of approximately -1.5). Although the peak of the plausibility function for distractor C obtained for Level 12 is above 0.25 , the expected chance or guessing level, it only reaches about 0.30 , and is only slightly higher than the peak for distractor B. The plausibility function for distractor C is also rather flat for Level 12 and generally gives weak support at best for C as an informative distractor. The identification of

distractor C as being informative comes primarily from the results of Level 11. Actually the results obtained for item 42 for the two levels are not that different, but Level 12 apparently does not have examinees of low enough ability needed to detect that distractor C is informative. It is only at the lower end of the ability scale for Level 11 that distractor C clearly distinguishes itself from distractor B.

Four of the 28 items administered to two levels were identified as having two informative distractors. For two of these items, items 56 and 80, the results appear to identify the same two informative distractors, in the same order with regard to the relative ability of examinees to which they are most plausible, although this conclusion is less clear for item 56 (p. 89) than for item 80 (p. 105). The estimated plausibility functions for item 56 are rather interesting, particularly for Level 12. The plausibility function for distractor B estimated for Level 12 appears to have two peaks, the higher of which corresponds to approximately -2.0 on the latent trait, and the lower of which corresponds to approximately 0.1 on the latent trait. Between these two points on the ability scale, the plausibility function of distractor B drops while the plausibility function of distractor C increases, with the latter reaching a peak corresponding to about -0.9 on the latent trait. It is evident from these results that B is the most appealing distractor both to relatively high ability levels (e.g., examinees above -0.5) and to rather low ability levels (e.g., examinees below -1.0), but that distractor C is essentially equally as appealing or plausible as B for moderately

low ability levels (e.g., examinees between -1.0 and -0.6 on the ability scale). Interestingly, the Level 11 results for item 56 do not show two peaks for the plausibility function of distractor B, although there is a slight indication at the lower end of the ability scale that the plausibility function is beginning to increase again as ability decreases, and that another peak would possibly have been found if the data had included enough low ability examinee abilities to allow estimation of the plausibility functions below -1.5 on the latent trait for Level 11. However, unlike Level 12, where the plausibility functions for distractors B and C nearly intersect at about -0.9 on the latent trait, the plausibility function for distractor B obtained for Level 11 is clearly above that of distractor C over the entire range of ability for which the functions were estimated.

The other two items identified as having multiple informative distractors that were administered to two different levels, items 50 and 85, provided different indications of informative distractors at each level. Item 50 (p. 86) was identified as having two informative distractors, B and C, based upon the analysis of the combined results for Levels 11 and 12, but Level 12 taken alone indicated only that distractor C was informative. For Level 11, however, the plausibility function of distractor B is generally as high or higher than that of distractor C over the range of ability, although the plausibility functions of both are relatively flat. For item 85 (p. 108), that was discussed earlier in this section, the same two distractors, D and C, were identified as being informative both for Level 12 and for Level

13, but the order of the two distractors with regard to the ability level for which they were most plausible was reversed. The peak of the plausibility function for distractor D occurs for higher ability examinees than that of distractor C for Level 12, while the peak for distractor C appears to occur at a relatively higher ability level than that of D for Level 13.

Of the nine items administered to all three levels, items 58 through 66, six items were identified as having informative distractors. For five of these six items, the informative distractors identified from the plausibility functions were the same for each of the three Levels, 11 through 13. This includes the only item administered to all three levels that had multiple informative distractors, item 65. For each of the three levels, distractor C was the informative distractor most plausible to relatively high ability examinees, and D was the informative distractor most plausible to lower ability examinees. Item 61 (p. 93), on the other hand, gave somewhat different results for Level 12 than for Levels 11 and 13. For Levels 11 and 13, alternative B was the only informative distractor indicated, while distractor D was also informative for Level 12. Distractor D, in fact, appeared for Level 12 to be more plausible than B for relatively high ability examinees. Table III.7 presents the informative distractors identified from the results for each separate level, as well as those determined from the combined results for all levels, for each of the 46 items that were administered to more than one level. The informative distractors based upon the combined results are the same as those presented

TABLE III.7 Informative Distractors Observed for Each Level and for Overall Results, and Identification of Levels for Which the Plausibility Functions Show a Consistent Pattern, for All Items Administered in Two or More Levels.

Item	Levels	Informative Distractors				Consistency of Rank of Plausibility Functions		
		Overall	Level			Levels		
			11	12	13	11,12	12,13	11,13
40	11,12	C	C	C		+		
41	11,12	C	C	C		++		
42	11,12	C	C	---				
43	11,12	---	---	---		+		
44	11,12	D	D	D				
45	11,12	D	D	D				
46	11,12	C	C	C		++		
47	11,12	D	D	---		+		
48	11,12	C	C	C				
49	11,12	C	C	C				
50	11,12	B-C	B-C	C				
51	11,12	B	B	B		+		
52	11,12	B	B	B				
53	11,12	D	D	D		++		
54	11,12	C	C	C		+		
55	11,12	B	B	B		+		
56	11,12	B-C	B-C	B-C		+		
57	11,12	---	---	---				
58	11,12,13	C	C	C	C			
59	11,12,13	---	---	---	---			
60	11,12,13	C	C	C	C	+		
61	11,12,13	B	B	D-B	B			
62	11,12,13	D	D	D	D	+	+	+
63	11,12,13	---	---	D	---			++
64	11,12,13	B	B	B	B			
65	11,12,13	C-D	C-D	C-D	C-D			++
66	11,12,13	---	---	---	---			++
67	12,13	D		D	D			
68	12,13	B		B	B			
69	12,13	---		---	---		+	
70	12,13	---		---	---		+	
71	12,13	C		C	C		+	
72	12,13	D		D	D			
73	12,13	D		D	D		+	
74	12,13	---		---	---		+	
75	12,12	---		---	---			
76	12,13	---		---	---			
77	12,13	D		D	D		+	
78	12,13	---		---	---			
79	12,13	B		---	B			
80	12,13	D-B		D-B	D-B			
81	12,13	B		B	B		+	
82	12,13	---		---	---			
83	12,13	C		C	C		+	
84	12,13	D		D	D			
85	12,13	D-C		D-C	C-D			

previously in Table III.6, for items 40 through 85. When two informative distractors were identified for an item, they are separated by a hyphen, with the distractor appealing to higher ability examinees appearing first.

Also presented in Table III.7 is an identification of the items for which the relative ranking of the alternatives with respect to the operating characteristics, or conditional probability of being selected, given ability, shows a consistent pattern between levels. For example, for item 41 (p. 81), at the lower end of the ability interval shown, for both of Levels 11 and 12, alternative A is the alternative with the highest probability of being selected, distractor C has the second highest probability of selection, distractor B is the third most probable selection, and distractor D is the least likely choice. As ability increases, these four alternatives remain in the same order in terms of the rank of their conditional probability of being chosen, given ability, up to approximately 0.5 on the latent trait for Level 11, and up to approximately -0.2 on the latent trait for Level 12. The plausibility functions for distractors B and C intersect at these points for the two levels, with distractor B being more likely than C to be selected by examinees above the point of intersection. At the high end of the ability scale (above 3.0 for Level 11 and above 2.0 for Level 12), the plausibility functions for the two distractors converge, with the conditional probability of selection approaching zero. Thus one sees basically the same pattern for the two levels, in terms of the relative rank of the alternatives with respect to likelihood of selection. General tendencies such as

these can be observed when comparing the results for two different levels, even though the precise relationship between the ability metrics for the two levels is not known, since no attempt was made in this study to equate the two scales. In the case of item 41, for example, it can be seen for both of Levels 11 and 12 that distractor C is informative, appealing primarily to lower ability examinees, while distractor B is slightly more appealing than C to relatively high ability examinees. For all ability levels except the upper extreme, for which none of the distractors is very plausible, distractor D is the alternative least likely to be selected. The results for Levels 11 and 12 for item 41 are therefore consistent in terms of the pattern of the rankings of the operating characteristics over the ability range. Table III.7 identifies items for which different levels provide consistent results, indicating such consistencies with two plus symbols ("++").

For some items the results between two levels are generally consistent, except for a slight discrepancy, or are consistent except for the extreme ranges of the latent trait. For example, the pattern observed for the plausibility functions of item 77 (p. 104) is very similar for Levels 12 and 13. For Level 12 the plausibility functions for distractors B and D intersect at the upper end of the latent trait shown in the figure, which is 4.0, whereas they intersect at approximately 2.9 on the latent trait for Level 13, and above this point of intersection the plausibility function for distractor B is higher than that of distractor D. The pattern of the rankings of the plausibility functions is generally consistent between Levels 12 and

13 up to the point on each ability scale where the plausibility functions of B and D intersect. The plausibility functions for Level 12, however, appear to be truncated at the upper extreme when compared to Level 13, which is not surprising since the ability distribution for the group of examinees taking Level 13 is expected to be higher than that of the group taking Level 12. Results such as these for item 77 were considered by the author to be generally consistent between Levels 12 and 13. Generally consistent results between levels such as these are identified by a single plus ("+") in Table III.7, to distinguish them from the more consistent results indicated by two plus symbols.

Of the 46 items administered to two or more levels, six items, items 41, 46, 53, 63, 65 and 66, were observed to provide consistent results between two or more levels. As can be further seen in Table III.7, an additional 17 items were found to have generally consistent results between two or more levels to which they were administered. Interestingly, of the nine items administered to three levels, three items (items 63, 65 and 66) had a consistent pattern of rankings of the operating characteristics for Levels 11 and 13, but not for Level 12. For item 63 (p. 95), for instance, the plausibility function for distractor D estimated for Level 12 is considerably higher than that of the other distractors at the low end of the ability range, and it does not intersect any other plausibility functions until the upper end of the ability scale. For Levels 11 and 13, on the other hand, the plausibility function for distractor D is below that of distractor C at the low end of the scale, but as ability increases slightly it

intersects the plausibility function of distractor C. Item 62 (p. 94) was the only item rated as having generally consistent results for all three levels. It should be noted, however, that the criteria used here in rating the consistency of the operating characteristics are rather strict, and that significant deviations between the patterns observed for different levels were rarely observed. The more times the plausibility functions for an item intersect with each other, the more complex the pattern of rankings becomes, and the less likely the item will yield consistent rankings following these criteria. A good example of this is item 61 (p. 93), for which the plausibility curves intersect frequently as ability increases. Although the results for the three levels are not very dissimilar, especially between Levels 11 and 13, they do not meet the criteria of consistency used here.

It was stated previously in this section that informative distractors are more likely to be detected with lower ability examinees, since they tend to be more attracted to distractors. While this may generally be the case, some interesting exceptions were observed. For items 58, 60 and 64, which were administered to all three levels, the informative distractor identified for each becomes more clearly observed as the level increases. For item 58 (p. 90), for example, the peak of the estimated plausibility function for distractor C is approximately 0.30 for Level 11, which is not too much higher than the peak for distractor D. The Level 12 results, obtained from an overall higher ability group than that of Level 11, show that the peak of the plausibility function for distractor C is

close to 0.35 , which is noticeably higher than that of any other distractor. For Level 13, the peak of the plausibility function for C is even higher, above 0.40 , and is much higher than that of either distractor B or distractor D. When comparing the estimated item characteristic functions and plausibility functions obtained from these three levels, it appears that while the difficulty of the item relative to the examinee group becomes easier as level increases (as we should expect), the plausibility of the non-informative distractors drops, while the plausibility of the informative distractor remains relatively high. If one looks at the relative frequency distribution of responses to the alternatives for each level, the tendency becomes a little clearer. For item 58, for instance, the relative frequency of examinees correctly answering the item increases from 0.38 for Level 11, to 0.50 for Level 12, and to 0.60 for Level 13. The relative frequency selecting distractor C drops only slightly, from 0.23 for Level 11, to 0.22 for Level 12, and to 0.21 for Level 13. For distractor D, however, the drop is more significant, from 0.21 (Level 11), to 0.14 (Level 12), and finally to 0.07 (Level 13). Distractor B drops in relative frequency from 0.15 , to 0.14 , to 0.11 , as level increases.

Item 83 (p. 107) also provides evidence that informative distractors are sometimes better detected when administered to higher ability examinees. In this example it is rather clear that distractor C is appealing or plausible to high ability examinees. This conclusion is somewhat clear from the results for Level 12, as the plausibility function for C reaches a peak of close to 0.30 ,

corresponding to the rather high ability level of approximately 2.7 . The plausibility of distractor C becomes even clearer for Level 13, however, as the estimated plausibility function is above 0.40 at the upper end of the latent trait, 4.0 , and appears to still be increasing with ability at the point the scale is truncated. This item seems to be a good example that a relatively high level of ability or achievement may sometimes be necessary before an examinee is drawn to a particular distractor.

It is rather clear that informative distractors are not likely to be detected when given to a group of examinees who find the item very easy and thus rarely select a distractor (Samejima, 1984a). Conversely, it should not be surprising that when an informative distractor is identified for an item, it is likely to have been selected by a relatively high frequency of examinees compared to the frequency selecting other distractors of the item. As it turned out, for 57 of the 61 Vocabulary items identified in this study as having at least one informative distractor, the distractor that was informative (or that appealed to the highest ability examinees in the case of multiple informative distractors) was also the distractor that was selected most frequently by examinees, across all levels administered. For all nine items having two informative distractors, the informative distractor appealing to the higher ability levels was the most frequently selected distractor, while the other informative distractor for the item was the second most frequently selected distractor. An item that clearly did not fit the above pattern, however, was item 79, for which the identified informative distractor,

alternative B, had the lowest relative frequency of selection of any alternative for both levels (12 and 13) to which it was administered. The relative frequency for B was slightly below that of the other distractors. For Level 12, the relative frequency of selection for B was 0.15 , compared to 0.17 for C and 0.17 for D, while for Level 13 the relative frequency for B was 0.13 , compared to 0.14 for C and 0.16 for D.

Estimated item characteristic functions. In comparing the estimated item characteristic functions obtained from different levels for the same item, it can be seen that in some cases the results for one level appear to indicate a lower asymptote at or above the chance level, 0.25 , while the results obtained from a lower level do not give such evidence. For item 63 (p. 95), for instance, the estimated item characteristic function for Level 13 appears to be approaching a lower asymptote that is slightly below 0.60 , while the Level 12 results for the same item indicate a lower asymptote approaching approximately 0.40 . The item characteristic function for Level 11 is approximately 0.20 at the lowest point on the ability scale for which it could be estimated, and it does not appear to be leveling off toward a positive lower asymptote. At the low end of the ability scale for Level 11, the estimated item characteristic function follows closely to the normal ogive curve shown in Figure III.2 . This is in contrast to the Level 12 and 13 results, that indicate the possibility that the item characteristic function follows the three-parameter logistic model.

Similar tendencies are seen for other items, such as items 55,

57, 60, 62 and 84. These results suggest that for some items that appear to follow the three-parameter logistic model, it may be the case that an accurate lower asymptote simply can not be obtained due to the ease of the item for the examinee group, and that possibly another model is better suited for the item characteristic function.

Frequency ratios. The frequency ratios which very roughly approximate the plausibility functions were obtained for the 43 items of Level 11 Vocabulary subtest. In order to compute the frequency ratios, the 2,361 examinees having finite maximum likelihood estimates for Level 11 Vocabulary subtest were divided into 30 subgroups of approximately equal size. The mean value of the maximum likelihood estimate of ability, $\hat{\theta}$, for each of the 30 subgroups and the number of examinees in each subgroup is shown in Appendix A, as Table A.1. The examinees were arranged in order of ascending maximum likelihood estimate before being divided into subgroups, and thus the mean of $\hat{\theta}$ increases from subgroup 1 to subgroup 30.

The frequency ratios estimating the plausibility functions for Level 11 Vocabulary are presented as Figure A.1 of Appendix A. As was done with the operating characteristics presented earlier in Chapter 3, the estimated item characteristic function following the normal ogive model is shown by a solid curve. The item parameters of the normal ogive item characteristic function, presented previously in Table III.3, were those obtained by the tetrachoric method. The non-parametric item characteristic function obtained by the frequency ratio procedure is shown as a dotted curve in each graph, and the frequency ratios estimating the plausibility functions of distractors

B, C, and D are presented as a short dashed curve, a dashed curve, and a long dashed curve, respectively.

The results with respect to the identification of informative distractors, as obtained both from the frequency ratios and from the estimated plausibility functions, for the 43 items of Level 11 Vocabulary subtest, are presented as Table A.2 of the appendix. Alternatives that were considered as possible informative distractors based upon frequency ratios, but that were not strongly supported, are shown in parentheses. In comparing the results obtained from the two different methods, it is observed that distractors which appeared to be informative based upon the frequency ratios were usually supported by the plausibility functions, but that many more distractors were found to be informative than were initially detected from the frequency ratios.

Based upon the frequency ratios, two items, items 56 and 65, gave evidence of having multiple informative distractors. In both cases, the results of the plausibility functions were in agreement with those of the frequency ratios. Close inspection of the frequency ratios for these two items, however, is needed to reach this conclusion, and even so the results are not clear. For item 65, in particular, it is very difficult to visually discriminate between the frequency ratio curves of the distractors, as they are quite jagged and intersect frequently. The frequency ratios for distractor C have a slightly higher peak than that of distractor D, although at many points on the maximum likelihood scale, the frequency ratio for D is higher than the corresponding frequency ratio for C. Both distractors

appear to be informative, but the results for distractor D are rather questionable. The picture is clearer when the estimated plausibility functions in Figure III.2 are observed. Unlike the jagged, intersecting frequency ratio curves, the plausibility functions yield smooth curves that more clearly indicate the relationship between ability and probability of selection of a given alternative. For item 65 it is seen that distractor C is more plausible for relatively high ability examinees, while distractor D is frequently selected by examinees of low ability. On the other hand, it is almost impossible to separate the frequency ratios of the three distractors at the low end of the scale based upon the frequency ratios. As will be seen in more examples to follow, this problem is not uncommon when observing results obtained from the frequency ratio procedure.

Of the 12 items for which the frequency ratios indicated a single informative distractor, 10 were consistent with the results obtained from the plausibility functions. A good example of such a case is item 46. Even though the frequency ratios are somewhat jagged, as is typically the case, it is clear that distractor C is likely to be informative. This alternative is selected with a relatively high conditional probability, given ability, over a wide range of the scale, and a fairly distinct peak is seen for the jagged curve corresponding to this distractor. The two items for which the frequency ratios inaccurately appeared to have one informative distractor were items 37 and 59. For item 37 a single informative distractor, alternative C, was identified from the frequency ratios, but a second informative distractor, D, was observed from the

plausibility functions. There was some evidence based upon the frequency ratios that distractor D was informative, as the frequency ratio curve reached a peak of approximately 0.4, corresponding to the lowest ability subgroup. The evidence was not considered strong, however, since it was highly dependent upon only one or two subgroups at the low end of the scale. For item 59 the frequency ratio curve for distractor D gave weak evidence of being informative but was not supported by the plausibility functions. This is not too surprising in looking more closely at the frequency ratios, which are relatively flat, despite being jagged, and that peak at about the chance level of probability.

Of the 14 items that were categorized as indicating a "possible" informative distractor based upon the frequency ratios, the distractor identified turned out to be informative for 10 of those items, while for the other four no informative distractor was observed from the plausibility functions. As an example of the former case, distractor D of item 62 was identified as a possible informative distractor, but was not considered to be strongly supported by the frequency ratios since they were quite erratic and did not consistently surpass the chance level of probability. The plausibility functions, however, clearly identified distractor D as being informative. For item 32, on the other hand, distractor B was selected to the possible distractor category but did not turn out to be informative based upon the estimated plausibility functions.

Five items that were categorized as not indicating informative distractors, given their frequency ratios, did not show informative

distractors in their plausibility functions. Ten items that were similarly categorized based upon frequency ratios, however, were later found to have informative distractors. Item 61, for example, did not indicate an informative distractor, as its frequency ratio curves are not very high and overlap so much that they are very difficult to discern. The plausibility functions indicated that distractor B was informative, appealing primarily to relatively low ability examinees. Item 50 is noteworthy, since no informative distractor was seen from the frequency ratios, yet two were later identified based upon the plausibility functions. For both of these distractors, B and C, the frequency ratio curves are clearly higher than that of distractor D, but they are relatively flat and intersect frequently with each other. Neither distractor appears to be particularly informative or plausible. Even in observing the plausibility functions, this is not an easy item for which to classify the distractors, as the curves corresponding to distractors B and C are not very steep or especially high. However, the plausibility functions do give a little more evidence than do the frequency ratios that these two distractors are informative.

Comparison of Content-based and Latent Trait Results

Information presented in Tables III.1 and III.6 earlier in this chapter is combined in Table III.8 in order to facilitate the comparison of the observed informative distractors with those predicted by the author, for all 82 items of Levels 11, 12 and 13 of the Vocabulary subtest. After the alternative letter of each informative distractor, the ranking assigned by the author to that

TABLE III.8 Comparison of Author's Predictions of Informative Distractors to Observed Results for 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest.

	Item	Predicted	Observed	Rule
1	24	D	---	---
2	25	B	---	---
3	26	C	C(2)	7
4	27	D-C	C(3)	1
5	28	B-C	B(2)	1,4,5
6	29	C-B	---	---
7	30	B	B(2)	3,4,6
8	31	D-B	C(4)	6
9	32	C-B	---	---
10	33	C-B	---	---
11	34	D	D(2)	1
12	35	C-B	C(2)	1,8
13	36	D	D(2)	1,4,6,8
14	37	C-D	C(2)-D(3)	3,4-5
15	38	B-C	B(2)	1,3
16	39	C-B	D(4)	5
17	40	C-B	C(2)	1,5
18	41	C	C(2)	1,3,4
19	42	C	C(2)	1,5
20	43	B	---	---
21	44	B-C	D(4)	1
22	45	D-B	D(2)	1,5
23	46	C-B	C(2)	3,4
24	47	D-B	D(2)	3,4
25	48	C-B-D	C(2)	6
26	49	B-D	C(4)	4,7,8
27	50	B	B(2)-C(3)	1,3,4-1
28	51	B-C	B(2)	1,3,4
29	52	B-C	B(2)	1
30	53	C-D	D(3)	3,5,7
31	54	B-D	C(4)	6
32	55	B-D	B(2)	1,5
33	56	B-C	B(2)-C(3)	1,3-3
34	57	C	---	---
35	58	B-C-D	C(3)	5
36	59	D-C	---	---
37	60	C	C(2)	1,3,4,5
38	61	C-D	B(4)	1,3,4
39	62	D	D(2)	1
40	63	D	---	---

TABLE III.8 (Continued)

	Item	Predicted	Observed	Rule
41	64	B	B(2)	1,5
42	65	C-B	C(2)-D(4)	1,4-4,8
43	66	C-B	---	---
44	67	B	D(4)	6
45	68	B	B(2)	2
46	69	B	---	---
47	70	D-B	---	---
48	71	C	C(2)	1
49	72	D	D(2)	3
50	73	D	D(2)	5
51	74	B	---	---
52	75	---	---	---
53	76	B	---	---
54	77	C-D	D(3)	1
55	78	---	---	---
56	79	B-C	B(2)	1
57	80	D-B	D(2)-B(3)	1-1,3
58	81	B	B(2)	4,5
59	82	B	---	---
60	83	C	C(2)	1,4,6
61	84	D	D(2)	2
62	85	C	D(4)-C(2)	1-3,5,6
63	86	D-B	D(2)	1,3,5
64	87	---	C(3)-B(2)	3-5
65	88	C	---	---
66	89	---	B(4)	8
67	90	B	C(3)-B(2)	1-3
68	91	---	---	---
69	92	---	D(3)	1,6
70	93	D	C(3)	5
71	94	C-B	---	---
72	95	D	D(2)	1,4
73	96	C	C(2)	1
74	97	B	B(2)	3,4
75	98	B-C	B(2)-C(3)	4,6-5
76	99	D	D(2)	7
77	100	C	---	---
78	101	C-B	B(3)	3,8
79	102	B-D	D(3)	5,7
80	103	D	D(2)	3,7
81	104	B-C	B(2)	1,4
82	105	C-B	B(3)	1

alternative is given in parentheses. When more than one informative distractor was predicted or was observed for an item, the alternative letters are separated by a hyphen. In addition, Table III.8 gives the number of the rule or reason identified by the content analysis corresponding to each of the observed informative distractors. These rules are numbered as they were when they were previously presented in the first section of this chapter.

It can be seen from Table III.8 that 76 of the 82 items were predicted to have informative distractors and that 61 items actually did. The predicted results matched the observed results exactly for 28 of the 82 items, or for 34.1 percent, including three items in which no distractors were predicted to be informative and none were observed. For 41 items, or 67.2 percent, of the 61 items having informative distractors, the distractor appealing to the relatively highest ability levels of examinees was correctly predicted. Of the nine items having two informative distractors, these distractors were correctly predicted, in the proper order, four times (44.4 percent).

Table III.9 presents the average or mean predicted rankings for the various categories of distractors with regard to being informative or non-informative. Also presented are the theoretical rankings that would have resulted from perfect prediction of informative distractors. The mean of the author's rankings for all distractors that were observed to attract the highest level of examinees was 2.492 . (Had the author predicted these 61 distractors with perfect accuracy, each would have had a ranking of 2 and thus this mean would have been 2.0 .) The mean of the rankings for the second highest

TABLE III.9 Mean of Rankings by Author for Various Categories of Distractors with Regard to Information Provided, and Theoretical Rankings for Each Category if Perfectly Predicted. Levels 11, 12 and 13 of Vocabulary Subtest.

Information	Theoretical	Observed
<u>Informative</u>		
Higher	2.000	2.492
Lower	3.000	2.778
All	2.129	2.529
<u>Non-informative</u>		
2 Distractors	3.500	3.250
1 Distractor	4.000	3.778
All	3.347	3.188

distractor of those nine items having two informative distractors was 3.145 . When the mean ranking was computed for all non-informative distractors of items having only one informative distractor, the result was 3.250 . For items having two informative distractors, the mean ranking of the non-informative distractor was 3.778 . From these results we see that the means of the author's rankings increase in the same direction as do the theoretical means.

Table III.10 specifies the number of times each of the eight rules identified from the content analysis appeared as a reason for selection of a distractor that was observed to be informative. The frequency distributions are given separately for the informative distractors appealing to the highest ability level and, for those items having two informative distractors, for the second highest informative distractor. Also presented is the total frequency distribution, for all informative distractors. As can be seen in Table III.10, for most of the observed informative distractors there were two or more rules identified for one distractor. The frequency of times each rule appeared by itself for an informative distractor is given in parentheses in Table III.10. We see from this table that rule or reason 1, that a distractor was similar in meaning, appears 36 times in total, and that for 14 informative distractors this was the only reason identified. The next most identified rules were numbers 3 and 4, which deal with the sensibility and readability of the distractor within a given context. Reason 5, that a distractor was similar in spelling to a key word was close behind in total frequency, and was actually the second most frequent reason given alone. The

TABLE III.10 Number of Times Each Rule Identified from Content Analysis of Vocabulary Subtest Appeared as Possible Reason for Selection of Distractors Observed to Be Informative.

Rule	Highest Informative Distractor	Lower of 2 Informative Distractors	All Informative Distractors
1	34 (13)	2 (1)	36 (14)
2	2 (2)	0	2 (2)
3	18 (2)	4 (2)	22 (4)
4	19 (0)	1 (0)	20 (0)
5	14 (4)	4 (3)	18 (7)
6	10 (5)	1 (0)	11 (5)
7	5 (2)	0	5 (2)
8	5 (1)	1 (0)	6 (1)

least frequent rule corresponding to informative distractors was number 2, that a word or part of a word was taken out of context. This rule was identified only twice for informative distractors.

The observed results for the 16 items ranked by faculty and teachers are compared to the predictions, or mean rankings for these evaluators, in Table III.11. The alternatives are first presented in the order of ascending mean ranking, or descending ability level of the examinees predicted to find them plausible, and then the observed results are presented, with the mean ranking for each observed informative distractor given in parentheses. The predicted order is consistent with the observed results for four of the 13 items having informative distractors. For some of the items the mean ranking of the observed informative distractors was very close to that of the distractor having the lowest mean ranking, and for item 104 the observed informative distractor, alternative B, was tied with alternative C as having the lowest mean ranking among distractors. There were two items, 53 and 65, for which the highest observed informative distractor was predicted to appeal to the lowest ability groups, or that had the highest mean rankings.

The means of the faculty/teacher rankings for the different categories with regard to information provided by the distractors were computed, as was done for the author's rankings, and the results are presented in Table III.12. As can be seen from this table, the mean ranking of all distractors observed to be informative is 2.859 , which is a little lower than the mean ranking for all distractors which were not informative, which was 3.047 .

TABLE III.11 Comparison of Predicted Informative Distractors Obtained by Faculty, Teachers and Author to Observed Results for 16 Selected Items of Vocabulary Subtest.

	Item	Predicted	Observed
1	27	D-C-B	C(2.500)
2	37	B-C-D	C(2.750)-D(3.625)
3	40	B-C-D	C(2.750)
4	53	C-B-D	D(3.375)
5	56	B-C-D	B(2.625)-C(2.750)
6	59	D-C-B	---
7	60	C-B-D	C(2.125)
8	64	B-C-D	B(2.500)
9	65	D-B-C	C(3.375)-D(2.750)
10	67	B-D-C	D(3.125)
11	77	B-D-C	D(2.875)
12	84	D-B-C	D(2.000)
13	86	B-D-C	D(2.625)
14	94	C-B-D	---
15	100	C-B-D	---
16	104	B,C-D	B(2.500)

TABLE III.12 Mean of Rankings by Faculty, Teachers and Author for Various Categories of Distractors with Regard to Information Provided, and Theoretical Rankings for Each Category if Perfectly Predicted. Sixteen Selected Items of Vocabulary Subtest.

Information	Theoretical	Observed
<u>Informative</u>		
Higher	2.000	2.817
Lower	3.000	3.042
All	2.188	2.859
<u>Non-informative</u>		
2 Distractors	3.500	3.075
1 Distractor	4.000	3.042
All	3.406	3.047

Linking Content-based Rules to Examinee Ability

Since the empirical results for many of the Vocabulary items clearly indicated that the relative plausibility of the distractors differed with examinee ability, the emphasis of the study shifted toward an investigation of the relationship between the ability range for which a distractor is highly plausible and the rules or reasons behind the selection of the distractor. The primary purpose of this part of the study was to determine if examinees of similar ability may have been consistently following the same line of reasoning when responding to a series of items. If so, certain rules should tend to be followed by lower ability examinees while others should tend to be followed by examinees of higher ability.

The question of whether the previously identified rules differed in terms of their likelihood of yielding informative distractors was first investigated. The content-based rules corresponding to the informative distractors were presented earlier, but the rules identified for non-informative distractors were not shown. Table III.13 presents the rules corresponding to all of the Vocabulary distractors, with informative distractors denoted by an asterisk. For some non-informative distractors none of the eight rules previously identified seemed appropriate, so dashed lines are shown rather than rule numbers for these distractors.

Further inspection of Table III.13 shows a total of 124 distractors for which a single rule was identified. Of these 124 distractors, 34 were found to be informative. This yields a percentage of 27.4 that were informative. There were 71 distractors

TABLE III.13 Rule Numbers Corresponding to Each Distractor for the 82 Items of the Vocabulary Subtest. Informative Distractors Are Denoted by an Asterisk.

Item	Distractor		
	B	C	D
24	---	---	1,4,6
25	1,3	8	---
26	6	7*	2
27	6,7	1*	1
28	1,4,5*	1	4
29	7	1	7
30	3,4,6*	2	3,4,7
31	4,6	6*	1
32	1,4,6	---	---
33	5	3,4,7	---
34	6	3,4,7	1*
35	1	1,8*	4,7
36	7	7	1,4,6,8*
37	5	3,4*	5*
38	1,3*	2	7
39	1,4	1,4,6	5*
40	1,4	1,5*	7
41	2	1,3,4*	8
42	---	1,5*	---
43	4,5	3	5
44	3,4	3	1*
45	---	7	1,5*
46	---	3,4*	---
47	3,4,6	---	3,4*
48	3,4	6*	3
49	1,3,4	4,7,8*	1
50	1,3,4*	1*	6
51	1,3,4*	3,4	5
52	1*	1	3,4
53	5	1	3,5,7*
54	1	6*	1
55	1,5*	1,3	2
56	1,3*	3*	3
57	---	3,4,5	---
58	1,4,5	5*	6
59	7	---	1
60	5	1,3,4,5*	1
61	1,3,4*	1,3,4,5	2
62	5	---	1*
63	---	---	6

TABLE III.13 (Continued)

Item	Distractor		
	B	C	D
64	1,5*	1,4	---
65	1	1,4*	4,8*
66	---	5	3,6
67	4,9	4,7	6*
68	2*	5,6	5
69	1	---	2
70	4	---	1,3
71	1	1*	6
72	1	5,7	3*
73	4,6	---	5*
74	1,6	---	---
75	---	---	---
76	1	7	3
77	---	1,4,5	1*
78	---	6	---
79	1*	5	---
80	1,3*	8	1*
81	4,5*	8	7
82	6	3	5
83	3	1,4,6*	1
84	2	---	2*
85	4	3,5,6*	1*
86	3	---	1,3,5*
87	5*	3*	---
88	7	4	6
89	8*	3	---
90	3*	1*	---
91	7	---	6
92	1	6	1,6*
93	---	5*	3
94	5	---	---
95	---	1	1,4*
96	8	1*	7
97	3,4*	6	---
98	4,6*	5*	---
99	6	---	7*
100	3,6	5	---
101	3,8*	1,5	---
102	1	6	5,7*
103	---	---	3,7*
104	1,4*	1	---
105	1*	1	---

corresponding to multiple rules. Of these 71, 36 distractors, or 50.7 percent, were found to be informative. A Pearson chi-square test of association based upon these proportions yielded a statistic of 10.63 (1 degree of freedom), which was significant at the 0.01 level. Thus there appears to be a relationship between multiple rules being identified for a distractor and the distractor being informative.

Table III.14 presents the total number of distractors identified with each rule or reason, the number of these distractors that were found to be informative, and the resultant percentage of distractors that were informative. Those distractors corresponding to multiple rules were included in the count for each of these rules in Table III.14. It can be observed in this table that some rules were much more frequently identified than others, ranging from a low of 10 for rule number 2 (meaning taken out of context), to a high of 74 for number 1 (similarity in meaning). Reason 8 (familiarity) had the greatest likelihood of yielding an informative distractor, with this being the case for 54.5 percent of the distractors. Reasons 1 (similarity in meaning), 5 (similarity in spelling), and 3 (logical in context) were close behind. On the other hand, reasons 2 (meaning taken out of context) and 7 (opposite in meaning) yielded an informative distractor less than 25 percent of the time. A Pearson chi-square test of association, however, did not indicate a statistically significant relationship between the rules identified with a distractor and whether or not the distractor was informative. Furthermore, these percentages likely depend upon the item difficulty, as easy items should be less likely to have empirically identified

TABLE III.14 Number of Distractors Corresponding to Each Rule, and the Number and the Percent Which Are Informative, for 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest.

Rule No.	No. of Dstrs.	No. Info.	Pct. Info.	Rule
1	74	36	48.6	Similarity in Meaning
2	10	2	20.0	Meaning Taken Out of Context
3	48	22	45.8	Logical in Context
4	49	20	40.8	Readability of the Phrase
5	40	19	47.5	Similarity in Spelling
6	35	10	28.6	Similarity in Length
7	26	6	23.1	Opposite in Meaning
8	11	6	54.5	Familiarity

informative distractors.

Table III.15 presents the mean item difficulty for items having no informative distractor, one informative distractor, or two informative distractors, by level of the test. It is seen that for all three levels the mean difficulty is lowest for items not having informative distractors, as should be expected. The mean difficulty for items having two informative distractors is higher than that for one informative distractor at Levels 11 and 12, but it is slightly lower at Level 13. It should be noted, however, that the number of items with two informative distractors is only between four and six for each level. A comparison of mean difficulty between items with one informative distractor and items without an informative distractor yields a t statistic of 2.259 at Level 11, 3.231 at Level 12, and 3.888 at Level 13, each of which is significant at the 0.05 level. A comparison of mean difficulty between the two informative distractor case and the single informative distractor case was significant ($p < .05$) only for Level 11, which yielded a t statistic of 1.929.

In order to study the relationship between examinee ability and the reasons behind the choice of distractor, the ability range for which an informative distractor is considered plausible must be defined, or quantified. One approach is to take the peak of the plausibility function as a single measure of ability for which the informative distractor is considered plausible. Another approach is to look at the range of ability for which the plausibility function is greater than the chance level, which is 0.25 for a multiple-choice

TABLE III.15 Number of Items and Mean Item Difficulty Parameter for Items Having No Informative Distractor, One Informative Distractor, and Multiple Informative Distractors, for Each of Levels 11, 12 and 13 of Vocabulary Subtest.

	No Informative Distractor	One Informative Distractor	Multiple Informative Distractors
<u>Level 11</u>			
Number of Items	10	29	4
Mean Item Difficulty	-1.144	-0.135	0.482
<u>Level 12</u>			
Number of Items	11	28	5
Mean Item Difficulty	-0.997	-0.252	0.037
<u>Level 13</u>			
Number of Items	12	27	6
Mean Item Difficulty	-1.325	-0.148	-0.169

item with four alternatives. This range gives a broader picture of the plausibility of the distractor, and in this paper this ability range will be referred to as the plausibility range. Table III.16 presents the plausibility range and the peak of the plausibility function for each level of the test, as well as the content-based rules corresponding to each informative distractor. In cases where the peak was apparently outside of the ability range for which the plausibility function could be generated, a "less than" (<) or "greater than" (>) sign is used in the table.

It can be seen from Table III.16 that for many items the peak of the plausibility function appeared to correspond to a point below the ability range that could be generated from the empirical data (see Appendix B for further discussion of this range). For this reason the peak alone may not be a good measure to discriminate between distractors in terms of the ability level of the examinee finding the distractor plausible. The midpoint or the upper bound of the plausibility range may be a better single measure. One limitation of the midpoint is that it depends upon the lower bound as well as the upper bound of the interval, and for almost all of the informative distractors the plausibility function is above the chance level at the lowest value of ability for which the function could be generated. Thus the lower bound is often restricted by the ability distribution of the examinees taking each level.

A frequency count, by rule and by level of the test, of all informative distractors corresponding to a single content-based rule is shown in Table III.17. Informative distractors for which multiple

TABLE III.16 Plausibility Range and Peak of Plausibility Function of Each Informative Distractor, for Each of Levels 11, 12 and 13 of Vocabulary Subtest.

Item	Dstr	Rule(s)	Level 11		Level 12		Level 13	
			Plaus.Range	Peak	Plaus.Range	Peak	Plaus.Range	Peak
26	C	7	-1.6,-0.1	-1.5				
27	C	1	-1.6, 0.4	<-1.6				
28	B	1,4,5	-1.6,-0.8	-1.2				
30	B	3,4,6	-1.6,-0.5	<-1.6				
31	C	6	-1.6, 0.0	-0.9				
34	D	1	-1.6,-0.5	<-1.6				
35	C	1,8	-1.6,-0.9	-1.2				
36	D	1,4,6,8	-1.6,-0.4	-1.1				
37	C	3,4	-1.6, 2.1	0.7				
	D	5	-1.6,-0.7	<-1.6				
38	B	1,3	-1.6,-0.1	-1.6				
39	D	5	-1.6, 0.3	<-1.6				
40	C	1,5	-1.6, 0.4	-0.4	-2.4,-0.1	<-2.4		
41	C	1,3,4	-1.6,-0.8	<-1.6	-2.4,-1.4	<-2.4		
42	C	1,5	-1.6,-0.3	<-1.6	-2.4,-0.8	-2.4		
44	D	1	-1.6, 2.8	-0.1	-2.4, 1.3	-1.2		
45	D	1,5	-1.6, 0.0	-1.0	-2.4,-0.7	-2.4		
46	C	3,4	-1.6, 0.3	-1.2	-2.4,-0.3	-2.4		
47	D	3,4	-1.6,-0.9	<-1.6	---	---		
48	C	6	-1.6, 0.0	<-1.6	-2.4,-0.8	-1.2		
49	C	4,7,8	-1.6,-0.6	<-1.6	-2.4,-0.9	<-2.4		
50	B	1,3,4	-1.6, 0.0	<-1.6	-2.4,-0.9	<-2.4		
	C	1	-1.6,-0.9	<-1.6	-2.4,-0.9	<-2.4		
51	B	1,3,4	-1.6,-0.6	<-1.6	-2.4,-1.1	-2.4		
52	B	1	-1.6, 0.0	<-1.6	-2.4,-0.7	-2.4		
53	D	3,5,7	-1.6, 2.0	-0.3	-2.4, 1.3	-0.8		
54	C	6	-1.6,-0.4	<-1.6	-2.4,-0.9	-1.6		
55	B	1,5	-1.6,-0.4	<-1.6	-2.4,-0.9	-2.4		
56	B	1,3	-1.6, 1.4	-0.3	-2.4, 1.1	-2.1		
	C	3	-1.6, 0.6	-0.4	-2.4, 0.0	-0.8		
58	C	5	-1.6,-0.2	-1.2	-2.4,-0.3	-1.1	-2.4,-0.5	<-2.4
60	C	1,3,4,5	-1.6, 0.7	-1.1	-2.4, 0.0	<-2.4	-2.4,-0.1	<-2.4
61	B	1,3,4	-1.6,-0.8	<-1.6	-2.4,-1.4	-2.4	-2.4,-1.3	<-2.4
62	D	1	-1.6,-0.7	<-1.6	---	---	-2.4,-1.7	<-2.4

TABLE III.16 (Continued)

Item	Dstr	Rule(s)	Level 11		Level 12		Level 13	
			Plaus.Range	Peak	Plaus.Range	Peak	Plaus.Range	Peak
64	B	1,5	-1.6, 0.6	-1.0	-2.4, 0.1	<-2.4	-2.4,-0.1	<-2.4
65	C	1,4	-1.1, 0.6	-0.1	-2.4,-0.1	-0.8	-2.4,-0.4	-1.2
	D	4,8	---	---	-2.4,-0.5	<-2.4	-2.4,-1.0	<-2.4
67	D	6			-2.4, 1.2	-1.3	-2.4, 0.4	-1.0
68	B	2			-2.4, 0.5	-2.4	-2.4,-0.2	-1.2
71	C	1			-2.4, 0.1	<-2.4	-2.4,-0.1	<-2.4
72	D	3			-2.4, 1.9	0.1	-2.4, 1.9	0.1
73	D	5			-2.4,-0.8	-2.4	-2.4,-0.8	-2.4
77	D	1			-2.4, 1.9	-0.1	-2.4, 1.2	-1.0
79	B	1			---	---	-2.4,-1.3	<-2.4
80	D	1			-1.1, 0.2	-0.8	-2.4,-0.8	-1.9
	B	1,3			-2.4,-1.2	<-2.4	-2.4,-1.4	<-2.4
81	B	2			-2.4, 0.9	0.0	-2.4, 0.4	-0.7
83	C	1,4,6			1.3, 3.3	2.6	-0.1, 4.0	> 4.0
84	D	2			-2.4, 2.0	-0.4	-2.4, 1.8	-1.0
85	D	1			-2.4, 0.9	-1.0	-2.4,-1.0	<-2.4
	C	3,5,6			---	---	-2.4,-1.0	-1.7
86	D	1,3,5					-2.4, 2.1	-0.2
87	C	3					-2.4, 0.2	-2.1
	B	5					-2.4,-1.4	<-2.4
89	B	8					-2.4,-0.4	-1.2
90	C	1					-2.4, 0.3	-2.4
	B	3					-2.4,-0.9	-2.4
92	D	1,6					-2.4,-1.5	<-2.4
93	C	5					-2.4,-1.0	<-2.4
95	D	1,4					-2.4, 0.5	-0.9
96	C	1					-2.4, 3.8	-0.8
97	B	3,4					-2.4, 0.5	-1.0
98	B	4,6					-1.3, 0.7	-0.2
	C	5					-2.4, 0.3	-0.6
99	D	7					-2.4, 1.7	-0.8
101	B	3,8					-2.4, 0.8	-1.1
102	D	5,7					-2.4, 4.0	2.3
103	D	3,7					-2.4, 0.8	-0.9
104	B	1,4					-2.4, 1.7	-0.2
105	B	1					-2.4,-0.3	<-2.4

TABLE III.17 Mean Value of Ability, θ , at the Upper Bound and at the Midpoint of the Plausibility Range and at the Peak of the Plausibility Function, by Rule, for Informative Distractors Corresponding to a Single Rule, for Levels 11, 12 and 13 of Vocabulary Subtest.

Rule	<u>No. Dstrs.</u>			<u>Mean at Upper Bound</u>			<u>Mean at Midpoint</u>			<u>Mean at Peak</u>		
	<u>Level</u>			<u>Level</u>			<u>Level</u>			<u>Level</u>		
	11	12	13	11	12	13	11	12	13	11	12	13
1	6	7	9	0.183	0.400	0.011	-0.708	-0.907	-1.194	-1.433	-1.500	-2.067
2	0	2	2	---	1.250	0.800	---	-0.575	-0.800	---	-1.400	-1.100
3	1	2	3	0.600	0.950	0.400	-0.500	-0.725	-1.000	-0.400	-0.350	-1.467
4	0	0	0	---	---	---	---	---	---	---	---	---
5	3	2	5	-0.200	-0.550	-0.680	-0.900	-1.475	-1.540	-1.533	-1.750	-2.100
6	3	3	1	-0.133	-0.167	0.400	-0.867	-1.283	-1.000	-1.433	-1.367	-1.000
7	1	0	1	-0.100	---	1.700	-0.850	---	-0.350	-1.500	---	-0.800
8	0	0	1	---	---	-0.400	---	---	-1.400	---	---	-1.200

rules were identified are not included. It can be seen that relatively small numbers of informative distractors correspond to a single rule, with rule number 1 (similarity in meaning) being the only rule for which more than five informative distractors were found within a given level. The mean value of ability, θ , at the upper bound and at the midpoint of the plausibility range is also presented in Table III.17, for each rule, by test level. Also shown is the mean value of ability at the peak of the plausibility function. The usefulness of the data in Table III.17 must be questioned, due to the small frequencies of informative distractors identified with a single rule. For some rules there were no informative distractors found within a given level. For these reasons it was decided to also look at data for informative distractors corresponding to multiple content-based rules or reasons, although this multiplicity might complicate the interpretation of the data.

Table III.18 presents the number of informative distractors at each test level corresponding to each rule when all informative distractors are considered. Thus distractors identified as having multiple rules will be counted more than once in this table. The mean value of ability at the upper bound and at the midpoint of the plausibility range and at the peak of the plausibility function, by rule and by test level, are also shown in this table. Although rules 2, 6, 7 and 8 still have rather small frequencies, the addition of informative distractors linked to multiple rules added considerably to the totals for rules 1, 3, 4 and 5.

The data in Table III.18 are rather difficult to interpret in

TABLE III.18 Mean Value of Ability, θ , at the Upper Bound and at the Midpoint of the Plausibility Range and at the Peak of the Plausibility Function, by Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest.

Rule	<u>No. Dstrs.</u>			<u>Mean at Upper Bound</u>			<u>Mean at Midpoint</u>			<u>Mean at Peak</u>		
	<u>Level</u>			<u>Level</u>			<u>Level</u>			<u>Level</u>		
	11	12	13	11	12	13	11	12	13	11	12	13
1	22	21	19	-0.014	-0.062	0.189	-0.795	-1.112	-1.045	-1.264	-1.800	-1.553
2	0	2	2	---	1.250	0.800	---	-0.575	-0.800	---	-1.400	-1.100
3	13	11	11	0.262	-0.182	0.145	-0.669	-1.291	-1.127	-1.108	-1.891	-1.527
4	13	11	10	-0.131	-0.218	0.500	-0.846	-1.141	-0.780	-1.223	-1.627	-0.760
5	11	10	11	0.145	-0.130	0.173	-0.727	-1.265	-1.114	-1.182	-1.900	-1.436
6	5	4	5	-0.260	0.700	0.520	-0.930	-0.388	-0.600	-1.420	-0.375	-0.260
7	3	2	3	0.433	0.200	2.167	-0.583	-1.100	-0.117	-1.167	-1.650	0.200
8	3	2	3	-0.633	-0.700	-0.200	-1.117	-1.550	-1.300	-1.333	-2.500	-1.600

this form, so these means are shown graphically in Figures III.3, III.4 and III.5. Figure III.3 plots the mean ability at the upper bound of the plausibility range corresponding to each rule, for each level of the test. Ability, θ , is given as the abscissa, but it should be kept in mind that the ability scale for each level is different. Each scale is linear and thus the relative distance between means for a given level is meaningful, but distances cannot be compared across levels. The data are presented in this form to look for general tendencies, both within and across levels. The mean on each ability scale corresponding to a particular rule is represented by the rule number of that rule.

One sees in Figure III.3, for example, that the mean ability level at the upper bound of the plausibility range was clearly the lowest at every level for informative distractors corresponding to rule number 8 (familiarity). It appears from this data that distractors composed of words relatively familiar to this age group tend to be most attractive to lower ability examinees. On the other hand, rule number 7 (opposite in meaning) attracts the highest ability group, on the average, for Levels 11 and 13, and has the third highest mean for Level 12. Rule number 2 (meaning taken out of context) also appears to be followed by examinees of relatively high ability, with the highest mean ability at Level 12 and the second highest at Level 13. There were no informative distractors corresponding to rule 2 among the Level 11 items.

Figure III.4 presents the mean ability at the midpoint of the plausibility range for each rule at each level of the test. Once

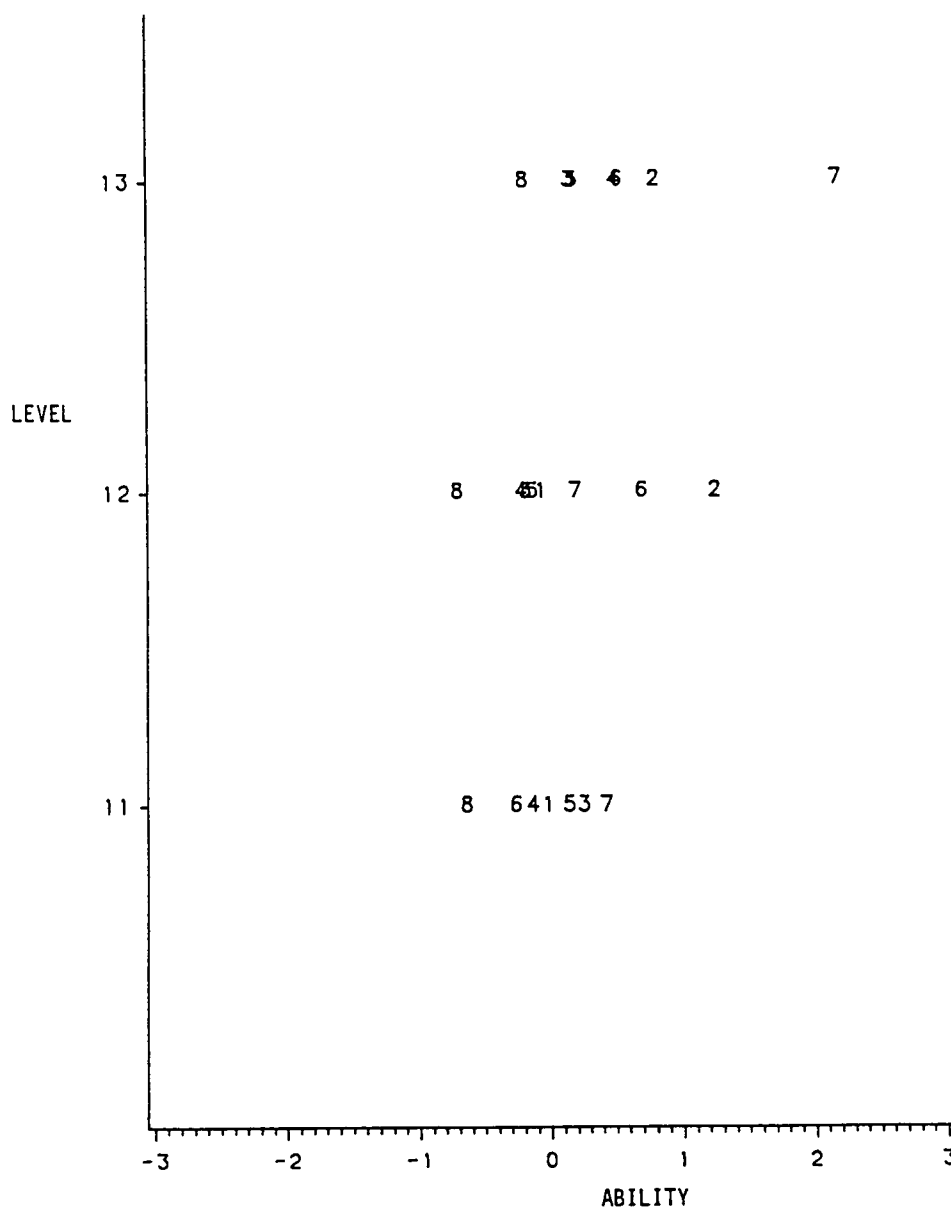


FIGURE III.3 Mean Value of Ability, θ , at the Upper Bound of the Plausibility Range Corresponding to Each Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest. Each Mean Is Represented by the Appropriate Rule Number.

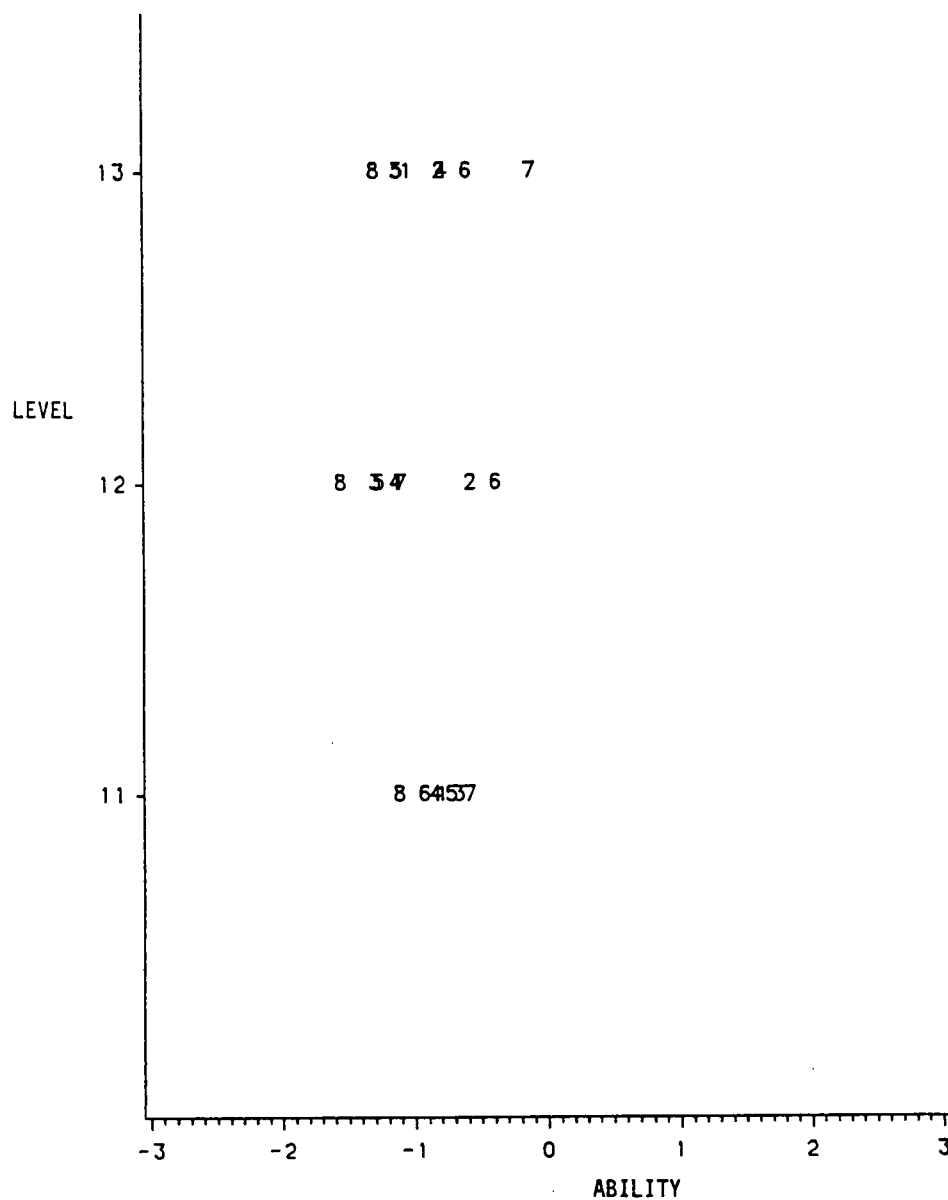


FIGURE III.4 Mean Value of Ability, θ , at the Midpoint of the Plausibility Range Corresponding to Each Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest. Each Mean Is Represented by the Appropriate Rule Number.

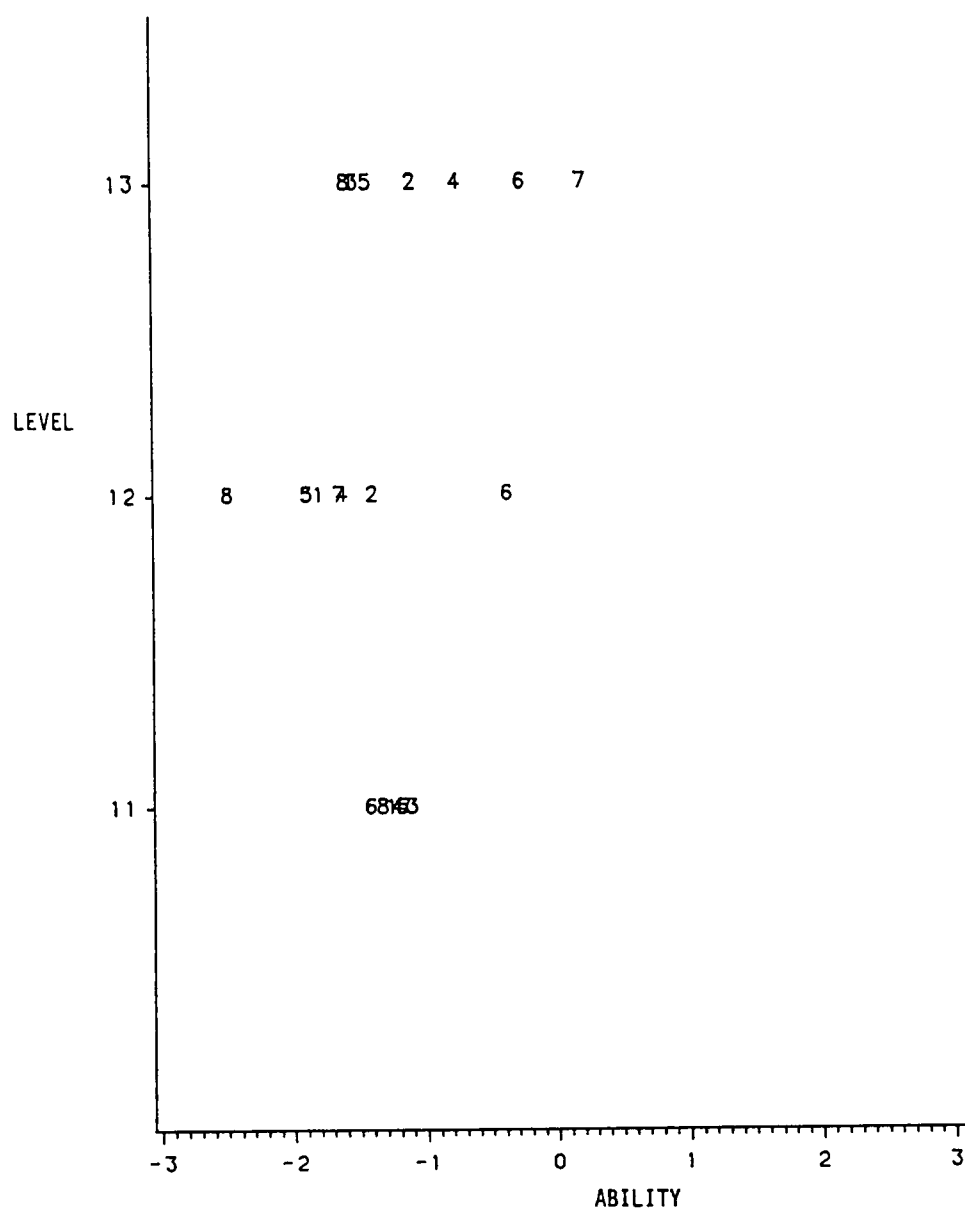


FIGURE III.5 Mean Value of Ability, θ , at the Peak of the Plausibility Function Corresponding to Each Rule, for All Informative Distractors, for Levels 11, 12 and 13 of Vocabulary Subtest. Each Mean Is Represented by the Appropriate Rule Number.

again it is seen that rule 8 has the lowest mean value at all three levels and that rule 7 has the highest mean ability for two of the three levels. Rule 6 (similarity in length) has the highest mean at Level 12 and the second highest at Level 13, but has the second lowest mean ability at Level 11.

The mean ability at the peak of the plausibility function corresponding to each rule is plotted in Figure III.5. This time rule 8 is lowest for Levels 12 and 13 and second lowest for Level 11. Rule 7 is the highest for Level 13 and second highest for Level 11, but it ranks only fourth for Level 12. Rule 6 has the highest mean for Level 12 and the second highest for Level 13, but it has the lowest for Level 11.

One of the limitations with the means presented in Table III.18 and plotted in Figures III.3 through III.5 is that all informative distractors for which a given rule was identified were equally weighted in computing the mean ability for that rule, regardless of whether the distractor was associated with a single rule or many rules. The purpose was to use as many distractors per rule as possible, since there were so few corresponding to some of the rules. However, these results are complicated by the fact that most of the distractors are linked to multiple rules, and while one rule might tend to attract lower ability examinees, another rule associated with the same distractor might tend to be attractive to higher ability examinees. The net effect of these multiple rules is difficult to interpret when a simple mean is computed in the above manner. In an attempt to account for the effects of multiple rules in a more

discernible manner, a multiple regression approach was used. Letting $\bar{\theta}$ represent the ability level defined by the upper bound or midpoint of the plausibility range, or the peak of the plausibility function, the multiple regression model may be expressed as

$$\bar{\theta} = \beta_0 + \beta_1 R_1 + \beta_2 R_2 + \cdots + \beta_j R_j + \cdots + \beta_8 R_8 + \epsilon, \quad (3.1)$$

where β_0 is the intercept, β_j is the multiple regression coefficient corresponding to rule j , ϵ is an error term, and R_j is an indicator variable used to identify whether rule j pertains to the distractor such that

$$R_j \begin{cases} = 0 & \text{when rule } j \text{ does not pertain,} \\ = 1 & \text{when rule } j \text{ pertains.} \end{cases} \quad (3.2)$$

Table III.19 presents the multiple regression coefficients corresponding to the eight content-based rules, for each of Levels 11, 12 and 13, when trying to predict the value of the upper bound, the midpoint, and the peak, respectively. These coefficients were obtained through the use of the REG procedure of the statistical package SAS. Also presented in Table III.19 is the average rank of the coefficients for a given rule across levels and ability measures (upper bound, midpoint and peak), from lowest to highest. First the coefficients in each column of the table were ranked. A rank of 1 represented the smallest multiple regression coefficient for each level and 8 represented the largest multiple regression coefficient for Levels 12 and 13. For Level 11 the highest rank was 7 since no

TABLE III.19 Multiple Regression Coefficients Corresponding to Each Rule, When the Upper Bound or the Midpoint of the Plausibility Range Is a Function of the Specific Rules Pertaining to the Informative Distractor, for Each of Levels 11, 12 and 13 of Vocabulary Subtest.

Rule	<u>No. Dstrs.</u>			<u>Upper Bound</u>			<u>Midpoint</u>			<u>Peak</u>			Avg. Rank Within Column
	<u>Level</u>			<u>Level</u>			<u>Level</u>			<u>Level</u>			
	11	12	13	11	12	13	11	12	13	11	12	13	
1	22	21	19	0.024	-0.143	0.324	0.017	0.162	0.194	-0.108	-0.625	0.068	3.6
2	0	2	2	---	0.943	1.173	---	0.643	0.624	---	-0.463	0.972	6.5
3	13	11	11	0.548	-0.439	-0.058	0.229	-0.328	-0.071	0.262	-0.913	-0.138	3.1
4	13	11	10	-0.478	0.022	0.572	-0.196	0.222	0.426	-0.101	0.683	0.988	4.3
5	11	10	11	0.086	-0.375	0.091	0.024	-0.196	0.026	0.100	-0.655	0.143	3.7
6	5	4	5	-0.069	0.422	0.528	-0.051	0.734	0.585	-0.199	0.548	1.389	4.9
7	3	2	3	0.558	1.071	2.528	0.285	0.728	1.322	-0.059	1.137	2.271	7.3
8	3	2	3	-0.475	-1.565	0.002	-0.282	-0.919	0.005	0.129	-2.815	0.189	2.2

informative distractors were found corresponding to rule number 2 at this level. These ranks were then averaged for each row in the table to yield an overall average rank for each rule. Although the different types of ability measures yield somewhat different results, some general tendencies can be seen. Most obvious is that rule 7 (opposite in meaning) consistently ranks high, or is apparently frequently followed by relatively higher ability examinees. It has the highest overall mean rank, although rule number 2 (meaning taken out of context) is not far below. At the other extreme, rule number 8 (familiarity) is consistently ranked low and has the lowest overall mean rank, with rule 3 (logical in context) ranking next lowest.

In order to further investigate the possibility that low-ability examinees tend to select vocabulary alternatives that are composed of relatively familiar words, a system of assigning a measure of familiarity to each alternative was sought that did not depend strictly upon the author's judgment. While typical sight vocabulary word lists include a relatively small number of words and reflect specific reading programs, Hillerich (1978) compiled a vocabulary list of almost 9,000 words based upon unrestricted writing samples of elementary students. The writing vocabulary list was obtained from weekly writing samples of almost 3,000 students in grades two through six in a suburban Illinois community. Samples were collected over the course of a school year and a frequency count was made of all the words written, by grade and in total. Excluding proper nouns, 8,925 different words were found among the 379,694 total words tabulated.

As a measure of the familiarity of a specific word, the

relative frequency of the word among all written words tabulated by Hillerich was computed. Thus the total frequency for the word across grades two through six was divided by 379,694. For alternatives consisting of one word the relative frequency of the word was used as a measure of the familiarity of the alternative. A decision was necessary with regard to alternatives composed of two or more words. If the average relative frequency or the relative frequency of the most common word was taken as a familiarity measure, the author felt that the very high frequency of use for a common article, conjunction or preposition might give a distorted picture of the familiarity of a particular phrase. For instance, five words--"I," "and," "the," "a," and "to"--account for about 18 percent of all words tabulated in Hillerich's study. A three word phrase composed of two very unfamiliar words and one of the five common words listed above might yield a relatively high measure of familiarity if the average frequency or frequency of only the most familiar word were used. However, this phrase would likely appear quite unfamiliar to the typical examinee. For these reasons, the relative frequency of the least frequently used word in the phrase was the measure taken in this study.

Table III.20 shows the percent usage (relative frequency multiplied by 100) for the least frequently used word in each alternative of each Vocabulary item. The alternatives are presented in descending order with regard to percent usage, with dashes drawn between alternatives of equal value. An asterisk is placed after each alternative that was previously identified as an informative

TABLE III.20 Percent Usage, As Determined from Hillerich's (1978) Writing Vocabulary, of the Least Frequently Used Word in Each Response Alternative, for Each of 82 Items of Levels 11, 12 and 13 of Vocabulary Subtest. Informative Distractors Are Denoted by an Asterisk.

Item	Alternative				Percent Usage			
24	C	A	B	D	0.3282	0.1412	0.1188	0.0756
25	A	C	B	D	0.0061	0.0055	0.0026	0.0000
26	B	D	A	C*	0.0271	0.0169	0.0161	0.0108
27	B	C*	A	D	0.5760	0.1786	0.1309	0.0764
28	B*	C	A	D	0.0108	0.0047	0.0016	0.0008
29	D	C	B	A	0.1156	0.0598	0.0100	0.0005
30	A	C	B*	D	0.1309	0.0485	0.0108	0.0008
31	D	B	C*	A	0.0269	0.0032	0.0005	0.0000
32	B	C---	D	A	0.1222	0.0011	0.0011	0.0000
33	D	B	A	C	0.0205	0.0137	0.0126	0.0097
34	B	D*	A---	C	0.0029	0.0013	0.0008	0.0008
35	C*	B	D	A	0.0774	0.0016	0.0003	0.0000
36	C	D*	B	A	0.1001	0.0598	0.0100	0.0000
37	B	A	C*	D*	0.0124	0.0058	0.0005	0.0000
38	C	D	A---	B*	0.0248	0.0097	0.0000	0.0000
39	B	A	C	D*	0.1298	0.1188	0.0532	0.0045
40	C*	B	A	D	0.0532	0.0287	0.0042	0.0013
41	D	C*	B	A	0.0219	0.0087	0.0076	0.0013
42	B	C*	A	D	0.0090	0.0013	0.0005	0.0003
43	D	C	A---	B	0.0203	0.0047	0.0000	0.0000
44	C---	D*	A---	B	0.0011	0.0011	0.0000	0.0000
45	C	A---	B---	D*	0.0011	0.0000	0.0000	0.0000
46	D	C*	B	A	0.0108	0.0011	0.0003	0.0000
47	C	B	A---	D*	0.0026	0.0008	0.0000	0.0000
48	D	B	A	C*	0.0345	0.0029	0.0003	0.0000
49	B	D	C*	A	0.0116	0.0111	0.0076	0.0008
50	B*	A	D	C*	0.0974	0.0356	0.0121	0.0016
51	C	B*	D	A	0.0248	0.0103	0.0021	0.0000
52	A	C	B*	D	0.0100	0.0097	0.0042	0.0000
53	A	C	B---	D*	0.0029	0.0016	0.0000	0.0000
54	A	B	D	C*	0.0216	0.0047	0.0008	0.0003
55	B*	A---	C---	D	0.0063	0.0000	0.0000	0.0000
56	A	B*	D	C*	0.0529	0.0514	0.0082	0.0018
57	D	C	A---	B	0.0029	0.0005	0.0000	0.0000
58	A	D	B	C*	0.1035	0.0208	0.0171	0.0016
59	B	C	D	A	0.1838	0.1140	0.0972	0.0303
60	B	A	D	C*	0.0153	0.0108	0.0021	0.0000
61	D	A	B*	C	0.0453	0.0171	0.0016	0.0011
62	B	A	C---	D*	0.0011	0.0008	0.0000	0.0000
63	A	C	B---	D	0.0137	0.0024	0.0000	0.0000

TABLE III.20 (Continued)

Item	Alternative				Percent Usage			
64	C	B*	D	A	0.0358	0.0298	0.0016	0.0005
65	D*	A---	C*	B	0.0063	0.0003	0.0003	0.0000
66	D	B	C	A	0.0029	0.0024	0.0021	0.0000
67	A	C	D*	B	0.0013	0.0008	0.0005	0.0000
68	B*	D	A---	C	0.0047	0.0003	0.0000	0.0000
69	D	B	A---	C	0.0263	0.0013	0.0000	0.0000
70	A	B	D	C	0.0108	0.0082	0.0003	0.0000
71	C*	B	A	D	0.0263	0.0013	0.0003	0.0000
72	A---	B---	C---	D*	0.0000	0.0000	0.0000	0.0000
73	C	D*	A	B	0.5315	0.4293	0.0005	0.0003
74	A	B	C---	D	0.0011	0.0008	0.0000	0.0000
75	B	D	C	A	0.0543	0.0147	0.0068	0.0000
76	B	A	C---	D	0.0087	0.0013	0.0000	0.0000
77	C	B	A---	D*	0.0018	0.0011	0.0000	0.0000
78	C	D	A	B	0.0013	0.0008	0.0003	0.0000
79	B*	D	C	A	0.1272	0.0163	0.0037	0.0008
80	C	B*	A---	D*	0.0058	0.0021	0.0000	0.0000
81	C	D	A---	B*	0.0137	0.0050	0.0000	0.0000
82	D	C	A---	B	0.0769	0.0008	0.0000	0.0000
83	D	B	A---	C*	0.0658	0.0198	0.0005	0.0005
84	C	B	D*	A	0.2547	0.1383	0.0242	0.0008
85	B	A	D*	C*	0.0079	0.0011	0.0005	0.0000
86	A	B	C---	D*	0.0163	0.0032	0.0000	0.0000
87	D	C*	A---	B*	0.0532	0.0274	0.0005	0.0005
88	A	D	C	B	0.1412	0.0374	0.0037	0.0005
89	D	A	B*	C	0.0208	0.0161	0.0108	0.0034
90	C*	B*	D	A	0.0774	0.0229	0.0018	0.0000
91	B	C	D	A	0.0097	0.0029	0.0003	0.0000
92	D*	B	C	A	0.0045	0.0013	0.0003	0.0000
93	D	A	B---	C*	0.0055	0.0013	0.0005	0.0005
94	A	D	B	C	0.0292	0.0018	0.0011	0.0000
95	D*	C	B	A	0.1338	0.0055	0.0040	0.0013
96	B	A	D	C*	0.5315	0.0382	0.0050	0.0026
97	B*	C	D	A	0.0013	0.0005	0.0003	0.0000
98	A	D	C*	B*	0.0184	0.0011	0.0005	0.0000
99	A	D*	C	B	0.0353	0.0068	0.0011	0.0008
100	D	B	C	A	0.1180	0.0211	0.0008	0.0005
101	A---	B*	D	C	0.0024	0.0024	0.0013	0.0000
102	C	B	D*	A	0.1148	0.0182	0.0111	0.0055
103	D*	A	C	B	0.1973	0.0100	0.0095	0.0016
104	D	B*	C	A	0.0018	0.0011	0.0003	0.0000
105	B*	C	D	A	0.0374	0.0074	0.0050	0.0008

distractor. Of the six informative distractors that the author had identified with rule 8 (familiarity), in three cases the informative distractor had the highest percent usage among the three distractors. In two of the three other cases, the informative distractor had the second highest percent usage. A Pearson chi-square test of association based upon all Vocabulary items did not indicate a significant relationship between the rank of the distractor with respect to usage and whether or not the distractor was informative. Investigation of informative distractors appealing to low-ability examinees did not reveal a disproportionate number ranking either high or low in terms of usage.

The correlation between the percent usage of the key word, or stimulus word in the stem, and the item difficulty parameter was computed for each level of the subtest. A negative correlation was anticipated since the higher the usage--and thus, hopefully, the greater the familiarity--the easier the item should be. Of course, the familiarity of the alternatives should also be an important factor in the item difficulty. Negative correlations were found for two of the three levels, but none was statistically significant. The correlations were -0.140 , -0.286 , and 0.068 for Levels 11 through 13, respectively.

Item data was reviewed with respect to the part of speech (e.g., noun, verb, adjective, adverb, preposition) represented by the key word and the response alternatives. For all but two Vocabulary items, items 33 and 43, all alternatives represented the same part of speech as the key word. For the two exceptions, neither of which had

an informative distractor, there was one alternative that differed from the other alternatives and the key word. The number of items having informative distractors, by part of speech, is presented in Table III.21. A Pearson chi-square test of association did not show a relationship between the part of speech of the distractor and whether it was informative. The number of items having multiple informative distractors is also shown in Table III.21. Although there are too few cases for meaningful statistical analysis, it is interesting that six of the nine items having multiple informative distractors had key words and alternatives that were verbs.

Table III.21 also presents the mean item difficulty parameter (normal ogive model) for each level, by part of speech. There does not appear to be any consistent trend across subtest level relating item difficulty to part of speech.

TABLE III.21 The Number of Items for Which the Key Word Represents a Particular Part of Speech, the Number and Percentage Having Informative Distractors, and the Mean Item Difficulty Parameter for Each of Levels 11, 12 and 13, by Part of Speech.

Part of Speech	Number of Items	No. Items with Info. Distractor	Pct. Items with Info. Distractor	No. Items with Mult. Info. Dist.	<u>Mean Item Difficulty</u>		
					Lev.11	Lev.12	Lev.13
Adjective	22	15	68.2	1	-0.202	-0.398	-0.491
Adverb	8	4	50.0	0	-0.947	-0.254	-0.341
Noun	25	20	80.0	2	-0.345	-0.631	-0.400
Preposition	2	2	100.0	0	-0.260	---	---
Verb	25	20	80.0	6	-0.134	-0.270	-0.543

CHAPTER IV

CONCLUSIONS

There are several conclusions that can be drawn from this study. Some of these conclusions follow from the content analysis, some from the latent trait analysis, and some are drawn from the comparison of the results of these two different approaches. The major conclusions, together with some suggestions for future investigations, are as follow:

1. There was generally much agreement among faculty members, teachers and the author, based upon the item content, as to which distractors were more likely to attract higher ability examinees and which would attract lower ability examinees. For many of these items, the same or similar reasons were given by these evaluators for their predictions, and several distinct reasons were identified that pertained to more than one item distractor in the subtest investigated, the Vocabulary subtest of the Iowa Tests. In some cases, evaluators found it difficult to express a specific reason for making a prediction, or even to make a prediction at all, and some of these selections appear to be more or less intuitive.

2. Most of the items investigated in this study had informative distractors. There was generally only one distractor per item that was observed to be informative, although a small percentage of items yielded two informative distractors. It is possible that additional distractors would have been found to be informative had there been a broader range in the ability distribution of the

examinees used in this study. For example, certain distractors might have appealed highly to examinees of lower ability than were included in this data. Several items in the Vocabulary subtest occurred in two or more levels and thus were administered to examinees of a wider range of ability. In many of these cases the item was relatively easy for at least one of these levels, and as has been noted by Samejima (1984a), informative distractors are not likely to be observed when the item is easy for the group and thus usually answered correctly. Indeed it was found that the mean item difficulty of items having informative distractors was significantly higher than that of items not having informative distractors. However, it was observed (e.g., item 83) that informative distractors can sometimes be better detected when administering an item to a higher ability group. It appears that a relatively high level of ability is necessary in order for certain distractors to be plausible to examinees, which is consistent with the concept of the informative distractor. Since most of the items tend to follow the Informative Distractor Model, Samejima's New Family of Models (Samejima, 1979) could in the future be applied to these items, in order to better use the available information and improve efficiency in ability estimation.

3. The normal ogive model generally was validated by the results. For the vast majority of items the non-parametric estimated item characteristic function fit closely to the estimated normal ogive item characteristic function. In addition, the chi-square test of the assumption of a joint bivariate normal distribution between each pair of response tendencies identified only a few items that gave

significant evidence of violating this assumption.

4. For Vocabulary subtest items that appeared in two or more levels, the results were generally consistent between levels, with the same distractors usually being seen to be informative. Although the plausibility functions for the different levels were on different scales and were thus not directly comparable, the same general tendencies were typically observed in the results at different levels for the same item. For many items it was seen that essentially the same patterns emerged for different levels when the distractors were ranked in terms of plausibility, over the range of ability. If the latent traits corresponding to the different levels were in the future adjusted to the same scale, it would become even clearer as to which distractors are informative and over what ranges of ability, particularly for those items in which the results are inconsistent between two levels. This would also make it easier to identify which specific types of errors tend to be made by relatively high level examinees, and which tend to be made by low ability examinees.

5. Some general tendencies were observed in the relationship between the rules behind alternative choice and examinee ability. Although considerable individual variability existed, individuals of similar ability often selected distractors reflecting the same rules or reasons. Distractors reflecting selection based upon some partial understanding (or misunderstanding) of the meaning, such as selecting the opposite meaning or taking the meaning out of context, tended to attract examinees of moderate ability. On the other hand, very low ability examinees appeared more likely to choose a distractor for

reasons less related to meaning, such as the familiarity of the word(s) or that it was logical in context.

6. It was found that informative distractors can frequently be correctly predicted based strictly upon their content. The predictions by the author tended to be somewhat closer to the observed results than the overall faculty/teacher rankings, but this may be explained at least in part by the increased time spent by the author in reviewing each item. In addition, items selected to be reviewed by faculty/teacher evaluators were chosen partly because they were considered to be interesting in content and typically had multiple distractors that were relatively highly selected. Therefore these items were likely more difficult to predict, on the whole, than were the other items in the subtests. It was also observed that for many items the distractors behaved quite differently than was expected either by the author or by the other evaluators.

7. In order to better determine which rules or reasons are most commonly followed by examinees, in the future a set of items could be constructed with relatively equal numbers of distractors reflecting the various types of rules identified. The distractors investigated in this study were likely to have been constructed in a more intuitive manner, rather than with regard to the specific types of errors that they indicate, and therefore they do not appear to occur in any systematic manner. In addition, the types of strategies or rules that examinees follow can be further investigated in the future by interviewing examinees with regard to their reasons for their answers, as has been done to some extent by other researchers

(e.g., Powell, 1968; Tatsuoka, 1981; Norris and King, 1984). Although Tatsuoka, Birenbaum and Arnold (1989) have shown that examinees do not always consistently follow rules during test sessions, there do appear to be general tendencies in the rules used by subgroups of examinees of similar ability.

8. There were studies cited in Chapter I of this dissertation in which distractors were selected for new multiple-choice items by the popular method of presenting item stems in a free-response situation and then choosing as distractors those incorrect responses that were most frequent (e.g., Owens, Hanna and Coppedge, 1970). However, none of these studies attempted to select distractors that would appeal to different levels of ability. Without considering ability level when selecting distractors, and looking only at overall frequencies, distractors appealing to low ability examinees are likely to be chosen, since errors will be made most often by these examinees. Consideration of ability level could have been taken on a rough but simple basis, for example, by dividing examinees into different subgroups based upon their test score and then looking within these subgroups at the frequency distributions for the incorrect responses. The estimation of plausibility functions could be made after the distractors have been selected to determine if indeed these response categories are attracting different levels of examinees when the items appear in a multiple-choice format.

9. The semantic closeness or similarity in meaning between a distractor and the key word appears to be positively related to the likelihood that a Vocabulary item will be informative. Words that are

opposite in meaning to the key word can also serve as informative distractors, but they are less likely to do so.

10. Context clues seem to be quite important for examinees in selecting an alternative for a Vocabulary item. Distractors that are logical and readable in the context presented are frequently informative, whereas it appears less likely that a distractor is selected because a part of a word or phrase is taken out of context.

11. When reasons involving semantic closeness and context clues do not attract an examinee, familiarity of wording and similarity in spelling between distractor and key word appear to be important factors in alternative selection for Vocabulary items.

12. The plausibility of a distractor appears to be increased when multiple reasons or rules can be identified that would lead to selection of that distractor. In fact, most of the informative distractors identified on the Vocabulary subtest had multiple reasons identified for them. However, having multiple reasons such as these can reduce the diagnostic information available, since it becomes less clear as to which reason the examinee's selection is based upon.

13. The results obtained by the frequency ratio procedure indicate that this method might be useful in identifying distractors that are highly informative (those that are chosen with a rather high conditional probability, given ability, over a certain range of ability), but that it is not so effective in detecting subtle differences in the plausibility functions which distinguish many informative distractors from non-informative distractors. Distractors identified as being informative based upon their frequency ratios

usually turned out to be informative when the results of the plausibility functions were observed, but many informative distractors were missed by the frequency ratio method, and were instead classified as being non-informative or as questionable. Frequently the jagged curves estimating the plausibility functions that are produced by the frequency ratio method are erratic and do not indicate a consistent tendency. For these reasons and the fact that the maximum likelihood scale is not a satisfactory substitute for the ability scale (Samejima, 1981), the frequency ratio procedure should only be used as a screening method in specific situations.

Summary and Implications

It was stated in the first section of Chapter I of this dissertation that the three main objectives of this research were to: (1) identify rules for the prediction of informative distractors based upon their content, (2) empirically determine the informative distractors by applying methods and approaches developed by Samejima (1981), and (3) to better describe the characteristics of informative distractors by integrating the results of the content analysis and the empirical analysis. The methods by which these goals were accomplished were described in Chapter II of this dissertation. Earlier in this chapter conclusions and suggestions for further research were presented. This study described some of the characteristics of informative distractors that were found with Vocabulary achievement items. It also raised several other questions that can only be answered through future research (e.g., see conclusions 7 and 8 earlier in this chapter). Although this study is

seen only as a starting point in the investigation of informative distractors, there are certain findings that have practical implications for test constructors and measurement specialists, as will be discussed in the rest of this section.

It may be stated from this and previous studies (e.g., Powell, 1968; Samejima, 1980; Levine and Drasgow, 1983) that not all distractors are created equal. Some tend to be more plausible than others and are more frequently selected by examinees. The selection of distractors is clearly not a random process, as the knowledge or random guessing principle states, yet the three-parameter logistic model, which is based in theory upon this principle, continues to be a very popular mathematical model with practitioners in the field of measurement. Practitioners and researchers should be aware that there are alternatives to the three-parameter logistic model that are not based upon the knowledge or random guessing principle and should make an effort to empirically test the model or models which they use. Furthermore, as it is seen from this research that a relatively high percentage of items have informative distractors, estimation can be made more efficiently using mathematical models that have been developed based upon the concept of informative distractors, such as Samejima's New Family of Models (Samejima, 1979).

This study showed a fairly high degree of agreement among evaluators in the predictions of informative distractors. There was only modest success in the accuracy of these predictions, however. While this success rate would be expected to improve with feedback from this type of study, item writers should take note of this finding

since their selection of distractors is frequently quite intuitive. What may seem to be a plausible distractor to an item writer or group of writers may not actually be attractive to the population for which the item is intended. There is no substitute for empirical analysis or field testing of the items.

Although the distractor selection procedure of administering item stems in a free-response situation and then choosing the most frequent incorrect responses as distractors has an empirical basis, consideration should be given to the ability levels of the examinees, as was discussed in conclusion 8 of this chapter. As was suggested in conclusion 8, the examinees could be divided into subgroups based upon their test score. For example, say that three distractors were needed for a particular item. The group of examinees could be divided into three subgroups (high, average and low) based upon the number of items correctly answered. For each of the three subgroups, the incorrect response most commonly given in the free response situation could be used as one distractor, rather than simply choosing the three most common responses by the entire group of examinees. This procedure would be more likely to give distractors appealing to different levels of ability.

For an item that contains a distractor that attracts examinees at the upper end of the latent trait (e.g., Vocabulary item 83), it seems that a high degree of ability or achievement may be required in order for the examinee to comprehend that the distractor is plausible. There may be times when it is not desirable to construct distractors that are especially attractive to high ability examinees, such as on a

minimal competency or proficiency test. If the purpose of the test is only to discriminate between examinees at the lower end of the ability scale, high level distractors should probably be avoided. If such distractors are used, some examinees may in a sense be penalized for knowing too much, since they may be drawn toward these distractors. Excluding such distractors would likely increase the probability that "proficient" examinees would correctly answer the item. The methods developed by Samejima (1981) could be used in this type of situation to estimate the plausibility functions in order to identify items having distractors that might be too appealing to examinees of high ability.

While some of the previous suggestions have stressed the importance of empirical analysis of response data, some of the findings stated previously in this chapter have implications for the initial construction of the items. These findings tend to be dependent upon the content of the test. For Vocabulary achievement items, similarity in meaning, context cues, familiarity, and similarity in spelling between the distractor and the key word in the stem appear to be factors to consider in writing informative distractors (see conclusions 9, 10 and 11). While considerations such as these do not guarantee the creation of informative distractors, they make it more likely that empirical analysis will reveal informative or plausible distractors, and will thus make the necessity of revising the items less likely.

Also relevant to item construction is the finding that examinees of similar ability may follow the same types of rules in

selecting an item alternative when the correct response is not positively known. If certain rules can be related to specific ability ranges, as it appears might be the case (see conclusion 5), then distractors appealing to different ability groups could be constructed with minimal field testing. The likelihood of writing items within a predetermined difficulty range would also be enhanced.

Items being considered for operational use in major testing programs are generally subjected to various forms of psychometric analysis. An item might be rejected for any of a number of reasons, such as poor discriminating power or bias towards a particular subgroup. Typically the item is returned to item writers for revision without any comment or recommendations regarding content. Unless a grammatical or logical error is detected, any type of revision will be mostly guesswork. As the link between content and plausibility to examinees is better understood, item construction and revision should become a more efficient process.

Linking rules or error types to ability level also has diagnostic implications. If examinees within a certain ability range are known to typically make the same types of mistakes, the ability estimate of the examinee can be used as a starting point in determining future diagnostic testing, with emphasis on these specific types of errors.

A number of reasons for further investigation of informative distractors have been discussed, such as increased estimation efficiency, practical use to item writers, and potential diagnostic benefits. In addition, better understanding of distractors--including

the relationship of content and format to probability of selection-- could be helpful in the application of judgmental standard setting procedures. These procedures, such as the Angoff (1971) and Nedelsky (1954b) methods, are frequently used to set standards or cut scores on competency examinations. Each item is reviewed by content experts who must predict the probability of success (Angoff method) or the plausibility of each distractor (Nedelsky method) if the item were administered to a minimally competent examinee. Research has shown that different standard setting procedures yield different standards and that there is often a discrepancy between predicted and empirical item difficulties (Smith and Smith, 1988). Considering the high stakes often associated with such competency tests (such as graduation, certification, promotion, employment selection, etc.), it seems clear that more research is necessary in this area. This study has given further indication that predicted and empirical response data do not always agree. The validity of judgmental standards will increase with the identification of rules that relate item content and format to the conditional probability of success, given ability.

There appears to be increased interest in the distractor in the psychometric literature. In a recent special issue of the Journal of Educational Measurement devoted to the test item, three articles emphasized the importance of utilizing distractor data in the item analysis process (Green et al., 1989; Thissen et al., 1989; Wainer, 1989). All three articles advocated the analysis of the functional relationship between distractor choice and ability level of the examinee. As stated by Thissen et al. (1989):

Analysis of the trace lines associated with each alternative on a multiple-choice item may lead to the accumulation of sufficient knowledge about the properties of distractors to allow alternatives to be designed to meet the specific needs of the test. Distractors could be designed to be equivalent if the test is to be scored using a binary (right/wrong) model or they could be designed to be informative if a multiple-category scoring system is contemplated. . . .

.
Item response models for multiple-choice items have come of age: They fit the data. They produce excellent item analysis, but they are complex and are best represented graphically--that breaks with the tradition of item analysis using a few numerical summaries, but this idea is entirely compatible with computerized item analysis system described by Wainer (1989) as "the future of item analysis."

LIST OF REFERENCES

LIST OF REFERENCES

- Abu-Sayf, F. K. Recent developments in the scoring of multiple-choice items. Educational Review, 1979, 31, 269-279.
- Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike, Educational Measurement. Washington, D. C.: American Council on Education, 1971.
- Angoff, W. H. Does guessing really help? Journal of Educational Measurement, 1989, 26, 323-336.
- Birnbaum, A. 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. Chapters 17-20. Reading, Mass.: Addison-Wesley, 1968.
- Birenbaum, M. and K. K. Tatsuoka. The use of information from wrong responses in measuring students' achievement. Office of Naval Research, Research Report 80-1, 1980.
- Birenbaum, M. and K. K. Tatsuoka. Effect of different instructional methods on error types and the underlying dimensionality of the test, Part I. Office of Naval Research, Research Report 81-3, 1981.
- Board, C. B. and D. R. Whitney. The effect of selected poor item-writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 9, 225-233, 1972.
- Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51, 1972.
- Brown, J. S. and R. R. Burton. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 1978, 2, 155-192.
- Chase, C. Relative length of options and response set in multiple choice items. Journal of Educational Measurement, 1964, 1, 38.
- Coombs, C. H., J. E. Milholland and F. B. Womer. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.
- Cross, L. H. and R. B. Frary. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

- Davis, F. B. and G. Fifer. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.
- Diamond, J. and W. Evans. The correction for guessing. Review of Educational Research, 1973, 43 181-191.
- Dressel, P. L. and P. Schmid. Some modifications of the multiple-choice item. Educational and Psychological Measurement, 1953, 13, 574-595.
- Ebel, R. L. Blind guessing on objective achievement tests. Journal of Educational Measurement, 1968, 5, 321-325.
- Evans, W. Testwiseness: An examination of cue-using strategies. Journal of Experimental Education, 1984, 52, 141-144.
- Glaser, R. The future of testing: A research agenda for cognitive psychology and psychometrics (Technical Report no. 3). Pittsburgh, PA: Learning Research and Development Center, 1981.
- Goldman, S. Information theory. Prentice Hall, 1953.
- Gorham, D. R. Proverbs tests. Missoula, Montana: Psychological Test Specialists, 1956.
- Green, B. F., C. R. Crone and V. G. Folk. A method for studying differential distractor functioning. Journal of Educational Measurement, 1989, 26, 147-160.
- Green, K. Effects of item characteristics on multiple-choice item difficulty. Educational and Psychological Measurement, 1984, 44, 551-561.
- Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst, The prediction of personal adjustment. New York: Social Science Research Council, 1941.
- Hakstian, A. R. and W. Kansup. A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures. Journal of Educational Measurement, 1975, 12, 231-239.
- Haladyna, T. M. and S. Downing. An analysis of knowledge about multiple-choice test item writing. Paper presented at the 1984 American Psychological Association Annual Meeting, Toronto, Canada, 1984.

- Haladyna, T. M. and S. Downing. A quantitative review of research on multiple-choice item writing. Paper presented at the 1985 AERA Annual Meeting, Chicago, Illinois, 1985.
- Hambleton, R. K. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48.
- Harris, D. Comparisons of 1-, 2-, and 3-parameter IRT models. Educational Measurement Issues and Practice, 1989, 8, 35-41.
- Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. Journal of Educational Measurement, 1971, 8, 291-296.
- Hieronymus, A. N. and E. F. Lindquist. Iowa test of basic skills teacher's manual. Iowa City, Iowa: University of Iowa, 1974.
- Hillerich, R. L. A writing vocabulary of elementary children. Springfield, Ill.: Charles C. Thomas, 1978.
- Horst, P. A generalized expression for the reliability of measures. Psychometrika, 1949, 14, 21-23.
- Jones, P. D. and G. G. Kaufman. The differential formation of response sets by specific determiners. Educational and Psychological Measurement, 1975, 35, 821-833.
- Kansup, W. and A. R. Hakstian. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring Procedures. Journal of Educational Measurement, 1975, 12, 219-230.
- Lawley, D. N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, 273-287.
- Lawley, D. N. The factorial analysis of multiple item tests. Proceedings of the Royal Society of Edinburgh, 1944, 62A, 74-82.
- Levine, M. V. and F. Drasgow. The relation between incorrect option choice and estimated ability. Educational and Psychological Measurement, 1983, 43, 675-685.
- Lord, F. M. A theory of test scores. Psychometric Monograph No. 7, Psychometric Society, 1952.
- Lord, F. M. Formula scoring and number-right scoring. Journal of Educational Measurement, 1975, 12, 7-11.
- Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N. J.: Lawrence Erlbaum, 1980.

- Lord, F. M. and M. R. Novick. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Mayberry, P. W. and J. C. Ory. Application of CAT misconception analysis in mathematical placement and proficiency decisions. Paper presented at the 1985 AERA Annual Meeting, Chicago, Illinois, 1985.
- McMorris, R. F., J. A. Brown, G. W. Snyder and R. M. Pruzek. Effects of violating item construction principles. Journal of Educational Measurement, 1972, 9, 287-295.
- Nedelsky, L. Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 1954a, 14, 459-472.
- Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954b, 14, 3-19.
- Norris, S. P. and R. King. The design of a critical thinking test on appraising observations. St. John's: Memorial University of Newfoundland: Institute for Educational Research and Development, 1984.
- Owens, R. E., G. S. Hanna and F. L. Coppedge. Comparison of multiple-choice tests using different types of distractor selection techniques. Journal of Educational Measurement, 1970, 7, 87-89.
- Patnaik, D. and R. E. Traub. Differential weighting by judged degree of correctness. Journal of Educational Measurement, 1973, 10, 281-286.
- Peterson, J. P. The Iowa testing programs. Iowa City: University of Iowa Press, 1983.
- Powell, J. C. The interpretation of wrong answers from a multiple-choice test. Educational and Psychological Measurement, 1968, 28, 403-412.
- Powell, J. C. and A. G. Isbister. A comparison between right and wrong answers on a multiple-choice test. Educational and Psychological Measurement, 1974, 34, 499-504.
- Powell, J. C. and H. Miki. Answering anomalies: How serious? Paper presented at the 1985 Psychometric Society Annual Meeting, Nashville, TN, 1985.
- Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph No. 17. Psychometric Society, 1969.

- Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39, 111-121.
- Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.
- Samejima, F. A new family of models for the multiple-choice item. Office of Naval Research, Research Report 79-4, 1979.
- Samejima, F. Research on the multiple-choice test item in Japan: Toward the validation of mathematical models. Office of Naval Research-Tokyo, Scientific Monograph 3, April, 1980.
- Samejima, F. Final Report: Efficient methods of estimating the operating characteristics of item response categories and challenge to a new model for the multiple-choice item. Office of Naval Research, Final Report, 1981.
- Samejima, F. Specification of the information provided by distractors of the multiple-choice test item and efficient ability estimation. Paper presented at the 1984 AERA Annual Meeting, New Orleans, Louisiana, 1984a.
- Samejima, F. Plausibility functions of Iowa Vocabulary test items estimated by the Simple Sum Procedure of the Conditional P.D.F. Approach. Office of Naval Research, Research Report 79-2, 1984b.
- Samejima, F. and P. S. Changas. How small the number of test items can be for the basis of estimating the operating characteristics of the discrete responses to unknown test items. Office of Naval Research, Research Report 81-3, 1981.
- Samejima, F. and R. L. Trestman. Analysis of Iowa Data I: Initial study and findings. Office of Naval Research, Research Report 80-1, 1980.
- Sarnacki, R. E. An examination of testwiseness in the cognitive domain. Review of Educational Research, 1979, 49, 252-279.
- Sato, T. Engineering techniques for analyzing educational data, No. 5. Central Research Institute, Nippon Electric Company, Ltd., 1977.
- Shiba, S. Construction of a scale for acquisition of word meanings. Bulletin of Faculty of Education, University of Tokyo, 17, 1978.
- Shuford, E. H., A. Albert and H. E. Massengill. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

- Smith, J. K. Converging on correct answers: A peculiarity of multiple-choice items. Journal of Educational Measurement, 1982, 19, 211-220.
- Smith, R. L. and J. K. Smith. Differential use of item information by judges using Angoff and Nedelsky procedures. Journal of Educational Measurement, 1988, 25, 259-274.
- Strang, H. R. The effects of technical and unfamiliar options on guessing on multiple-choice test items. Journal of Educational Measurement, 1977, 14, 253-260.
- Tatsuoka, K. K. An approach to assessing the seriousness of error types and predictability of future performance. Office of Naval Research, Research Report 81-1, 1981.
- Tatsuoka, K. K., M. Birenbaum and J. Arnold. On the stability of students' rules of operation for solving arithmetic problems. Journal of Educational Measurement, 1989, 26, 351-361.
- Tatsuoka, K. K., M. Birenbaum, M. M. Tatsuoka and R. Baillie. Psychometric approach to error analysis on response patterns of achievement tests. Office of Naval Research, Research Report 80-3, 1980.
- Thissen, D. M. Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 1976, 13, 201-214.
- Thissen, D., L. Steinberg and A. R. Fitzpatrick. Multiple-choice models: The distractors are also part of the item. Journal of Educational Measurement, 1989, 26, 161-176.
- Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.
- Wainer, H. The future of item analysis. Journal of Educational Measurement, 1989, 26, 191-208.
- Wang, M. W. and J. C. Stanley. Differential weighting: A review of methods and empirical studies. Review of Educational Research, 1970, 40, 663-705.
- Weiten, W. Violation of selected item construction principles in educational measurement. Journal of Experimental Education, 1984, 52, 174-178.

APPENDICES

APPENDIX A

TABLE A.1 Mean of Maximum Likelihood Estimate of Ability, $\hat{\theta}$, and Number of Examinees in Each of 30 Subgroups Used in Obtaining Frequency Ratios to Approximate Plausibility Functions.

Subgroup	Number of Examinees	Mean
1	79	-1.978
2	79	-1.522
3	78	-1.352
4	79	-1.228
5	79	-1.117
6	78	-1.001
7	79	-0.894
8	79	-0.797
9	78	-0.700
10	79	-0.586
11	79	-0.486
12	78	-0.376
13	79	-0.290
14	79	-0.181
15	78	-0.096
16	79	-0.011
17	79	0.087
18	78	0.184
19	79	0.291
20	79	0.393
21	78	0.491
22	79	0.615
23	79	0.757
24	78	0.902
25	79	1.049
26	79	1.233
27	78	1.412
28	79	1.647
29	79	1.966
30	79	2.962

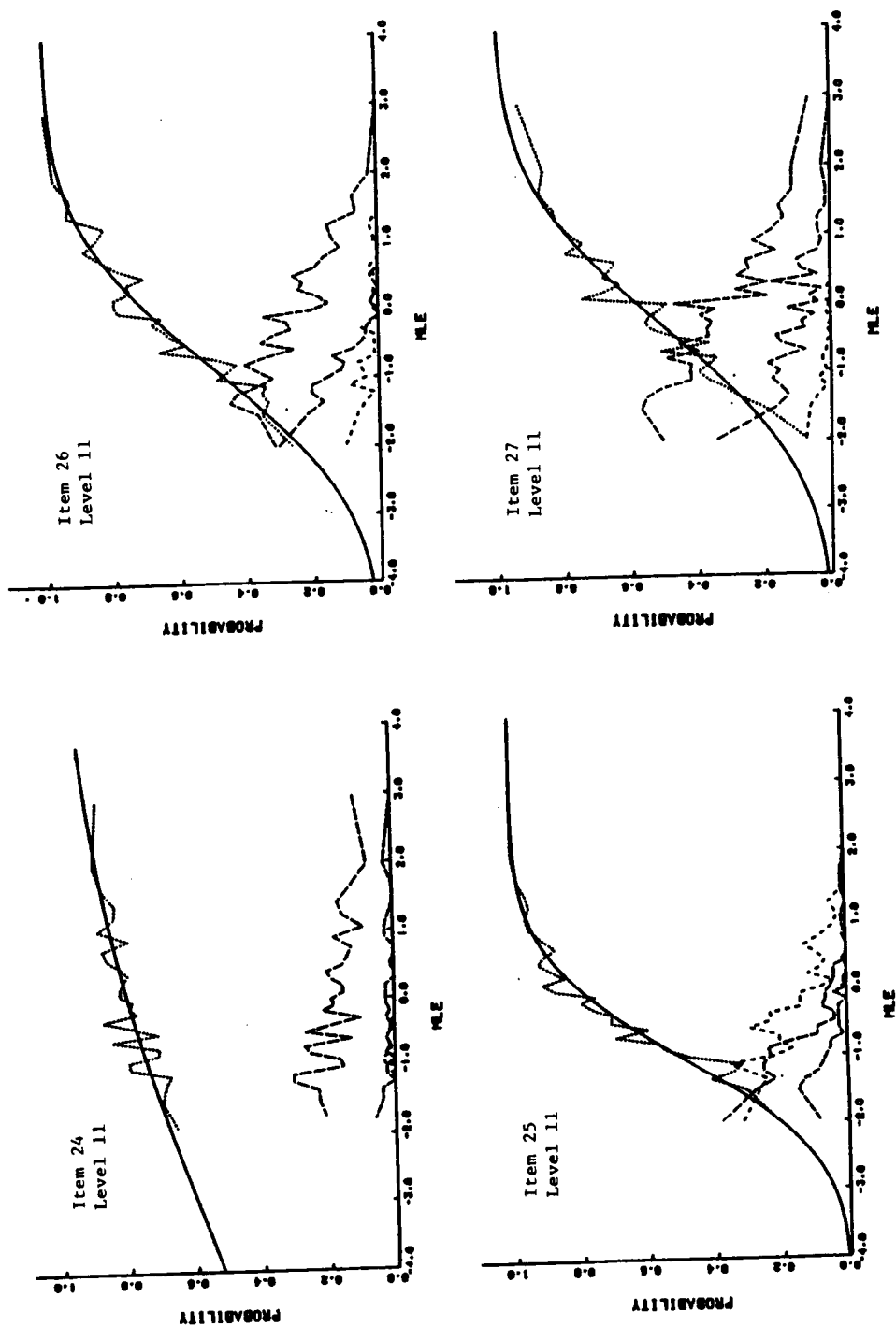


FIGURE A.1 Estimated Item Characteristic Function Following the Normal Ogive Model (Solid Curve) and the Frequency Ratios Roughly Approximating the Non-parametric Item Characteristic Function (Dotted Curve) and Plausibility Functions (Dashed Curve), for Each of 43 Items of Level 11 Vocabulary.

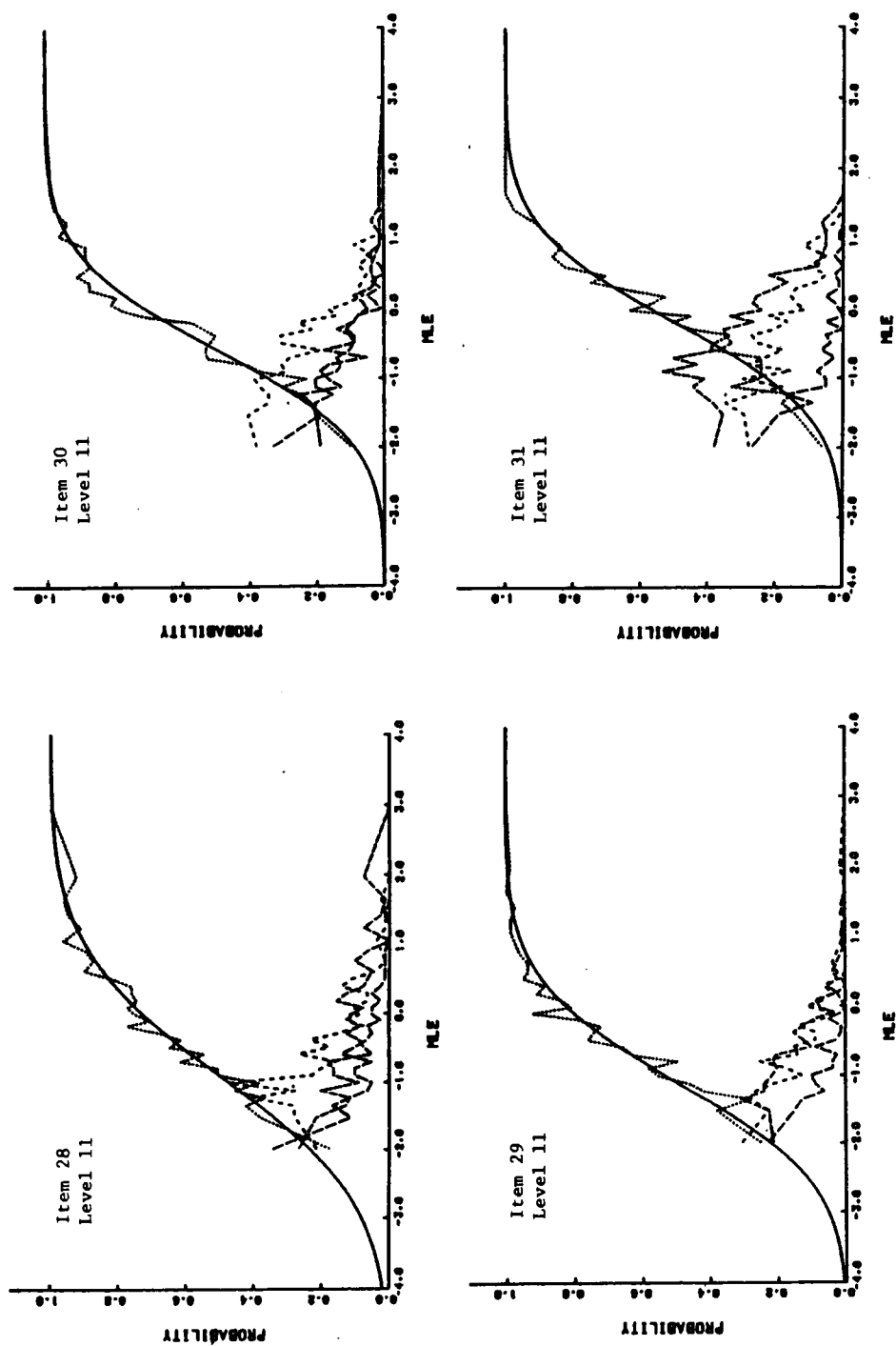


FIGURE A.1 (Continued)

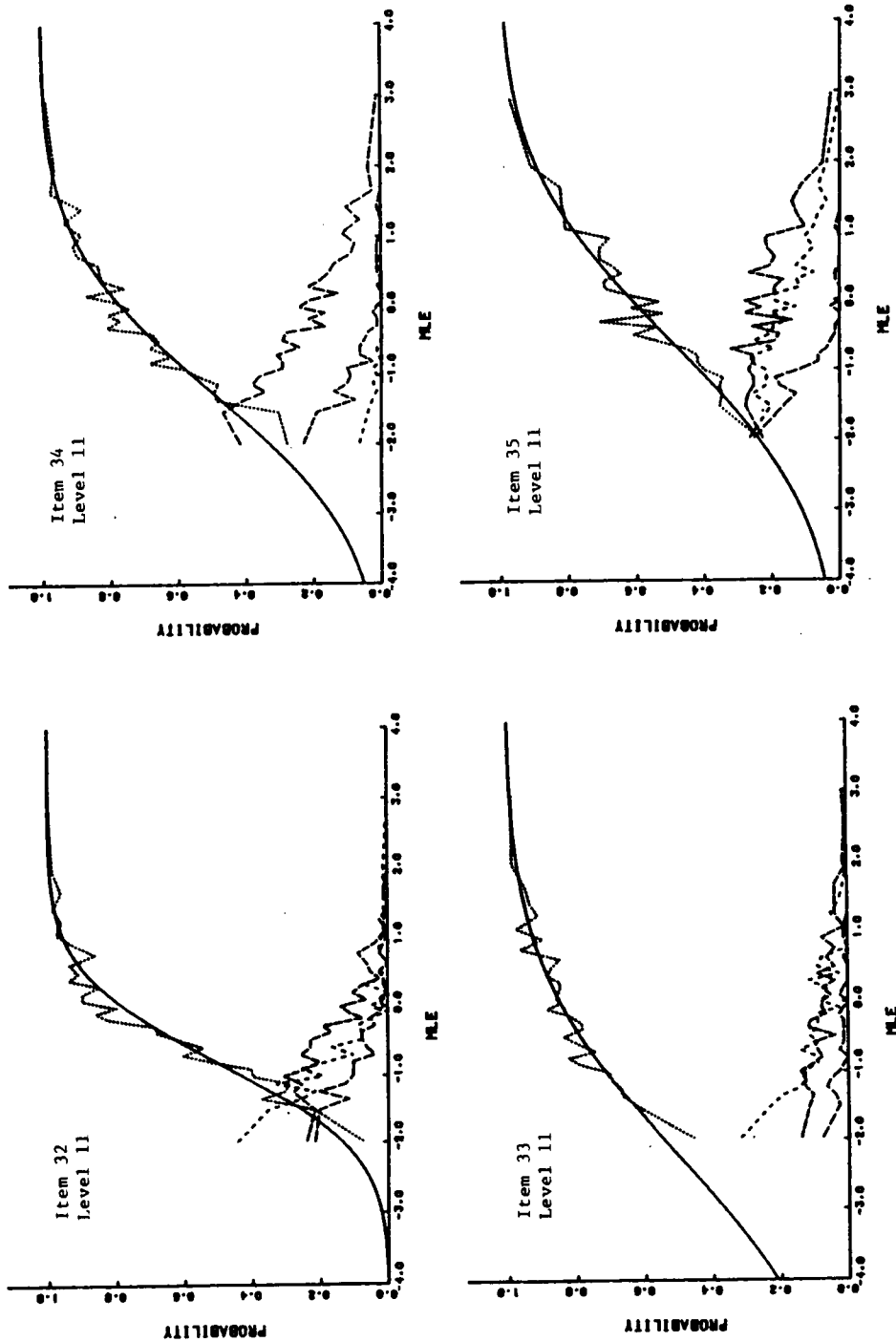


FIGURE A.1 (Continued)

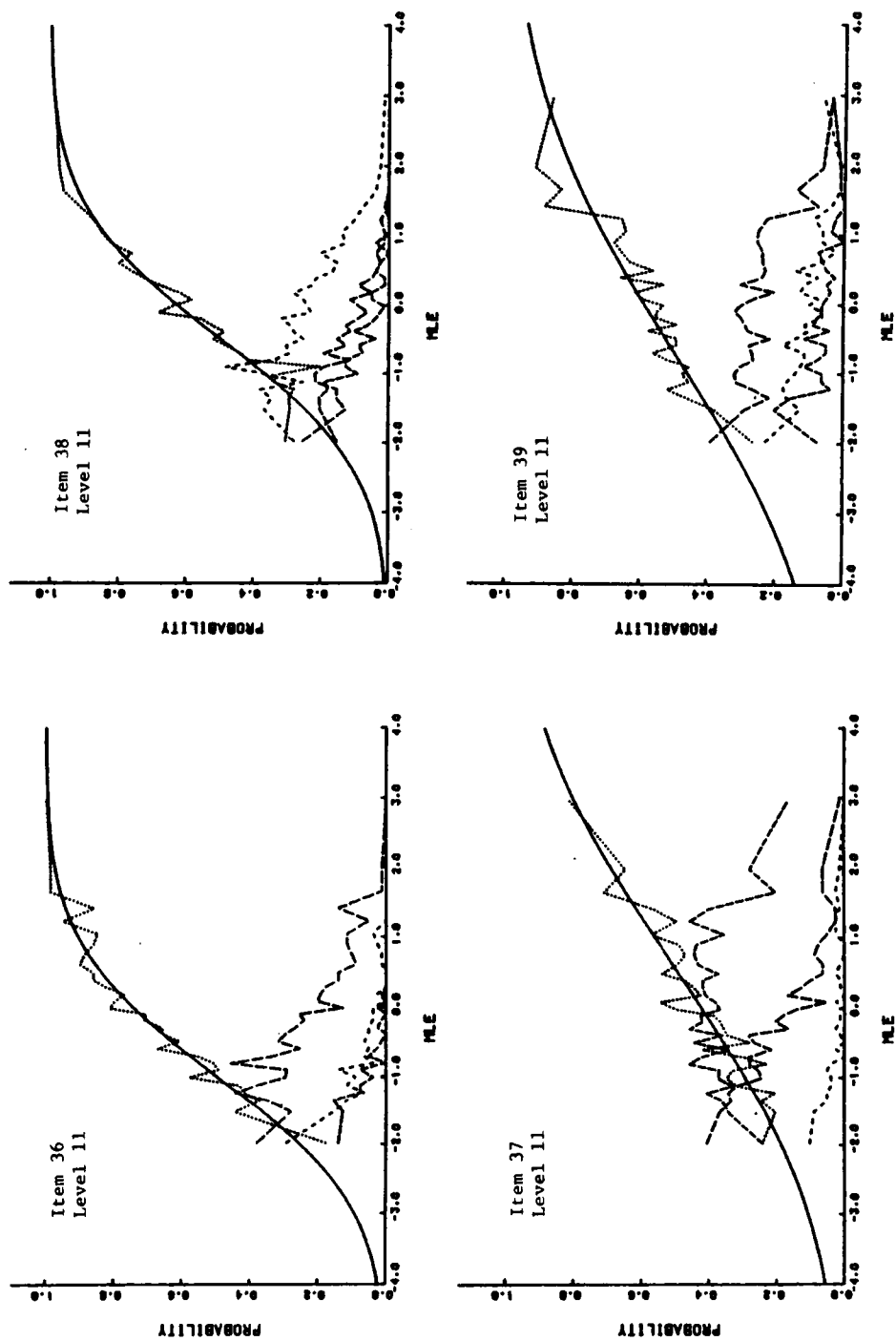


FIGURE A.1 (Continued)

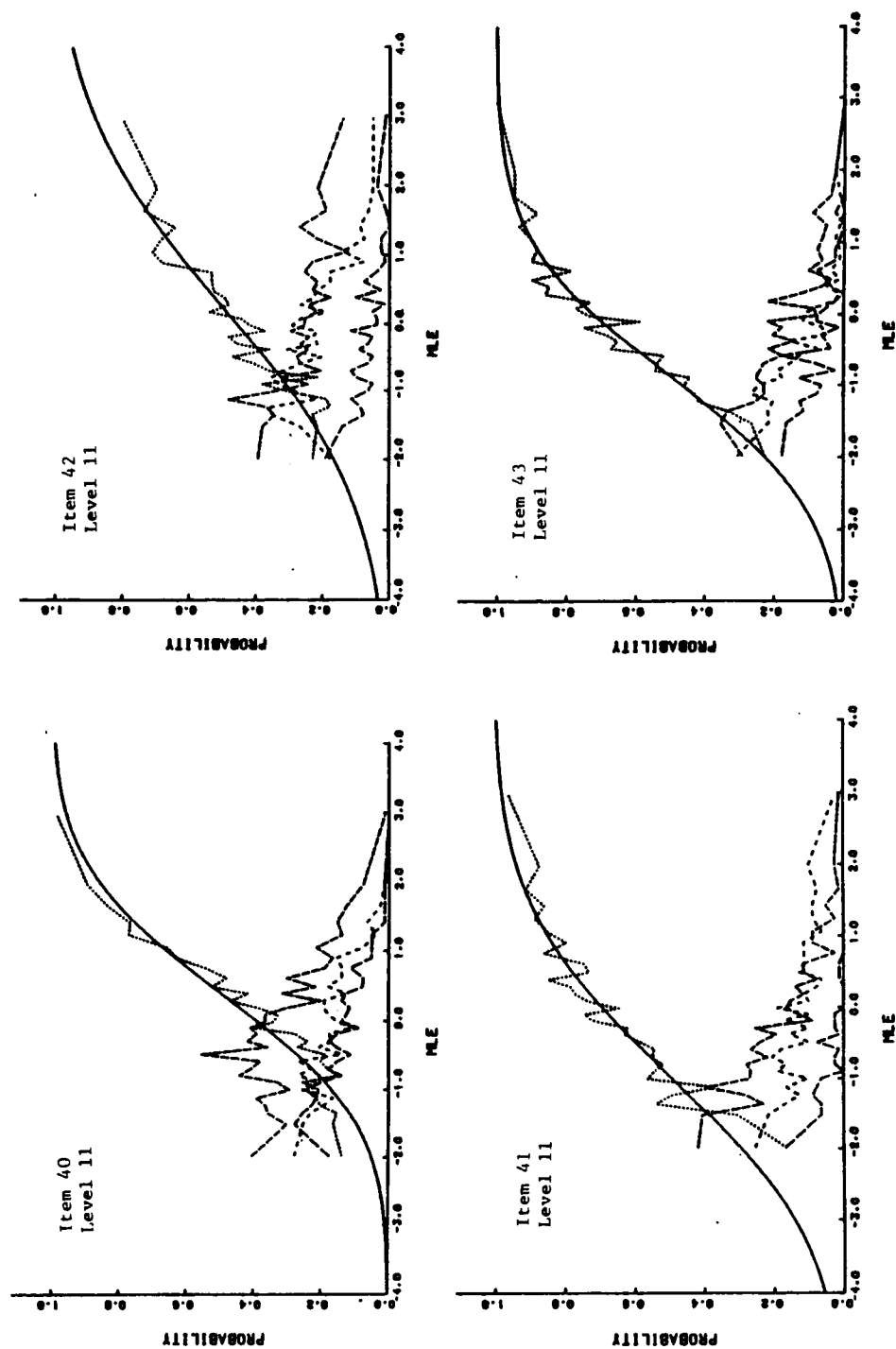


FIGURE A.1 (Continued)

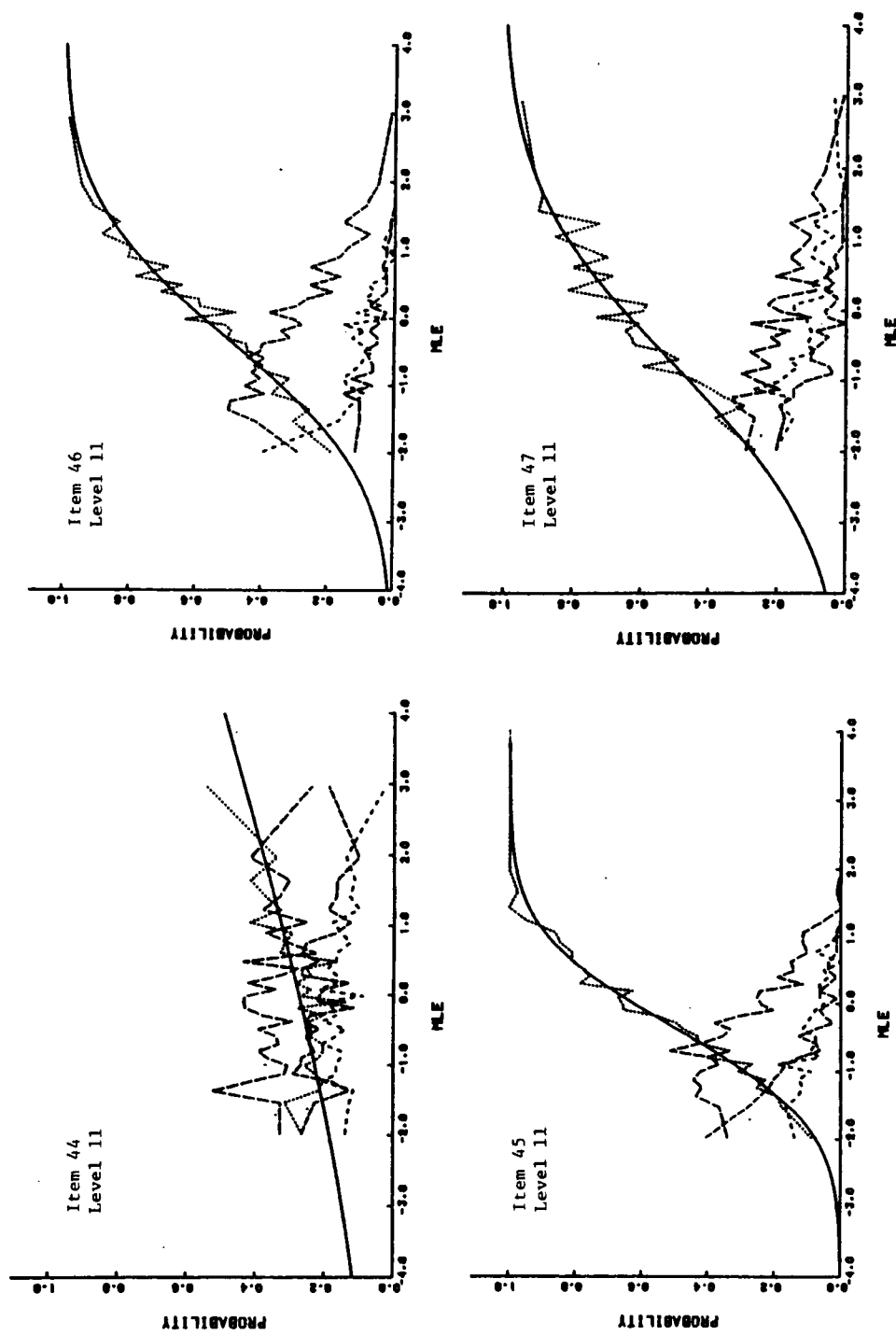


FIGURE A.1 (Continued)

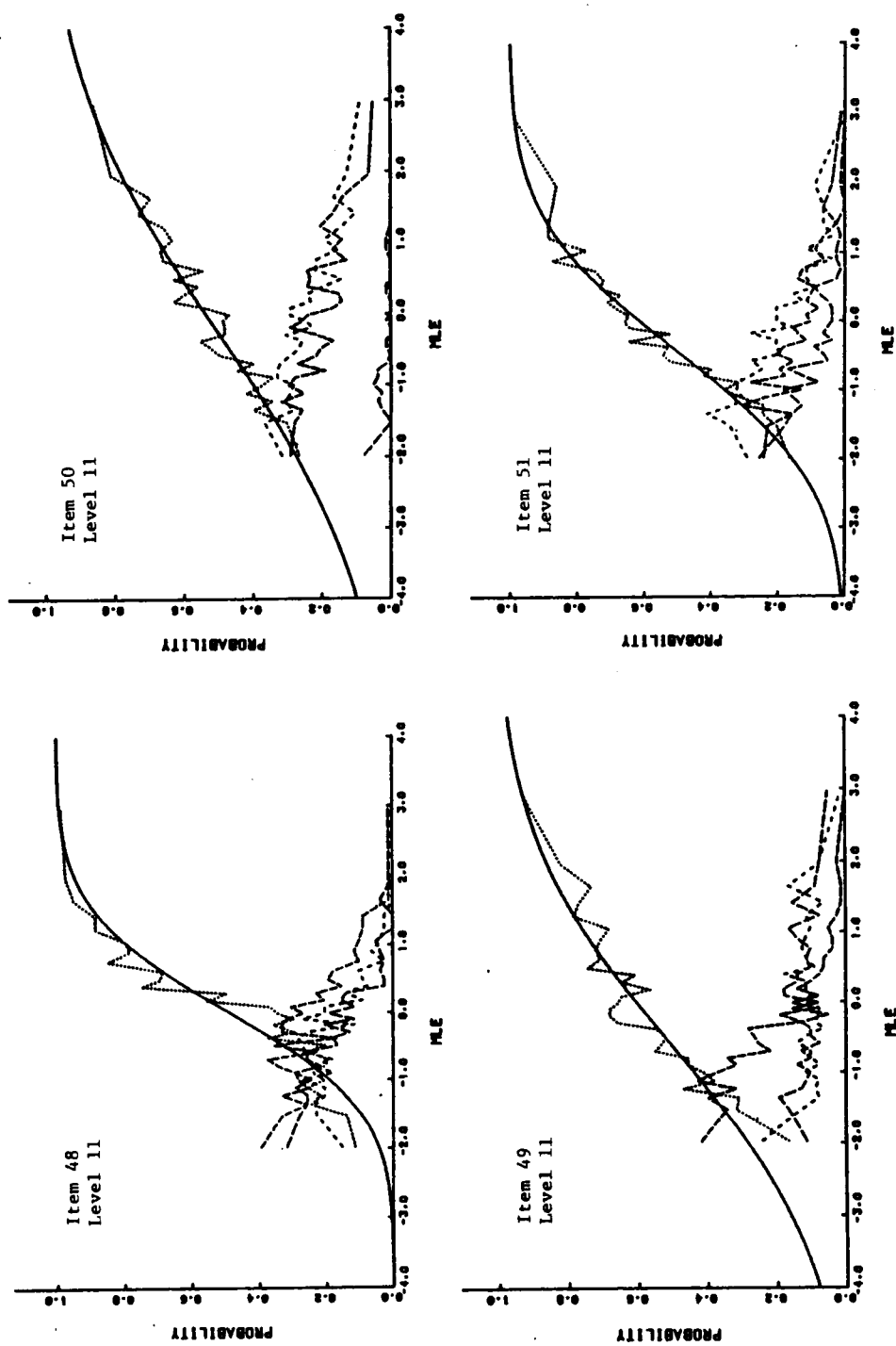


FIGURE A.1 (Continued)

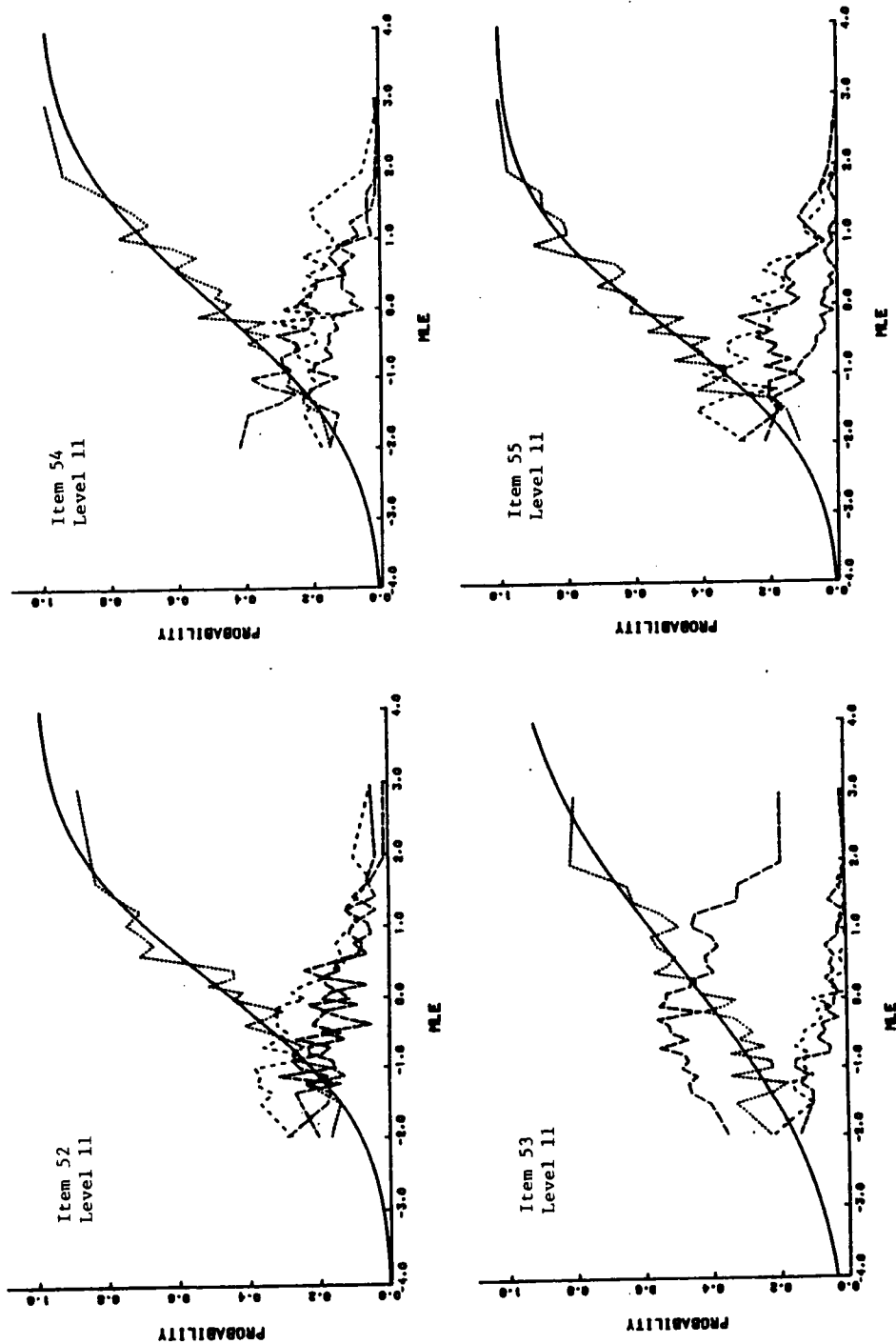


FIGURE A.1 (Continued)

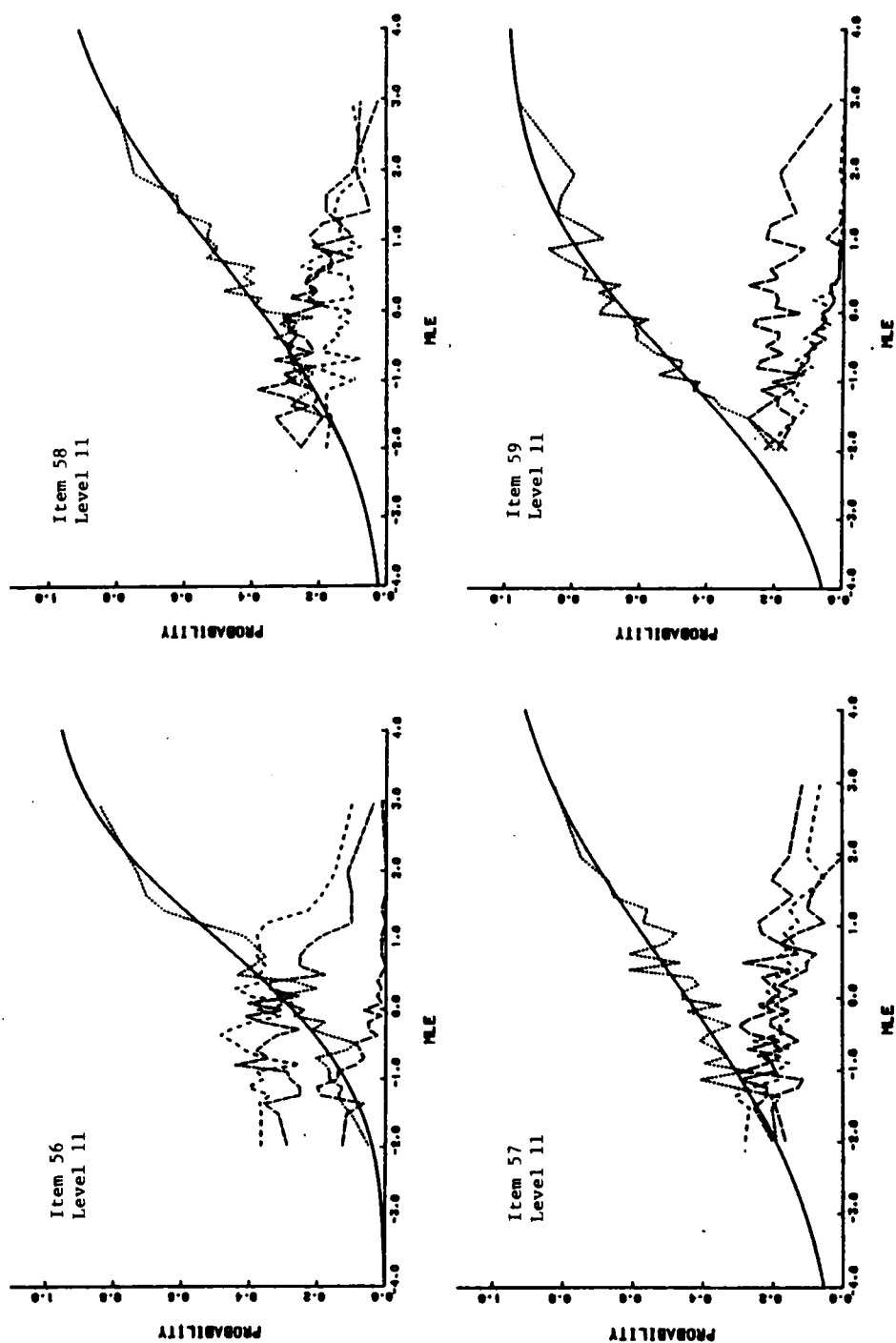


FIGURE A.1 (Continued)

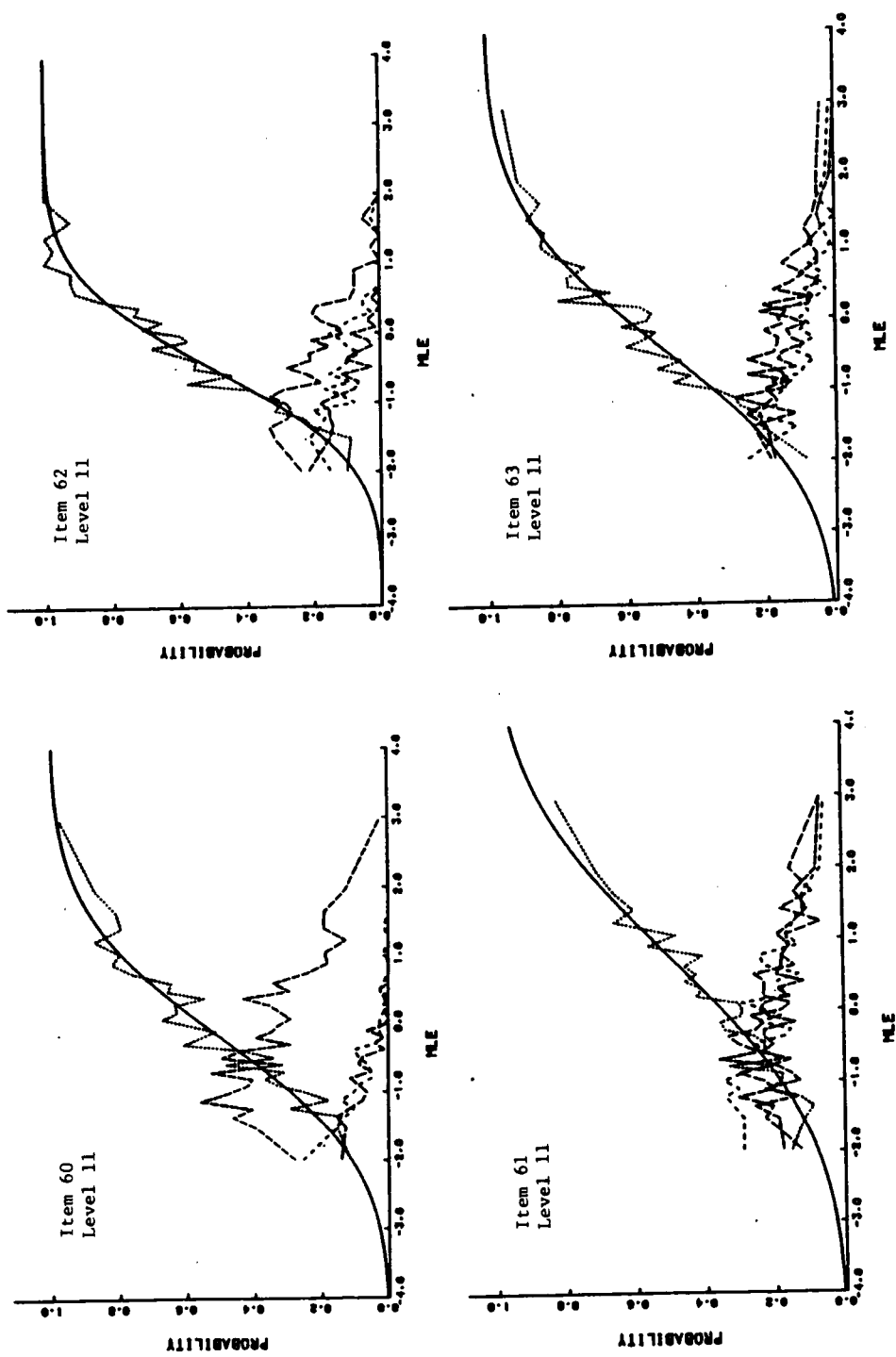


FIGURE A.1 (Continued)

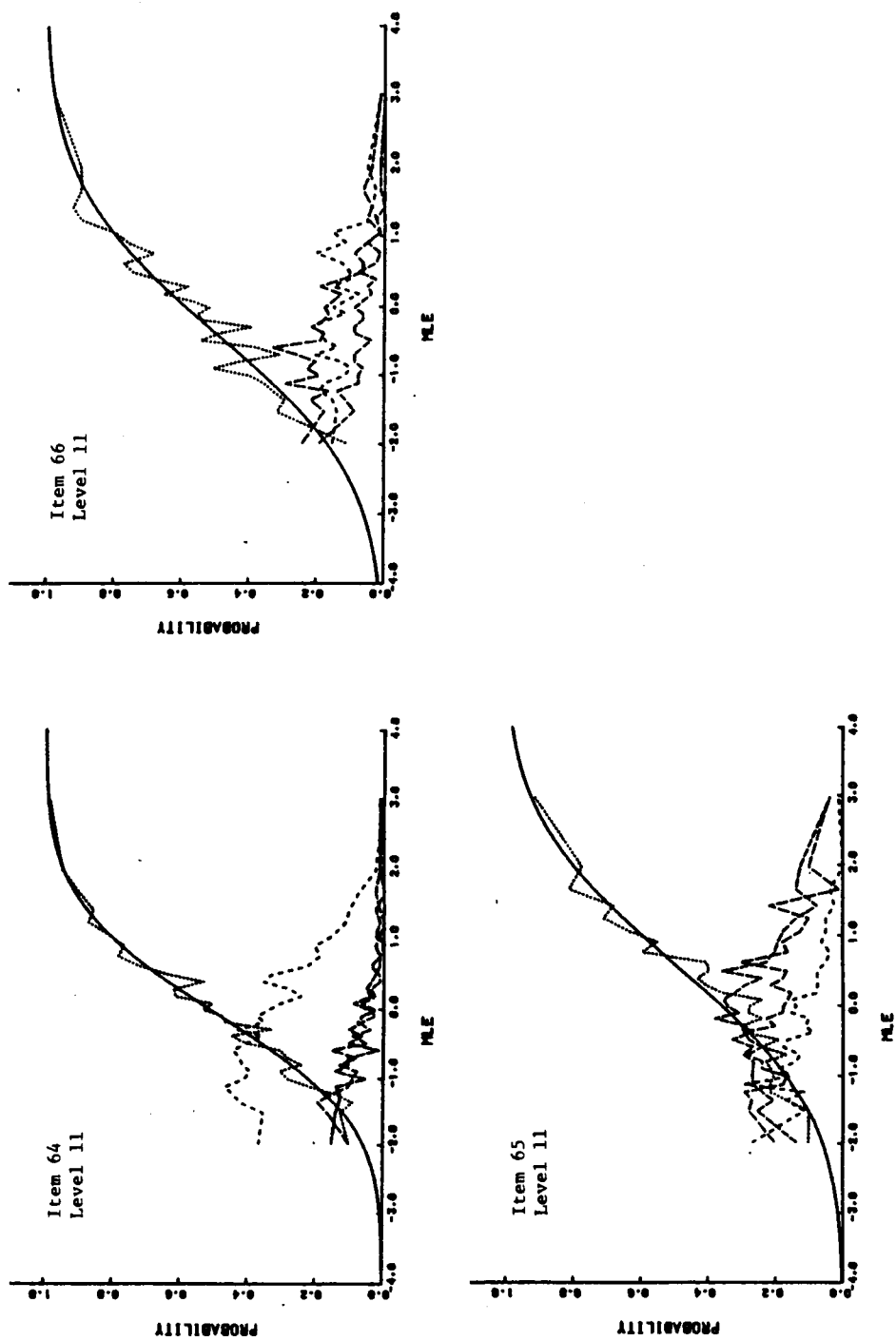


FIGURE A.1 (Continued)

TABLE A.2 Comparison of Informative Distractors Identified from Plausibility Functions to Those Identified from Frequency Ratios for Level 11 Vocabulary.

Item	Frequency Ratios	Plausibility Functions
24	(D)	---
25	---	---
26	(C)	C
27	(C)	C
28	B	B
29	(B)	---
30	---	B
31	C	C
32	(B)	---
33	---	---
34	(D)	D
35	---	C
36	(D)	D
37	C	C,D
38	(B)	B
39	---	D
40	C	C
41	---	C
42	(C)	C
43	---	---
44	---	D
45	(D)	D
46	C	C
47	---	D
48	(C)	C
49	---	C
50	---	B,C
51	B	B
52	---	B
53	D	D
54	C	C
55	B	B
56	B,C	B,C
57	---	---
58	C	C
59	D	---
60	C	C
61	---	B
62	(D)	D
63	---	---
64	(B)	B
65	C,D	C,D
66	(D)	---

APPENDIX B

In evaluating the information provided by the plausibility functions, care must be taken in the selection of the interval of the latent trait upon which conclusions are to be based (Samejima, personal communication, 1991). Generally there is greater likelihood for estimation error and/or bias in the plausibility functions corresponding to the extreme ranges of the ability distribution--both high and low--of the population used in the estimation process. In order to determine the ability range for which one can feel confident in making inferences, Samejima recommends that the interval be found which satisfies the following conditions:

1. The test information function is sufficiently high.
2. The discrepancy between the square root of the test information function of θ and its polynomial approximation is small.
3. The discrepancy between the target square root of the test information function of the transformed latent trait τ and its polynomial approximation is small.
4. The conditional distribution of τ , given $\hat{\tau}$, approximates the normal distribution.
5. The ability distribution of the population used in the estimation process is sufficiently high.

Before a suitable interval of the ability scale for each subtest investigated in this dissertation study could be found, the author had to operationally define each of the five requirements listed above. For instance, what value of the test information function could be considered sufficient as a minimum standard? A

review of the literature gave the author some insight--although no definitive answers--to this and other related questions. The operational definitions for these conditions or requirements follow.

1. The square root of the test information function of the original latent trait, $[I(\theta)]^{-1/2}$, is greater than or equal to 2.5. Samejima (1977, p. 236) explored the issue of a minimum standard for the test information function in individual ability estimation with adaptive, or tailored, testing:

It might be asked: How large should $I(\theta)$ be in order to tolerate the approximate normal distribution for the error ϵ ? To answer this we must depend, more or less, on our subjective judgment.

In this same paper Samejima cited a study using monte carlo data which yielded adequate results when the test information was as low as 7.4 ($[I(\theta)]^{-1/2} = 2.7$).

In another monte carlo study Samejima and Changas (1981) investigated the minimum number of Old Test items for which the Simple Sum Procedure/Normal Approach Method successfully estimated item characteristic functions. It was found that adequate estimation was possible with as few as eleven items. The square root of the test information function ranged from approximately 2.5 to 2.8 over the entire interval of the latent trait.

2. The absolute difference between the square root of the test information function, $[I(\theta)]^{-1/2}$, and its polynomial approximation is less than or equal to 0.05. Visual inspection of the graphs of these two functions will show that the discrepancy is almost imperceptible when this condition is satisfied. While a less conservative standard could likely be supported, the very small discrepancy over a broad

interval of ability found in this study makes discussion of such a standard unnecessary.

3. The absolute difference between the target square root of the test information function of the transformed latent trait τ and its polynomial approximation is less than or equal to 0.05 . This criterion was selected for the same reasons discussed for condition 2 above.

4. The conditional distribution of τ , given $\hat{\tau}$, should be symmetric for examinees within the appropriate interval. More specifically, this conditional distribution will be identified as normal or Pearson's System Type II in the results obtained by the computer program Conditional Moments MLE. (See Samejima, 1981, for further description of the Pearson System Method.)

5. The estimated density of the transformed maximum likelihood estimate, $\hat{\tau}$, denoted as $\hat{g}^*(\hat{\tau})$, is greater than or equal to 0.05 . While this criterion is somewhat arbitrary, it does reduce the likelihood that too few examinees are represented in any subinterval.

The interval of the latent trait θ and the transformed latent trait τ which satisfies each of these five conditions is presented in Table B.1, for each of Levels 11-13. For conditions 4 and 5, which involve the maximum likelihood estimate, $\hat{\tau}$, the conditional mean of τ , given $\hat{\tau}$, was used to determine the corresponding interval of $\hat{\tau}$.

Condition 4 was generally found to be the most restrictive factor in determining an appropriate interval. The possibility of defining a second criterion for condition 4 was investigated, with the hope of obtaining a broader interval of the latent trait than one

TABLE B.1 Interval of Latent Trait θ and Transformed Latent Trait τ for which Each of Individual Conditions 1 through 5 is Met, for Levels 11, 12 and 13.

Condition	Latent Trait	Level 11	Level 12	Level 13
1		[-2.1,1.2]	[-2.3,0.9]	[-2.3,1.0]
		[-1.6,0.9]	[-2.3,0.7]	[-2.3,0.9]
2		[-4.7,4.7]	[-4.7,4.7]	[-5.2,5.6]
		[-2.6,2.2]	[-3.6,2.7]	[-3.8,3.9]
3		[-3.8,3.7]	[-4.5,3.7]	[-4.3,4.3]
		[-2.4,2.0]	[-3.6,2.5]	[-3.5,2.7]
4		[-1.1,1.1]	[-1.1,1.2]	[-1.1,1.1]
		[-0.9,0.9]	[-1.2,1.1]	[-1.1,1.0]
5		[-1.6,2.9]	[-1.7,2.1]	[-1.7,2.0]
		[-1.3,1.7]	[-1.8,1.7]	[-1.7,1.7]

could still feel relatively confident in drawing conclusions from.

The computer program Conditional Moments MLE defines the conditional distribution of τ , given $\hat{\tau}$, as normal if the Pearson's system coefficient β_2 is between 2.975 and 3.025, and as Type II if β_2 is less than 2.975 and Pearson's criterion κ is less than or equal to 0.025 in absolute value. If criterion κ is less than -0.025, the conditional distribution is defined as Type I, which is a general Beta distribution. If Type I cases which are slightly asymmetric--let us say $-0.035 \leq \kappa < -0.025$ --are accepted in addition to Type II and the normal distribution, we can establish a broader interval of ability that should still provide satisfactory results. To distinguish these two intervals, let us refer to the original, more conservative interval as Case 1, and to the broader interval as Case 2. Thus Case 1 satisfies all five conditions operationally defined initially in this section, while Case 2 satisfies a modified version of condition 4.

Table B.2 presents the intervals of the original latent trait θ and the transformed latent trait τ yielded by Case 1 and Case 2, for each of Levels 11-13 of the Vocabulary subtest. For Level 11 Vocabulary, for example, one can be most comfortable with the results of the plausibility functions over the interval of θ from -1.1 to 0.8, but there is a greater likelihood of estimation error and/or bias in the results outside of this interval, particularly outside of the Case 2 interval, [-1.3, 1.2]. We see that there is not a great difference in the intervals for cases 1 and 2, particularly at the lower boundary of the interval for any given level.

TABLE B.2 Interval of Latent Trait θ and Transformed Latent Trait τ for which All Conditions, 1 through 5, Are Satisfied, for Levels 11, 12 and 13. Two Different Criteria, Case 1 and Case 2, Are Established for Condition 4.

Case	Latent Trait	Level 11	Level 12	Level 13
1	θ	[-1.1,0.8]	[-1.1,0.5]	[-1.1,0.4]
	τ	[-0.9,0.6]	[-1.2,0.5]	[-1.1,0.4]
2	θ	[-1.3,1.2]	[-1.3,1.6]	[-1.3,1.3]
	τ	[-1.0,0.9]	[-1.4,1.3]	[-1.3,1.2]

The graphs of the plausibility functions for Levels 11 through 13 were reviewed with the consideration that there may be greater estimation error outside the Case 1 and Case 2 intervals of the latent trait. In general, the intervals defined by Case 1 are sufficient to identify the vast majority of informative distractors. For all but a few of these distractors, the plausibility function rises above the chance level within the Case 1 interval. In the majority of cases where the peak of the plausibility function (see Table III.16, p. 152) is within the interval plotted, it falls within the Case 1 interval. This happens 12 of 18 times for Level 11. Two other informative distractors at Level 11 peak within the Case 2 interval. Of the 25 Level 12 informative distractors for which the peak can be identified in the graphs of the plausibility functions, the peak falls within the Case 2 interval 13 times, and ten of these are within Case 1. The peak is within the Case 2 interval for 18 of 25 such informative distractors at Level 13, 15 being within the smaller Case 1 interval.

There are only two informative distractors at Level 11 which are not clearly identifiable based solely on the plausibility functions within the Case 1 interval of θ . These two exceptions occur for items 41 and 42. For Level 12 there are three informative distractors that can not be determined from the Case 1 results. Two of these three distractors correspond to item 50 and the other to item 42. There are five such cases for Level 13. Informative distractors for items 61, 62, 79, 80 (distractor B only), and 92 are not clearly identified using the Case 1 interval.

It appears that the conclusions drawn earlier from the

plausibility functions in Figure III.2 generally hold well up when one looks at a smaller but more reliable interval of the latent trait. The selection of a smaller interval, whether it be Case 1 or Case 2, does not necessarily mean that we must ignore the results outside of this interval. Further research into the effects of violating any of the five conditions listed earlier is necessary to more accurately assess the degree of estimation error that may occur. However, one should be careful in interpretations made from certain extreme ranges of the ability scale. Of particular concern should be the temptation to overgeneralize from small differences in the plausibility functions at these extremes.

VITA

Paul Chngas was born in Knoxville, Tennessee on April 30, 1956. He attended public schools in Knoxville until his graduation from West High School in 1974. He then enrolled at the University of Tennessee and graduated in 1978 with a Bachelor of Arts degree in Psychology. During this time he was selected to the Phi Beta Kappa Honor Society.

Following graduation he entered the Experimental Psychology program at the University of Tennessee and served as a graduate research assistant in a psychometrics laboratory. From 1981 to 1983 he served a School Psychology internship with the Knoxville City School System. He continued his psychometric research assistantship through 1985.

In 1986 he began working as a statistician for the Tennessee State Testing and Evaluation Center. He currently serves as Testing Consultant for the Center.

He has been married to the former Kathy Wallis of Knoxville since 1982.