1-1-1981

The fit of empirical data to two latent trait models.

Leah R. Hutten
University of Massachusetts Amherst

THE FIT OF EMPIRICAL DATA TO TWO LATENT TRAIT MODELS

A Dissertation Presented

By

LEAH R. HUTTEN

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

DOCTOR OF EDUCATION

September 1981

School of Education

THE FIT OF EMPIRICAL DATA TO TWO LATENT TRAIT MODELS

A Dissertation Presented

By

LEAH R. HUTTEN

Approved as to style and content by:

Ronald K. Hambleton

Ronald K. Hambleton, Chairperson

H Swaminathan

H. Swaminathan, Member

Harry Schumer

Harry Schumer, Member

H Swaminathan

Hariharan Swaminathan, Acting
Associate Dean for Academic Affairs
School of Education

ACKNOWLEDGEMENTS

Ronald Hambleton, acting in the capacity of my committee chairperson, gave me the incentive to initiate this project. His continued guidance and faith in me were the primary factors that supported me through this work. Dr. Hambleton also generously supplied funding sources, secretarial support, research assistantships, his private library, and friendship throughout. I also wish to thank the other members of my thesis committee, H. Swaminathan and Harry Schumer for the knowledge and inspiration they provided.

This study would not have been possible without the typing support provided by Bernie McDonald and editorial assistance by Stephen Miles.

Many people at Educational Testing Service provided assistance to me during this study. I especially want to thank Frederic Lord for critiquing the methodology and Marilyn Wingersky for help with LOGIST. ETS was also kind enough to provide SAT databases for the study.

I was fortunate to obtain a wide variety of data for this study. I wish to thank Dr. Eric Gardner, Psychological Corporation, Dr. Robert Rentz, and Educational Progress Corporation in this regard.

Throughout the course of the study many persons provided criticism and shared new ideas with me. I am especially grateful to Benjamin Wright, Mark Reckase, Linda Cook, Daniel Eignor, Jim Douglas, and Anne Fitzpatrick.

Initial funds were made available for this study from a contract with the U.S. Air Force Human Resources Laboratory. This support allowed for a much richer study than would otherwise have been possible.

ABSTRACT

The Fit of Empirical Data to Two Latent Trait Models

September 1981

Leah R. Hutten, B.A., University of Wisconsin-Madison

Ed.D., University of Massachusetts

Directed by: Ronald K. Hambleton

Fit of data to the Rasch and three-parameter logistic latent trait models was explored with 25 empirical datasets. Deviations in data from latent trait model assumptions were the primary variables of interest. The study also investigated estimation precision for small samples and short test lengths and evaluated costs for latent trait parameter estimation by the two latent trait models.

Ability and item parameters were estimated under the assumptions of the Rasch and three-parameter models for tests with 40 items and 1000 examinees. Estimated parameters were substituted for true parameters to make predictions about number-correct score distributions. When ability is known, a theorem by Lord (1980) equates ability with the conditional distribution of number-correct scores. Predicted score distributions were compared to observed score distributions with statistical and graphical techniques. Both Kolmogorov-Smirnov and Chi-square test statistics were obtained. The importance of three latent trait model assumptions, unidimensionality, equality of item discrimination indices, and no guessing were assessed with correlation analyses. Estimation precision for short tests of only 20 items, and small samples of 250 examinees were evaluated with correlation methods and average absolute differences between estimates. CPU time and cost were tallied for estimations by each model and summary statistics were gathered for comparison purposes.

Both the Rasch and three-parameter models demonstrated reasonably good fit to most of the 25 tests. Only one test deviated greatly from the two models. Five tests did not appear to fit very well when the chi-square was employed as the criterion. The chi-square test was more rigid than the Kolmogorov-Smirnov test and tended to be very sensitive to irregularities and lack of normality in observed score distributions. Graphic results tended to support outcomes of the Kolmogorov-Smirnov test.

Overall, the Rasch model fit data as well as the three-parameter model. Average K-S statistics across the 25 tests were 1.304 for the Rasch model and 1.289 for the three-parameter model. For 65 percent of the tests, the three-parameter model fit data better than the Rasch model, although in most cases, fit statistics for the two models were very close. Similar results were obtained with chi-square measures, although these statistics favored three-parameter model fit somewhat. Graphic evidence demonstrated how analogous fit was for the two models.

Lack of unidimensionality was found to be a primary cause for misfit of models to the data. Correlations between fit statistics and indices of unidimensionality were significant at the .05 level of probability for both the Rasch and three-parameter models. A weak relationship was found between equality of item discrimination indices and fit to the Rasch model. Generally, data with more equal item discriminations fit both models slightly better than other data. Underestimation of the amount of guessing for both models resulted in less adequate model fit. Sample sizes were not sufficient for obtaining accurate estimates of guessing.

Ability estimates from short 20-item tests were somewhat more precise for the Rasch model than for the three-parameter model. Generally, good estimates of ability from short tests were obtained from both models. Correlations between ability estimates on short and longer tests were .923 for the Rasch model and .866 for the three-parameter model.

Estimates of item difficulty made on samples of 250 examinees in contrast to larger samples (N=1000) were very good for both models. Estimates of other item parameters from samples of 250 examinees were not very accurate. Item discrimination estimates from small samples were reasonable, but estimates of guessing were very poor. The results indicated that samples of at least 1000 examinees are needed to obtain stable estimates of parameters for the three-parameter model. Smaller samples suffice for obtaining Rasch difficulty estimates.

The cost and computer time for simultaneous estimation of ability and item parameters for the Rasch model was one-third that for the three-parameter model. For 25 tests, the average Rasch model estimation cost $12.50 in contrast to $35.12 for the three-parameter model. When item parameters were known in advance, and only abilities were estimated, the cost of estimation by the two models was identical. These results suggest that in the long run, the differences in cost between estimation by the Rasch and three-parameter models is negligible.

TABLE OF CONTENTS

CHAPTER

CHAPTER

LIST OF TABLES

Table

Table

LIST OF FIGURES

Figure

Figure

CHAPTER I

INTRODUCTION

Item response theory, or latent trait theory as it is commonly known, was proposed over twenty-five years ago by Frederic Lord (1952) in the United States and coincidentally by Georg Rasch (1960) in Denmark. Lord's work focused on exploring relationships between ability and the probability of responding correctly to items designed to measure ability. A function describing examinee success in terms of ability was called an item characteristic curve (ICC) by Lord and formed the basis for latent trait theory.

Lord postulated the shapes of ICC's to be normal ogives and showed that their form could be derived if certain assumptions are made (Lord & Novick, 1968). These models are characterized in the general case by three parameters: one describing the point of inflexion in the curve, one describing the slope, and one characterizing the lower asymptote. Practically these parameters translate into item difficulty, item discrimination, and guessing, respectively. Later, in conjunction with Birnbaum (1968), Lord determined that substituting logistic curves of the form $e^x/(1+e^x)$ for normal ogives reduced calculation difficulties and made the models mathematically tractable. Throughout most of its early development, latent trait theory remained a theoretical description with little practical relevance due to mathematical and computational complexities in estimating latent trait parameters.

Concurrently with the development of the two- and three-parameter models in the United States, Georg Rasch (1960), a Danish mathematician, independently derived a theory of test scores which turned out to be a very interesting, albeit a special case of the work in progress by Lord and Birnbaum. As a mathematician, Rasch had studied scales of measurement in the physical world. He believed that mental measurement could be as objective as physical measurement. Mental measures were classically derived from sample-based statistics, a tenet which forms the basis of classical test theory. Conventional scores on mental measures are reported as relative positions in some reference group composed of items and people. Rasch proposed sample-invariant measurement on an objective scale which described both items and people. The scale when applied to people measured ability. When applied to items, the scale measured difficulty. Rasch proposed a theory of measurement which associated the probability of examinee success on items with underlying ability. Although logically, the Rasch model is derived from the contention that measurement should be objective, mathematically, the Rasch model is the simplest case of the more general logistic models. The slopes of Rasch ICC's are equal (equal item discrimination) and the lower asymptotes are all zero (no guessing).

Purpose of the Research

Because of the historically separate origins of latent trait models, little comparative research has been performed with the models. The current study was undertaken to highlight some similarities and differences between two latent trait models. Specifically, the purpose of the study was to compare the Rasch (or one-parameter) logistic model with the Birnbaum (or three-parameter) logistic model by fitting the models to empirical data.

Proponents for both models have asserted that their model is most appropriate for describing test behavior, yet little empirical evidence has emerged to confirm these claims. Practitioners have relied primarily upon theoretical assertions to select from the latent trait models. This research was designed to provide concrete information about the dynamics of the latent trait models particularly as they apply to estimating ability.

The study examined fit of the models to empirical data. Model fit was systematically analyzed in terms of deviations from latent trait model assumptions occurring in the data. This comparison is important because of the potential ramifications, legal or otherwise, that could result when the assumptions of the models have not been met. Because imprecise parameter estimates may be another cause for misfit of models, the study also examined the suitability of the models in situations where few examinees were available for estimating item parameters. Although there have been many applications of latent trait theory in nationwide standardized testing programs, there has been increasing interest in their use by local school systems, the military and by other small scale testing programs. This part of the study also provided information on the precision of ability estimates based on short tests, tests typically used in the classroom. Finally, comparative cost information for estimating parameters by the two models was collected. While cost should not usually be the primary reason for selecting one model over the other, expenses are an important issue today because of shrinking federal, state, and local education budgets. Because certain models may be more desirable than others for certain applications, e.g., equating test scores, the information provided by this study can help practitioners make informed rather than arbitrary decisions about latent trait model selection.

Research Questions

Because this study was exploratory in nature, no specific hypotheses were tested, rather the study sought to provide information in the following areas:

1. What methods can be used to determine that empirical data meet the underlying assumptions of latent trait models? The assumptions include unidimensionality (and equivalently, local independence), equality of item discriminations (Rasch model), and no guessing (Rasch model). Information in this area was obtained from a review of the literature. Various procedures were explored on a trial basis, and those selected were critically analyzed. Recommendations were made for how model assumptions can be tested.

2. How is model fit defined and what statistical, graphical, and practical procedures can be employed to determine model fit? Three measures of fit were used in the study. Outcomes based on each measure were compared and suggestions offered for future research.

3. Do latent trait models fit tests developed by conventional methods? Which model demonstrates better fit to empirical data? Fit statistics and graphical evidence of fit of the Rasch and three-parameter models to 25 empirical data sets were obtained. Results based on the various methods of fit were compared.

4. How do deviations from latent trait model assumptions affect fit of data to the latent trait models? Are the models robust to violations in their assumptions? For both models, fit was explored in terms of unidimensionality. For the Rasch model, fit statistics were examined when equality of item discriminations and guessing assumptions were violated in the data. Correlation and partial correlation techniques were used to provide information in this area.

5. How precise are estimates of ability made on short tests? Three measures of precision for short tests were used: Pearson correlations, Spearman rank order correlations, and average absolute differences (AAD).

6. How precise are estimates of item parameters from small samples of examinees? Pearson correlations, Spearman correlations, and AAD statistics were used to explore precision of item parameters from small samples.

7. What are the comparative costs (in terms of computer time and expense) for obtaining parameter estimates of the one- and three-parameter latent trait models? CPU time and cost were tallied and compared for parameter estimation under each model.

Concepts Utilized in Latent Trait Theory

A *latent trait* is a skill or ability (or attitude or perception) which is not directly measurable but can be inferred from examinees' responses to test items. Conventional estimates of ability, the *raw score* or *number-correct* score, differ from the latent trait ability estimate because the latter is measured on a standard score scale independent of the number of items on a test. The *true score* can be seen as a transformation of ability onto a number-correct score scale.

Ability estimates in latent trait theory are based on probabilistic models. When the difficulty level of an item is known it is possible to draw inferences about ability from scores on single items because difficulty and ability can be represented on the same scale. Assume that an item has a difficulty level of "γ_i". If an examinee obtains a correct response on the item, it is probable that ability is greater than or equal to γ_j ($\theta \geq \gamma_j$), whereas if the examinee fails the item, it is probable that ability is less than γ_j ($\theta < \gamma_j$). When such ability estimates are made on a sufficient number of items, it is possible to obtain a good measure of ability. Consistent estimates of ability can be found when test length is reasonably long. Figure 1 illustrates differences between conventional and latent trait methods for estimating ability. Three hypothetical six-item tests are shown in the figure: the top line of the figure depicts a test with items of mixed difficulty; the middle line shows an easy test; and the bottom line illustrates a hard test. Since latent trait item difficulty estimates are measured on the same scale as ability, a single scale ranging from 0 to 10 has been arbitrarily chosen for difficulty and ability. An examinee, with an ability score of 5 (θ=5) on this scale, is illustrated in the figure. The conventional number-right scores for this examinee on the three tests are 3, 5, and 1, respectively. The conclusions about examinee ability drawn from conventional scoring of these tests differ significantly, but the latent trait ability estimate for this examinee would be the same (disregarding measurement error) regardless of which test the examinee was administered because of the use of item difficulty in estimating ability. Items which are tailored to examinees'

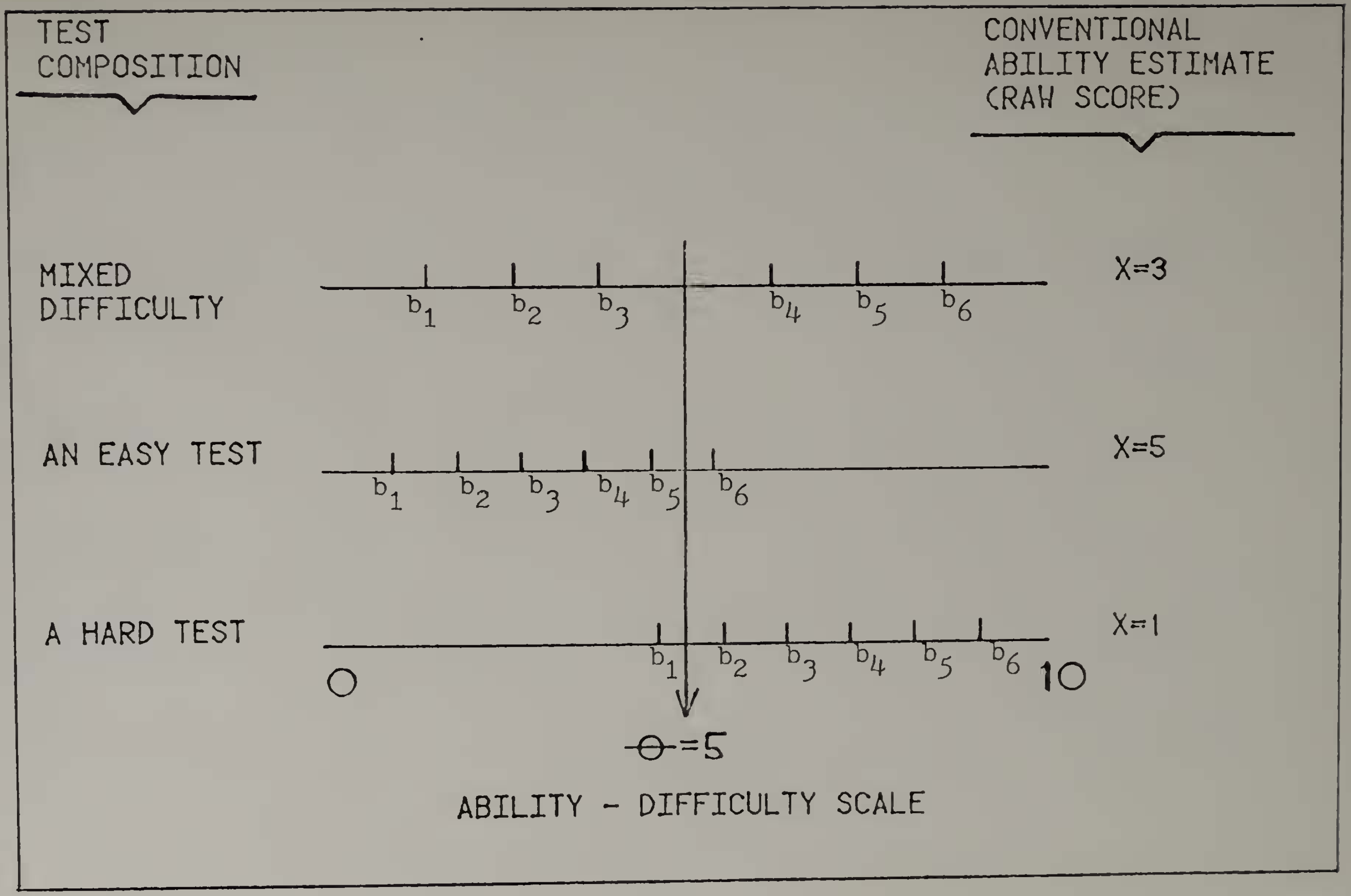

TEST
COMPOSITION
CONVENTIONAL
ABILITY ESTIMATE
(RAW SCORE)
MIXED
DIFFICULTY
b_1 b_2 b_3 b_4 b_5 b_6
X=3
AN EASY TEST
b_1 b_2 b_3 b_4 b_5 b_6
X=5
A HARD TEST
b_1 b_2 b_3 b_4 b_5 b_6
X=1
0
10
$\theta=5$
ABILITY - DIFFICULTY SCALE

Figure 1. A Comparison of Conventional and Latent Trait Estimates of Ability

ability levels provide excellent estimates of ability. Because latent trait ability estimates do not depend on the specific sample or number of items in a test, ability can be estimated with different sets of items.

Conventional number-right scores are derived without regard to item characteristics, such as item difficulty. The Rasch model incorporates item difficulty into estimating ability. Two additional item characteristics are considered in the three-parameter model: item discrimination and guessing. Item discrimination operates as a weight such that better (more discriminating) items have greater importance for estimating ability than items which are not very discriminating. Items can be selected so that they are most discriminating at particular locations on the ability continuum. The Rasch model makes the assumption that all items are equally discriminating, a proposition somewhat difficult to meet in practice. If present, equal item discrimination would be signalled by equal item-total score correlations.

The guessing or chance level parameter is the third parameter in the three-parameter model. It is assumed in the model that the probability of success on an item may be greater than zero when items are multiple choice. The chance level parameter is particularly important for estimating ability at the low end of the ability continuum where guessing is most likely to occur. When the parameter is not included, such as the case of the Rasch model, ability estimates for low ability examinees tend to be too high.

Two important assumptions of both models discussed here are unidimensionality and the equivalent assumption of local independence. Unidimensionality means that a test includes items which tap only a single underlying trait. Although there are multivariate extensions to latent trait models, these are not considered in this study. Local independence means that responses to items are statistically independent: for examinees of the same ability, the porbability of success on an item is not related to the probability of success on any other item. Local independence is indicated by a lack or correlation between items for examinees at the same bility level.

Model Descriptions

The latent trait models compared in this study have the logistic form:

$$P = e^{x}/(1+e^{x}), \qquad [1]$$

where P is the probability of a correct response. For the Rasch or one-parameter model, the probability function is given by:

$$P_g(\theta) = \frac{e^{(\theta-b_g)}}{1+e^{(\theta-b_g)}} \quad , \qquad [2]$$

and for the three-parameter model the probability of a correct response is:

$$P_g(\theta) = c_g + (1-c_g)\,\frac{e^{Da_g(\theta-b_g)}}{1+e^{Da_g(\theta-b_g)}} \quad . \qquad [3]$$

These functions relating ability to the probability of a correct response on an item are known as item characteristic curves (ICC's). The constant, D, in the three-parameter model is set to 1.7 to equate the model to the form of the normal ogive function introduced by Lord in 1952. Ability, θ, is measured on a standard score scale with practical values in the range -3 to +3. which can be linearly transformed to any arbitrary scale such as the LOGIT scale which is seen commonly in practice. Item difficulty, b_g, in the equations, is measured on the same scale as ability, with a practical range from -2 to +2. Item difficulty is the point on the ICC where the probability of a correct response is .5 if there is no guessing. Item discrimination and guessing parameters apply only to the three-parameter model. Item discrimination, a_g, is measured on a scale ranging from 0 to +2, although in theory the values can be considerably higher. Negative values of discrimination are possible but usually such items are deleted from tests. Item discrimination is proportional to the slope of the ICC at the point of its inflexion. Since item discrimination values are assumed to be equal, the a_g term does not appear in the Rasch ICC. Some developments of the Rasch model include the mean item discrimination value, $\bar{a}$, in the equation. In this instance the power of the exponent is: $D\bar{a}(\theta-b_g)$. The guessing parameter, c_g, or chance probability level, ranges from 0.0 to 1.0. This parameter forms the lower left asymptote to the ICC and represents the probability of success by chance alone. Practical limits for guessing are 0.0 to 0.5 and are related to the number of item choices. Since no guessing is assumed by the Rasch model, the parameter has the value zero and does not appear in the ICC.

Two item characteristic curves for the three- and one-parameter models are pictured respectively in Figures 2 and 3. The slopes of Rasch ICC's are identical and hence all one-parameter ICC's are parallel. The lower asymptotes of Rasch curves are all zero indicating no guessing. Three-parameter ICC's usually have different slopes and may vary in their lower asymptotes.

Importance of Latent Trait Theory

Research in latent trait theory is significant because of the advantages the theory has over classical test theory. Lord and Novick (1968) draw a distinction between weak and strong true score theory. Latent trait theory is strong because many assumptions are made about data. Because classical test theory makes no assumptions about the items composing a test, generalizations from conventional tests can only be made to parallel forms. Scores on conventional tests are sample dependent because they are derived on a specific set of items and a specific sample of examinees. A consequence of this limitation has been that classical test theory has failed at providing solutions to a variety of measurement problems. Because of the particular choice of strong assumptions in latent trait theory, item and ability parameters can be estimated which are sample-invariant. The sample-free nature of latent trait parameters provides solutions to many problems handled inadequately by conventional testing methods. Test equating, detection of item bias, and tailored testing are easily

Figure 2
Three-Parameter Model ICCs

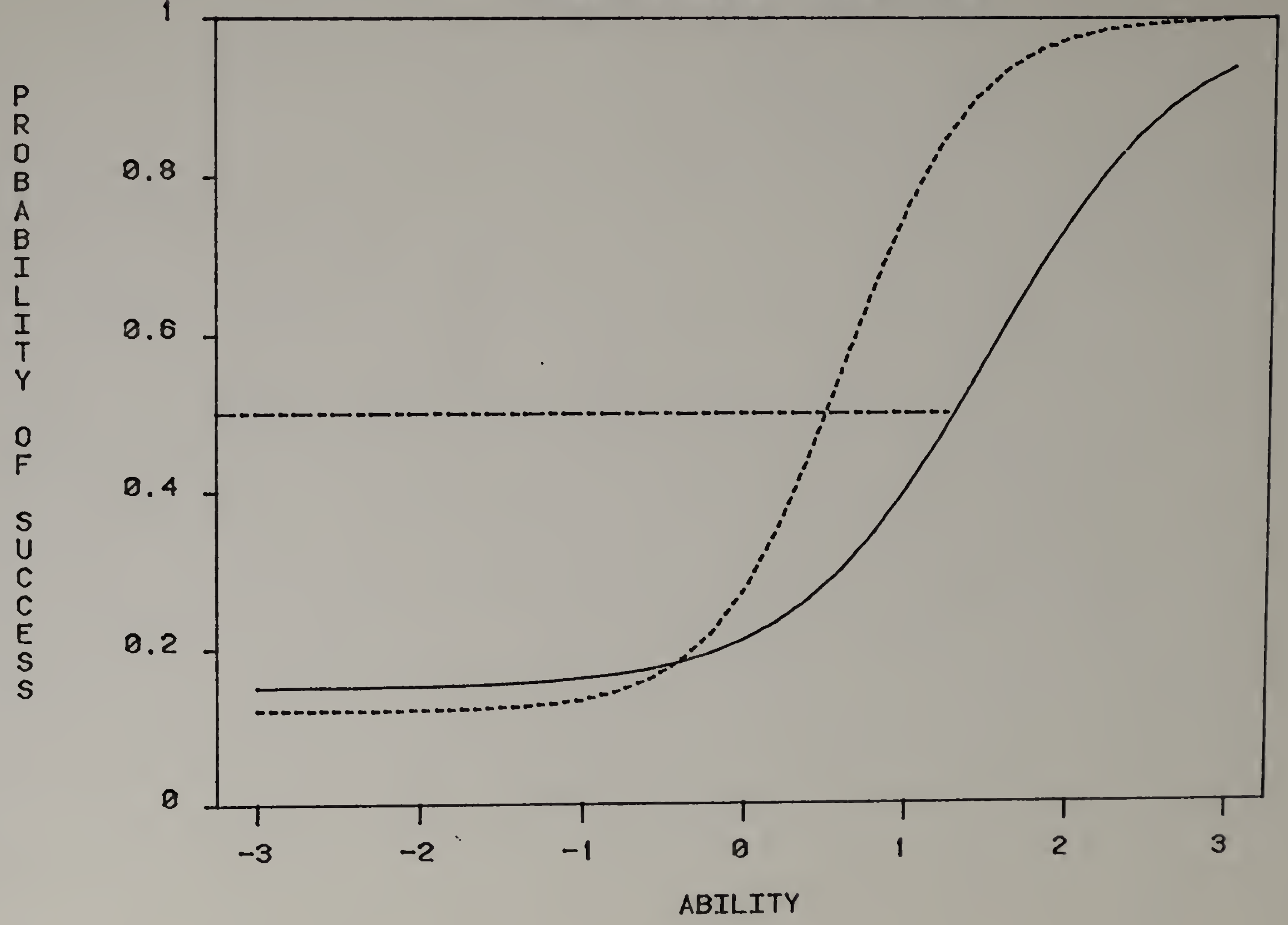
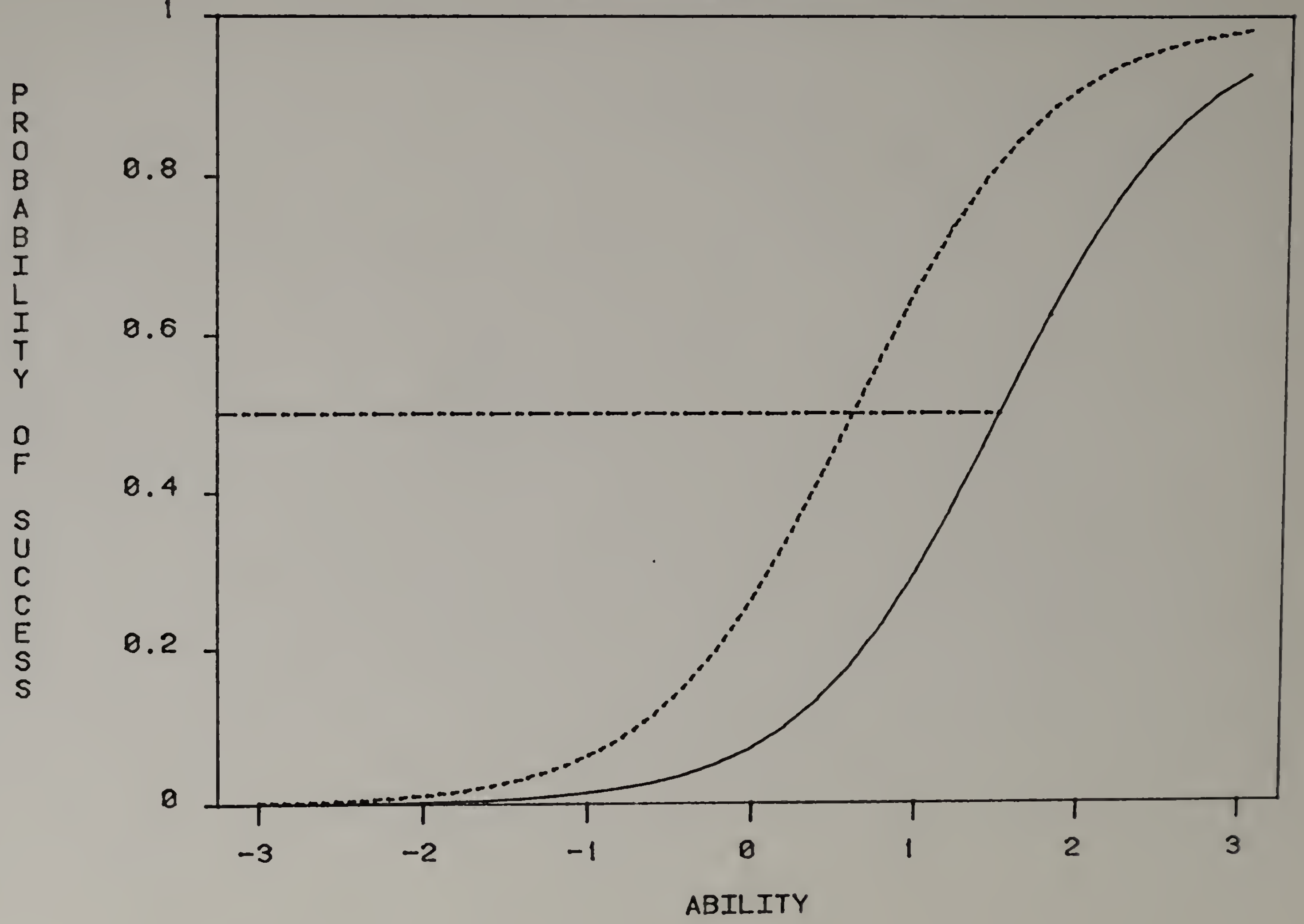


Figure 3
One-Parameter Model ICCs



managed with the results provided in latent trait theory. An excellent discussion of test equating with ICC theory was given by Marco, Peterson and Stewart (1979). Cowell (1979) provides another good source on equating. Pine (1976) provided a good description of application of latent trait theory to the study of item bias. Lord's (1980) recent book includes many applications of item response theory to test equating, study of item bias, and tailored testing. Hambleton et al. (1978), Lord (1977), and Wright and Stone (1979) provide reviews of many additional areas in which latent trait theory has been applied, including: test development, optimal scoring weights, mastery testing, handling omitted items, formula scoring, and item banking.

While latent trait theory provides flexible tools for solving measurement problems, the theory also has some limitations. One of these is that the assumptions made about data may be too strong for tests to be easily constructed to meet these requirements. A second shortcoming is that computation costs may be substantially higher than those incurred by conventional methods, thus prohibiting many applications. Another disadvantage of the theory is its mathematical complexity. Likelihood equations cannot be solved directly, and iterative solutions using Newton-Raphson techniques are required. Many restrictions are imposed in the process, especially when estimating item discrimination parameters and guessing. Another drawback to the theory is that the models are rather difficult for school personnel, students, and parents to comprehend. This lack of understanding has resulted in some resistance on the part of school systems to use the models in their testing programs.

Despite such drawbacks, Hambleton and Cook (1977) noted the acceptance of the theory by psychometricians and practitioners alike. In the summer of 1977, the Journal of Educational Measurement devoted an entire issue to latent trait theory. Two major review articles have appeared in the Review of Educational Research (Hambleton et al., 1978; Baker, 1977), and frequent articles on latent trait theory have appeared in Psychometrika and Applied Psychological Measurement. Sessions on latent trait theory have been very popular at the recent annual meetings of the American Educational Research Association. A major section of Lord and Novick's (1968) Statistical Theories of Mental Tests is devoted to latent trait theory and two books on the topic (Lord, 1980; Wright & Stone, 1979) have recently been published. Applications of the theory are numerous. These include the Key Math Test (Connally, Natchman, & Prichett, 1971), the Woodcock Reading Mastery Test (Woodcock, 1974), test equating at Educational Testing Service, and civil service examinations in the State of New York, to name a few. The theory has been applied to both achievement and aptitude tests for both norm and criterion-referenced testing situations.

Organization of the Study

This chapter has provided an introduction to latent trait theory and a discussion of its importance in solving measurement problems. The next chapter presents a review of tests for model assumptions and a discussion of issues revolving around model fit. Chapter III contains a description of the methodology for the study. The chapter includes a description of data sets, sampling information, methods for detecting violations in model assumptions, techniques for assessing model fit, and methods of comparison utilized in the study.

Model fit results are provided in Chapter IV. Descriptive information and conventional item statistics are presented for 25 data sets. Then the results of overall and comparative model fit are given. This is followed by a section containing correlations between fit and indicators of deviation from model assumptions. The next section examines precision of parameter estimates from short tests and small samples, and the final segment presents comparative cost information for the two models.

In the final chapter, the significance of the findings are discussed. The chapter includes a set of guidelines for latent trait model selection and a critique of the methodology used in the study. The study concludes with recommendations for related research.

CHAPTER II

ISSUES AND METHODS FOR TESTING LATENT TRAIT MODEL ASSUMPTIONS AND GOODNESS OF FIT

Methods for testing latent trait model assumptions and model fit are reviewed in this chapter. A discussion of previous comparative research studies which contrasts various methodologies for model fit is also included in the chapter. A presentation of issues concerning sample size and test length as they relate to parameter estimation concludes the chapter.

Tests for Latent Trait Model Assumptions

Unidimensionality (Local Independence)

Although multivariate extensions to latent trait theory have been developed, an assumption made for the models investigated in this study is that they are unidimensional. Unidimensionality means that all items in a test are designed to measure the same underlying trait or ability.

The assumption of local independence is predicated upon that of unidimensionality. The condition of local independence states that, "within any group of examinees all characterized by the same values θ_1, θ_2, ..., θ_k, the (conditional) distribution of the item scores are all independent of each other" (Lord & Novick, 1968, p. 361). Simply stated, for examinees of fixed ability, θ_k, success on any pair of items is uncorrelated. If the items in a unidimensional test were not stochastically independent, this would imply that among examinees of identical ability some would have a better chance of success than others for these items. If this were the case, then more than one ability would be needed to account for success on these items. This would clearly contradict the fact that the test was unidimensional. Goldstein (1980, p. 239) expressed doubt that the assumption of local independence can ever be met: "The assumption of local independence is such a strong assumption that it would be surprising if it were true other than in a few specially contrived circumstances." Goldstein asserted that local independence is not necessarily a logical consequence of unidimensionality and criticized the definition because it fails to account for the conditional distribution of other items. Despite this contention, the premise that local independence follows from unidimensionality is accepted in this study and a test of unidimensionality is considered sufficient for accepting that the condition of local independence has been met.

The viability of unidimensionality has been examined by a variety of techniques. Lumsden (1961) reviewed five methods for assessing unidimensionality in the test development framework. From item analysis techniques (magnitudes of item-test biserials), Loevinger's homogeneity criterion, the local independence criterion, Guttman's reproducibility criterion, and factor analysis, Lumsden concluded that the factor analytic method was superior. With this method, dimensionality is typically assessed by comparing the ratio of primary to secondary factor variances. Lord and Novick (1968) also advised that unidimensionality be investigated by factor analytic methods. Some factor analytic methods used in the latent trait context are described next. Issues arising from the use of factor analysis for assessing unidimensionality are also discussed in this section.

Bejar, Weiss, and Kingsbury (1977) used a method for determining unidimensionality which is attributed to Horn (1965). Test data were factor analyzed by a principal axis method. Then random item response data were generated which matched the test under investigation both in number of examinees and test length. The simulated data were factor analyzed and the resulting eigenvalues were compared to the test eignevalues with a graphic technique. With this procedure the number of test eigenvalues was reduced by the number of random roots which surpassed the actual roots in value. The method purportedly eliminates random factors attributed to correlations inflated by sampling fluctuation.

A number of other methods for factoring data have been used in assessing dimensionality by latent trait researchers. Principal components analysis was employed by Koch and Reckase (1978), principal factoring was done by Slinde and Linn (1979), maximum likelihood factor analysis was used by Bejar (1977), and a combination of principal components and principal axis common factor analysis was employed by Hambleton and Traub (1973). This last study used principal components analysis to determine the number of factors, and the principal axis solution gave estimated item-total biserial correlations. In the studies listed here, dimensionality was typically assessed by an eigenratio criterion, the ratio between the first and second latent roots (eigenvalues).

Lord and Novick (1968) suggested that tetrachoric correlations be employed in factor analysis for assessing dimensionality because of problems found to emerge with phi coefficients. The primary difficulty with phi coefficients is that they approach unity only when the marginal distributions of item scores are identical; otherwise phi coefficients are less than one even if a perfect relationship between items exists. Another problem with phi coefficients is that they vary with item difficulty levels and guessing and are therefore unstable across sample groups. The phi coefficient is a measure of relationship between two dichotomous variables and is easily obtained with the Pearson correlation formula. The tetrachoric correlation, which represents a relationship between two assumed latent variables scored dichotomously, is more appropriate for use in assessing dimensionality of the latent space, but is less easily obtained. First, there is little agreement in the literature on how to estimate tetrachoric correlations. A second drawback to the tetrachoric measure is that it makes the restrictive assumption that the underlying latent variables are normally distributed. Finally, when used for factor analysis, tetrachoric matrices are often singular and therefore impossible to invert. When such difficulties are overcome, the emergence of one common factor in a factor analysis of tetrachorics is a sufficient, but not necessary condition for assuming unidimensionality of the latent space (Lord & Novick, 1968, p. 382).

Christofferson (1976) has developed an alternate solution for factor analysis of dichotomous variables "based on marginal distributions of single and pairs of items." Generalized least square estimates are used, and a modified chi-square test examines the number of factors resulting as part of the procedure. More recently, Muthén (1978) offered a method for factoring dichotomous variables. Both methods overcome problems encountered with other methods for factor analyzing dichotomous variables, but unfortunately the methods have computational complexities that have not yet been satisfactorily resolved (Gustaffson, 1980).

Although there is agreement that factor analysis is the most adequate statistical tool for assessing dimensionality, the procedure is sample-dependent and may fail to determine that a set of items is unidimensional for all possible examinee samples. Lord (1980) emphasized the need for a statistical method for determining unidimensionality that is sample-invariant. A common sense technique that has been employed to assure unidimensionality for various examinee pools is the method of expert judgment, but this method can only be applied during test development. Hartke (1978) suggested a procedure for employing content experts to detect items which do not fit with an item set. The well-known Q-sort is another such technique.

Few alternatives to factor analysis for assessing unidimensionality have been suggested although Bejar (1980) described a new method which appears quite promising. In this method, unidimensionality is defined as a linear relationship between parameter estimates obtained from subsets of items arranged by content area and the full set of items. The item subsets are formed based on a priori hypotheses about content. "It follows that both sets of parameters should not differ unless one or more of the content areas is tapping a component which is unique to that content area" (Bejar, 1980, p. 284). The equivalence of item parameter estimates is verified with bivariate plots of the two sets of parameters. Unidimensionality is indicated when the points lie along a 45 degree line through the origin. Mean distances are computed to determine the extent of departure from the line. Bejar also provided a second technique for assessing dimensionality based on the intercorrelation of ability estimates resulting from tests with only single or multiple content structures. Both procedures avoid the time consuming and computationally complex procedures required for factor analysis of dichotomous variables.

Some latent trait researchers have argued that: "There are no separate adequate tests of unidimensionality. The direct test is the test of fit to the model" (Rentz & Rentz, 1978, p. 12). Gustaffson (1980) indicated that lack of unidimensionality cannot always be detected by certain tests of model fit. Gustaffson (1980) determined that the conditional likelihood ratio test of ICC slopes detects departures from unidimensionality only when slopes are equal for all items. The Martin Löf (1973) Person Characteristic Curve

(PCC) slope test, based on the chi-square method, was offered as an alternative test. This test also groups items together a priori by content and the subset parameters are compared to those obtained from the total test.

Prior research has shown that when the assumption of unidimensionality has been violated, data have not fit the latent trait models very well (Hambleton & Traub, 1971). Accurate tests for unidimensionality are therefore quite important both for doing research with and applying latent trait models.

Equality of Item Discriminations

The Rasch model is a special case of the Birnbaum (1968) logistic test model in which all item discrimination indices are equal. The viability of this assumption has been challenged: "The assumption that all item discriminations are equal is restrictive and substantial evidence is available which suggests that unless items are specifically chosen to have this characteristic, the assumption will be violated" (Hambleton et al., 1978, p. 26). Birnbaum (1968, p. 403) examined empirical data to explore this assumption and claimed that in most instances item discrimination indices varied considerably. In two sets of empirical data, Ross (1966) reported variations in item discrimination (a_g) from .47 to 1.99 (range 1.52) in one set, and from .30 to 1.97 (range 1.67) in the other. Using an approximate estimate of item discrimination, Hambleton and Traub (1973) reported three tests with discrimination ranges of .66, .74, and .69. Lack of fit to the Rasch model in these studies was attributed in part to heterogeneous item discriminations.

The Rasch model was reported to be robust to heterogeneous item discriminations in a study by Dinero and Haertel (1977). Model fit was explored for five data sets generated under two-parameter model assumptions. Five ranges of item discrimination were simulated from uniform and normal distributions with variances ranging from .05 to .25. Tests of fit to the one-parameter model were conducted with the Wright and Panchapakesan (1969) standard deviate. The conclusion of the study was that: "the lack of item discrimination parameters in the Rasch model does not result in poor calibration in the presence of varying item discriminations" (Dinero & Haertel, p. 589). They suggested that difficulties which might arise in ability estimation with the Rasch model due to the presence of non-homogeneous discriminations can be counteracted by increasing test length. Dinero and Haertel's results were based on rather small samples and should therefore be viewed with caution.

In a simulated comparison of the Rasch two- and three-parameter models, Hambleton and Cook (1978) found that the presence of heterogeneous item discrimination values had little affect on fit of data to the Rasch model. The criterion for model fit was the rank ordering of examinees by ability. The study used item discrimination ranges of zero, .81 to 1.43, .50 to .74, where the maximum range was selected to reflect the range of item discrimination values in the verbal section of the SAT. Earlier results by Hambleton (1969) demonstrated that increasing the range of discriminations simulated from a uniform distribution caused significant reduction in fit of data to the Rasch model. Similar results were obtained by Hambleton and Traub (1971) for data simulated from a normal distribution of ability. Discrimination parameters in the 1971 study ranged from .2 to .8 and were generated from a uniform distribution of a_g. Ranges of discrimination beyond .2 were found not to be tolerated by the Rasch model. This range corresponded to biserials in the .44 to .58 band. The variable results of these studies from those reported by Hambleton and Cook (1978) and Dinero and Haertel (1977) may be due either to the differing distributional assumptions or to the alternative methods for determining fit.

Few statistical procedures have been developed for testing equality of item discrimination indices. Panchapakesan (1969) provided a test for unequal item discriminations based on examination of probability plots of items. Departures from unity in slope indicated items with non-homogeneous discriminations. Birnbaum (1968) suggested a method based on magnitudes of conventional item discrimination parameters. Gustaffson (1980) suggested using Martin-Löf's chi-square test for explaining variable slopes in ICC's due to heterogeneous discriminations. Mead (1976) applied a residual approach to detecting a variety of deviations from Rasch model assumptions including non-homogeneous discriminations.

Wright (1977) claimed that it is impossible to estimate item discrimination. Wright showed that without severe restrictions, item discrimination indices tend to drift toward infinity. The BICAL procedure (Wright & Mead, 1978) for estimating parameters of the

Rasch model, includes a calculation of what is called "residual" item discrimination. The quantity is determined after data is fit to the model and is used as one of a number of statistics to assess lack of model fit. Such statistics are used to delete items with unequal discrimination values from a test. In a study of vertical equating with the Rasch model, Slinde and Linn (1979) used residual estimates of item discrimination to assess the equality of item discrimination indices. Item discrimination values within the range .80 to 1.20 were considered equal.

Guessing

The assumption that guessing does not occur is made with the Rasch and two-parameter models. When items are administered in an open-ended, or free-response, format this appears to be a reasonable assumption, but the assumption does not seem tenable for multiple choice tests. It seems plausible that examinees with little or no knowledge guess, unless cautioned otherwise, when presented with difficult items. Examinees with some or partial knowledge could make educated guesses by eliminating obvious erroneous choices, but examinees with no knowledge could select answers completely at random. In this latter case, the chance probability of obtaining a correct response is 1/C, where C is the number of response alternatives. Since correct responses, whether obtained by chance or through knowledge, are used to estimate ability, resulting ability estimates for low ability examinees tend to be too high unless some method is utilized to correct for guessing. Formula scoring, a common method for penalizing guessers (Lord & Novick, 1968, p. 307) is only applicable to conventional scoring, so other methods must be used to adjust for guessing in latent trait estimates of ability.

The lower asymptote of the three-parameter ICC, seen as a measure of the chance probability level, is used to correct ability estimates for random guessing. Since the lower asymptote of the Rasch ICC is zero, ability estimates are not adjusted for guessing. Guessing is viewed as an item characteristic in the three-parameter model, but there are other approaches which assume that guessing is an interaction of both item and person characteristics, and these approaches provide other methods for removing the influence of guessing from examinee ability. It is difficult to estimate guessing directly, although some attempts have been made. These are discussed next.

A straight forward method for estimating guessing, suggested by Lord (1970), was to examine visual plots of ICC's. Using this technique, Lord determined that lower bounds of ICC's for SAT items were typically below the chance level, 1/C. The efficacy of this method for estimating guessing is clearly dependent on the accuracy of the three-parameter model.

Urry (1974) developed a hueristic, or intuitive, method for estimating guessing which was based on regressing the percent of examinees passing an item on raw score, adjusted for the item under investigation. The lower left asymptote of the regression curve was taken as an estimate of guessing. Lord (1970, 1980) has also shown that when a sufficient number of examinees is used, item-test regression approximates the form of item-ability regression. Unfortunately, these methods are only accurate when thousands of examinees' scores are considered.

Jensema (1974, p. 74) criticized the approaches taken toward guessing in latent trait theory because: "A more basic question, which directly challenges the model, is whether the guessing parameter is constant over all levels of ability." Jensema postulated that guessing is a person- or sample-related characteristic or the product of some person-sample-item interaction. The Lumsden (1977) latent trait model includes a second person characteristic called "sensitivity" which reflects guessing among other person attributes. Approaches to guessing, based on these assumptions, have attempted to remove the effects of random guessing from ability estimates without estimating a guessing parameter, per se.

Waller (1974a, 1974b, 1976) outlined a procedure which can be applied to the Rasch and two-parameter models. "This is accomplished through a modification of the free response model removing those item-person interactions characterized by the item being too difficult for the person and therefore likely to invite guessing" (Waller, 1974b, p. 2). With simulated and empirical data, Waller found improvements in model fit when the ARRG (ability removing random guessing) procedure was applied. The ARRG method deletes items too difficult for an examinee and estimates ability from the remaining item subset. A number of passes through the data are needed to find items difficult for each examinee.

Since the guessing parameter is not estimated in the Rasch model, other methods are used to detect guessing. Mead (1976) examined guessing departures from the Rasch model by a method of standardized residuals. The residual between the expected and obtained ICC were plotted against the quantity, "θ_a - b_i" (the differential between ability and difficulty), and deviations from linearity in the plots signalled guessing. Gustaffson (1980) indicated that the Martin-Löf (1973) conditional likelihood ratio test can be used to detect irregularities in slopes of person characteristic curves (PCC) which purportedly indicate guessing.

With the exception of the Urry procedure, all of the methods for estimating or detecting guessing described above are based on the assumptions of latent trait theory. Conventional approaches, which use item difficulties, have also been suggested for estimating guessing. These approaches are severely limited because of the sample-independent of item difficulty, but do offer some means, independent quality of latent trait theory itself, for estimating guessing. Such methods estimate guessing by computing the item difficulties for hard items from low ability examinees' scores. The difficulty levels indicate the percentage of low ability examinees who passed items which were supposedly too difficult.

Results of studies which have assessed the effect of guessing on Rasch model fit have been somewhat variable. Ross (1966) used plots, similar to those suggested by Mead (1976), to visually inspect the impact of guessing on Rasch model fit. Although guessing was indicated on 11 out of 95 items, Ross claimed that the Rasch model demonstrated adequate fit to data. For the purpose of test score equating, Slinde and Linn (1978) concluded that guessing was not tolerated by the Rasch model. Gustaffson (1979) suggested that equating results might have been inadequate in Slinde and Linn's study because the presence of guessing would produce spurious correlations between item difficulty and discrimination. Upon analysis of the same data, Gustaffson determined that there were substantially high negative correlations between discrimination and difficulty in the data.

Speededness

Lord and Novick (1968) make a distinction between speed and power tests. A speed test is one based on an examinee's ability to answer as many items as possible within a fixed time limit. The score on a speed test depends on the rate of response. A power test is one with no time limit or a very liberal time limit. Latent trait theory does not apply to speeded tests, "but the theory can be still used to analyze answer sheets obtained in timed test administrations" (Lord, 1974, p. 248). Lord (1980) refers to consecutively omitted items at the end of a test as "not reached." Incomplete test response patterns may be attributed in part to speeded conditions.

If an examinee answers less than one-third of the items in a test, it can be assumed that the test was speeded for that examinee and, consequently, no ability estimate can be obtained. For other examinees, who answer a substantial proportion of items, it is possible to obtain ability estimates. LOGIST (Wood, Wingersky, & Lord, 1976) adjusts ability estimates from timed tests by a modification in the likelihood fucntion which was given by Lord (1974).

Because of the assumption of local independence, ability estimates can be obtained from any random set of homogeneous items designed to measure the ability. If an examinee answers items in the order presented, it can be assumed that those items answered constitute a small homogeneous set of items administered under non-speeded conditions for that examinee. Thus, ability can be estimated from the set of items reached, ignoring the set of items following the last item reached. When too few items are included in this set, there is a substantial loss of precision in estimation for the examinee. The procedure is also used to obtain ability estimates for examinees who omit intermittent items in a test.

There have been no empirical studies which have investigated the efficiency of ability estimation from timed tests, but Lord (1974) verified the maximum likelihood estimates of ability when data were characterized by omitted response patterns.

Significant research for detecting speededness has come from within the framework of classical test theory. Donlon (1978) provided an excellent review of these methods. No attempt was made in this study to evaluate speededness of tests because the LOGIST estimation procedure handled incomplete response patterns quite adequately. All tests used in the study were reported to have been administered as non-speeded tests.

Testing Model Fit

There has been little agreement among latent trait theorists concerning the measurement of model fit for latent trait models. There has been no consensus on the operational definition of fit, which has resulted in a variety of alternative methods for assessing fit, each based on a different definition.

Many view latent trait model fit in terms of item fit, but other researchers define fit based on the concept of test fit. Still other approaches define model fit in terms of ability or person fit. Separate methods for testing fit have evolved from each of these definitions.

Lord and Novick (1968, p. 383) described a generalized method for determining the adequacy of psychometric models. The procedure consists of the following steps:

1. Estimate the parameters of the model assuming it to be true;

2. Predict various observable results from the model substituting the estimated for true parameters;

3. Consider whether the discrepancies between predicted results and actual results are small enough for the model to be useful ("effectively valid") for whatever practical application the investigator has in mind; and

4. If the discrepancies found in step 3 are too large, then it may be useful to compare them with the discrepancies to be expected from sampling fluctuations.

Methods for testing model fit (step 3) have varied to some extent because they have been developed to explore fit in the context of rather different uses of models.

Birnbaum (1968) considered many of the statistical measures often employed for testing model fit to be unsound. For example, likelihood ratio test statistics which are asymptotically chi-square, were often assumed to be distributed as chi-squares despite the fact that they had been calculated on very small samples. For very large samples, most data are rejected by statistical tests even though fit may be adequate in a practical sense. Graphical techniques for inspecting fit have an element of subjective judgment, and few practical (non-statistical) measures of fit have been devised.

Rentz and Bashaw (1975) and Rentz and Rentz (1978) provided an excellent discussion of model fit. They viewed fit in terms of applications (which they called operations): test development and test analysis. They suggested that during test construction, the focus be on item fit, where "_item fit_ can be defined as the extent to which items can be characterized according to those antecedent conditions derived from the model's assumptions" (Rentz & Bashaw, 1975, p. 17). Based on the model's premises, graphic representations of items could be used to determine departures from the model, for example, inspection of plots of ICC's to evaluate the presence of non-zero lower asymptotes which would indicate guessing. During the test analysis phase, the focus switches to overall test fit: "_Test fit_ can be defined as the extent to which the test achieves those consequences specifiable from the concept of specific objectivity" (op, cit., p. 17). Specific objectivity, a concept originated by Rasch, means that ability and item difficulty can be estimated independently of one another. An instrument encompassing the quality of specific objectivity offers item-free person measurement and person-free object measurement. To measure test fit, Rentz and Bashaw suggested the use of a chi-square test based on the mean square fit criterion developed by Wright and Panchapakesan (1969). Rentz and Bashaw caution the user with respect to statistical tests of fit: "We do not believe that a routine application of some statistical test is adequate or even correct" (op. cit., p. 92). A second definition for test fit given in the same work is "the extent to which the test contains fitting items" (op. cit., p. 17). The mean square fit statistic can be applied to items individually or to the test as a whole.

Another approach to model fit is in terms of person fit. In this case, sample item parameter estimates are assumed to be true (i.e., representing population parameters), and fit is assessed in terms of person or ability parameters. Studies basing fit on this definition frequently have compared observed and predicted distributions of ability (or some monotone transformation of ability) by means of an approximate or exact chi-square test. Studies by Ross (1966) and Hambleton and Traub (1973) used this definition of model fit. The Wright and Panchapakesan mean square criterion is also used to evaluate person fit.

A simple technique for testing item fit is a graphic method (Rasch, 1960). This procedure entails regressing percents of examinees passing an item on raw scores (essentially an item-test regression). Departures from linearity in a plot can be statistically tested. Anderson, Kearney, and Everett (1968) developed a more sophisticated version of this test based on likelihood ratios. They applied the method to testing model fit of intelligence test items. Ross (1966) used a method for assessing guessing which was based on plots drawn on logistically scaled probability paper.

Other visual methods for exploring model fit include comparisons of frequency distributions (observed verses predicted) of ability, raw scores, true scores, or sufficient statistics. Most studies of model fit have included some graphic test. Cumulative distributions (item or test characteristic curves) can similarly be visually inspected for model departures. Hambleton (1980) described a method for comparing a predicted ICC with an actual ICC. For each item, the observed ICC is found by plotting the examinee performance level (i.e., percent of examinees obtaining a correct response) for various levels of ability. The predicted ICC is based on the estimated item parameters. The plot is explained by a second figure which shows the positive and negative discrepancies between the two ICC's. The magnitude of these discrepancies could be calculated using a squared distance formula. Gustaffson (1980) reports a graphic method similar to those discussed in this paragraph for application to fit of the Rasch model.

The primary statistical test used to measure model fit has been a chi-square test based on mean square deviations or likelihood ratios. Wright and Panchapakesan (1969) developed a widely used statistic for testing fit to the Rasch model.

For each item a standardized deviate is formed between the predicted and observed item score. The standardized deviate is usually expressed in the relative frequency metric. Deviates are summed across persons (or score groups) to obtain an approximate chi-square statistic of item fit, or across items to obtain an approximate chi-square statistic for person fit, or across both to measure overall test fit.

For score group i and item j the standard deviate is given by:

$$z_{ij} = ((f_{ij}) - E(f_{ij})) / V(f_{ij})^{\frac{1}{2}} \quad , \qquad [4]$$

where f_{ij} is the observed frequency of examinees in score group i who answered item j correctly; $E(f_{ij})$ is the expectation of f_{ij}; and $V(f_{ij})^{\frac{1}{2}}$ is the standard deviation of f_{ij}. Since f_{ij} has a binomial distribution with parameter P_{ij} (the probability of a correct response), the expectation is found by taking the mean of the binomial:

$$E(f_{ij}) = r_i P_{ij} \quad , \qquad [5]$$

where r_i is the number of examinees in score group i. The variance of the binomial is given by:

$$V(f_{ij}) = r_i P_{ij}(1-P_{ij}) \quad . \qquad [6]$$

The z_{ij} are normally distributed with mean zero and standard deviation one and can be summed across items, or people, or both. With sufficiently large sample size, the sums of the standard deviates approximate chi-square statistics. The total test chi-square is given as:

$$X^2 = \sum_{i=1}^{n-1} \sum_{j=1}^{n} z_{ij}^2 \quad , \qquad [7]$$

which has (n-1) (n-2) degrees of freedom, where n is the number of items, and n-1 is the number of score groups.

An alternate formulation of the standard deviate was given by Wright and Stone (1979) in which the deviate is formed between the actual and expected item score, u_{ij}, where $u_{ij}=1$ when item j is correct, and $u_{ij}=0$ when item j is incorrect. A standardized residual, found for each person-item combination given by:

$$z_{ij} = (u_{ij} - P_{ij}) / (P_{ij} (1-P_{ij}))^{\frac{1}{2}} \quad , \qquad [8]$$

is distributed normally with mean zero and unit variance. The sum of squared residuals, across items, persons, or both, approximates the chi-square distribution, or alternatively, the mean square can be found as:

$$v = \Sigma z^2 / df \quad , \qquad [9]$$

which is an approximate F-statistic with "(N-1) (n-1)/N" degrees of freedom for person fit or "(N-1) (n-1)/n" for item fit. Since the item score, u_{ij}, can only assume the values 0 or 1, equation [8] reduces to exp(θ-b) for a correct response or exp(b-θ) for an incorrect response.

George (1979) has shown that meaningless results can be obtained with the mean square statistic when samples employed are too small. Under these conditions, the chi-square test is inappropriate since distributional assumptions of the test are not met. Application of the test to small samples results in significant errors in interpretation. George also notes that for very large samples, the chi-square test, like other statistical tests, rejects all data even though fit may be more than adequate from a practical point of view. A generalized version of the chi-square test which can be used with all of the latent trait models, was offered by Ross (1966) and later by Hambleton and Traub (1973). The procedure uses estimated item parameters to predict distributions of number-correct scores or weighted number-correct scores. These are compared with actual distributions of number-correct scores using a standard chi-square test:

$$X^2 = \sum_{i=1}^{k} (f_i(o) - f_i(e))^2/f_i(e) \quad , \qquad [10]$$

where $f_i(o)$ are observed frequencies for score group i and $f_i(e)$ are expected frequencies for score group i, and the summation is across k score groups. Alternatively, the Kolmogorov-Smirnov statistic, which makes fewer assumptions than the chi-square, can be employed to compare actual and predicted score distributions.

Likelihood ratio tests for the normal ogive model were devised by Bock and Lieberman (1970) and for the Rasch model by Anderson (1973). Anderson's test assumes that parameters were estimated by a conditional maximum likelihood approach. Likelihood ratio statistics are well-suited for assessing differences in fit due to alternative models and for testing parameter invariance. The test statistic used with likelihood ratios is approximated by a chi-square for large samples. Versions of the test have also been formulated for estimates based on the unconditional maximum likelihood method.

A more recent development in likelihood ratio tests for the Rasch model is found in Anderson and Masden (1977).

The likelihood ratio tests cited above were all devised for use with a single model. Waller (1980) has formulated a likelihood ratio test which is claimed to be able to test deviations in fit for the Rasch, two-, and three-parameter models. The test is an extension of Bock's (1972) likelihood ratio test for the nominal response model. A likelihood ratio is formed based on r item parameters, and another likelihood ratio is formed based on a subset s of the r parameters from a model with fewer parameters. The log likelihood of the difference (r-s) is formed. Waller claims that the log likelihood of the difference is distributed as a chi-square statistic. The method is based on rank ordering examinees by ability and grouping them into i fractiles, or ability groups. The same number of ability groups are formed for each model. Then, for each item, the test statistic is given by:

$$l_j = C + \sum_{i=1}^{t} (r_{ij} \log P_{ij} + (N_i - r_{ij}) \log (1-P_{ij})) , \qquad [11]$$

where P_{ij} is the probability of item success, N_i is the number of examinees in fractile i, r_{ij} is the number of examinees in group i who obtain a correct response on item j, and:

$$C = \log (N_i! / r_{ij}! (N_i - r_{ij})!) \quad . \qquad [12]$$

Another test based on likelihood ratios is the binomial test offered by Divgi (1980). Divgi also claims that his test is applicable to all of the latent trait models of interest. The method purportedly detects model fit even when the parameters from a specific sample may have been estimated in error. The procedure is given as follows: first item and ability parameters are estimated for the two models of interest. Then, for a validation sample having approximately 100 examinees, maximum likelihood ability estimates are obtained based on the two sets of item parameters. Two likelihoods are calculated for the observed patterns of response. If $\underline{P}$ is the proportion of cases for which calibration by one model provides better fit, then the test is based on the null hypothesis, H_o: $\underline{P}$=.5. Since $\underline{P}$ is a binomial with mean .5, the test results in exact probabilities for $\underline{P}$. Divgi notes that when more than 50 examinees are in the validation sample, the normal approximation for the binomial can be used. The validation sample in this method is selected to represent a specific population of interest. The results of the test supposedly demonstrate model fit in terms of specific applications.

A variant approach to model fitting is one designed by Mead (1976). This technique uses the standardized residual between the actual and observed frequencies of examinees for item i in some score group j, as given earlier in equation [4]. The residual statistics are plotted against the quantity (θ-b) for a visual test of fit or can be used to perform t-tests or analyses of variance between models. A method bearing similarity to Mead's method is one used by Koch and Reckase (1978). In this case, the deviate is formed between the item response, u_g (zero or one) and the expected probability of item response, $P_g(\theta)$. The deviate is given as:

$$MSD = (\sum_{j=1}^{N} (u_g - P_g(\theta))^2/N \quad . \qquad [13]$$

The authors claim that the statistic is normally distributed, but no empirical evidence exists to support this claim. Until such evidence exists, the statistic should be used with caution. The MSD statistic was developed to replace frequently used chi-square test statistics which are inappropriate for small samples.

Lord (1970) provided a method for model fitting that has intuitive appeal. In this method, the ICC is estimated by two methods and the resulting curves are compared. One method assumes no special mathematical form for the ICC. It is based on the regression of item score on estimated true score (minus the item in question). The second estimate is an ICC from one of the latent trait models. Ability is transformed to the true-score metric for comparison by visual or statistical means. The method can be used to test model fit and to detect parameter invariance. Lord (1974) utilized this method to compare two maximum likelihood estimates of ability for data with omitted responses.

Gustaffson (1980) described a number of new methods for testing fit of the Rasch model when parameters have been estimated with the conditional maximum likelihood (CML) approach. Because Gustafsson has overcome some of the problems in CML estimation, it is claimed that the method can be practically applied more easily. One of the statistics discussed by Gustafsson is the Anderson (1973) method noted earlier. He suggests that the method is primarily appropriate for detecting deviations in slopes of ICC's. Another method is attributed to Martin-Löf (1973), who devised two tests of fit. One test statistic is asymptotically equivalent to the Anderson likelihood ratio test, but is constructed using frequency data for persons in different raw score groups. The second test statistic was designed to detect differences in Person Characteristic Curves (PCC) such as those described in the model by Lumsden (1978). The statistic uses the maximum of the log likelihood function, where twice the log function has been shown by Martin-Lof to be asymptotically distributed as a chi-square. The test can be applied to detecting person differences including item bias, speededness, guessing, and person sensitivity, which is defined as varying person reliabilities. Martin-Löf also developed a measure called redundancy which supposedly provides an absolute index of model fit and can be applied when there are large samples.

When item and ability parameters are known, such as in simulation research, a number of additional techniques for testing model fit can be employed. Lord (1974) and Hambleton and Cook (1978) used correlational analysis to compare estimated with true parameters. Both Pearson and Spearman techniques have been applied. In addition Hambleton and Cook formed the average absolute difference (AAD) between estimated and true parameters to explore fit.

Studies of Comparative Model Fit

This section reviews studies which have compared the Rasch model with the two- and three-parameter logistic models. One of the earliest empirical comparisons between the Rasch and two-parameter models was reported by Hambleton and Traub (1973). This study was based on the earlier work of Hambleton (1969). Comparisons between the two models were made for the verbal and math subtests of the Ontario Scholastic Aptitude Test, and for the verbal section of the SAT. The method of comparison employed a chi-square statistic to determine deviations in predicted from observed distributions of weighted raw scores, which are the sufficient statistics for ability. Since computerized techniques for parameter estimation had not been developed at the time of these studies, approximate solutions were used for obtaining item parameter estimates. These approximations requiredthat ability be normally distributed which added an additional restriction to the data. Weighted raw scores were constructed using estimated parameters as weights. With this technique, Hambleton and Traub found that when item discriminations were heterogeneous, the two-parameter model showed improved fit over the Rasch model for the three tests.

Koch and Reckase (1978) and Reckase (1978) compared fit of the one- and three-parameter models for both empirical and simulated data. Real data included a vocabulary test, an aptitude test, and four classroom achievement measures. The results of both studies indicated improvement in model fit when additional parameters (item discrimination and guessing) were considered in the model equations. A mean square deviation (MSD) statistic was employed item by item to detect fit. The MSD statistic reflects the deviation in the item response, u_{ig} (scored zero or one), from the predicted item response, $P_{ig}(\theta)$, which is a probability ranging from zero to one. T-tests were conducted between models based on average MSD statistics.

Because sampling properties of the MSD statistic are unknown, the conclusions of this study must be viewed cautiously.

Rentz and Rentz (1978) fitted the Rasch model to aptitude, achievement, and criterion-referenced test data. The study used Wright and Panchapakesan's (1969) fit statistic. Although the model was not compared to others, the study showed that the Rasch model can fit many diverse forms of tests.

Hambleton and Cook (1978) made comparisons in fit for the Rasch, two- and three-parameter models using simulated data. With simulated data, comparisons can be made between estimated and true parameters (from which the data was generated). Measures of fit were based on Pearson and Spearman correlations and the average absolute difference (AAD) between true and estimated ability. Hambleton and Cook found significant improvements in model fit at the lower end of the ability continuum for the more general models especially for tests which had few items. Although it is reasonable to anticipate improved fit to data when a model is less restrictive, unfortunately studies involving simulated data do not provide a check on the adequacy of models for describing the real world.

Douglas (1980) compared Rasch model equating with equating based on the two- and three-parameter models. Douglas used parameter estimates to predict raw scores and then compared estimated to true values with bivariate plots. For the purposes of equating, Douglas found Rasch equatings to be better and more consistent than those using the two-parameter model. The three-parameter model was not considered in the final comparison because lower asymptotes were judged to be unacceptable.

A more encompassing comparison of latent trait equating methods was done by Marco, Peterson, and Stewart (1980). An anchor test method was used to equate verbal SAT scores. The Rasch and three-parameter ICC equating methods were compared to each other and to a variety of additional methods. Contrary to the results of Douglas (1980), this study showed that the three-parameter ICC method gave superior results to all other methods of equating, although the authors pointed out that the SAT data may violate Rasch model assumptions since opportunity to guess on the test is considerable.

Issues Relating to Sample Size and Test Length

There has been some controversy in the latent trait area concerning the number of examinees and items that are required for obtaining precise ability and item parameter estimates. Within the context of the Rasch model, Wright (1977, p. 224) purported that "calibration sample sizes of 500 are more than adequate" and goes so far as to say that useful information can even be gained on samples of 100 examinees. In Wright and Stone (1979), an example of calibrating with the Rasch model is repeated throughout the text using only 35 pupils. Whitely and Davis (1974) and Whitely (1977) disagreed with Wright, and contend that samples of at least 1000 examinees are needed to effectively use Rasch techniques. For the three-parameter model, Lord (et al., 1976) stated that precise item parameter estimates cannot be obtained with fewer than 1000 examinees.

Parallel questions to these have been raised regarding test length. The Rasch model is often used with 20 to 30 items. Lord (op. cit.) advised that at least 40 items be used to estimate ability with the three-parameter model.

In conventional measurement, the reliability of a test is closely tied to test length. In theory, a test with an infinite number of items would be perfectly reliable. In latent trait theory, test length has a bearing on <u>precision</u> of estimation. Tests must be of sufficient length to obtain precise ability estimates. Consequently, precision of item parameter estimates is a function of sample size. As an alternative to reliability coefficients, latent trait theory uses the information function as a measure of precision (Hambleton, 1979; Lord, 1980).

Sample size has frequently been varied in simulation studies in the process of exploring other issues, but Ree (1980) was the only study designed to systematically assess the effects of varied sample size on item parameter estimates. Ree generated samples of 250, 500, 1000, and 2000 to explore effects on three item parameter estimates in the context of linear equating. Good estimates of difficulty were obtained from 250 examinees, but 1000 examinees were needed to get good discrimination estimates. Although little variation in guessing estimates was observed (as measured by average absolute differences between estimates from different sample sizes), correlations of guessing estimates across sample sizes were negligible. Ree's results suggest that 2000 examinees are needed to estimate guessing so that a sufficient number of low ability examinees are represented in the sample.

Hambleton and Cook (1978) found surprisingly small gains in model fit for both the Rasch and Birnbaum models when test length was extended from 20 to 40 items. This study, like the previous one, was based on simulated data.

Little other information has been reported on precision of latent trait parameter estimates. This is an area which has not been adequately researched.

Summary

This study was undertaken because evidence about latent trait model fit, especially for real data has been inconclusive. The study differs significantly from previous research in a number of ways. First, the data utilized were real and not simulated. Secondly, ability and item parameter estimates were determined using sophisticated computer methods rather than by approximation. Thirdly, fit statistics with known sampling properties were utilized to make comparisons, and fourthly, twenty-five data samples were employed, more than twice the number used in any previous comparative research study. Finally, the current study used three measures of fit to substantiate the results, rather than a single measure.

CHAPTER III

METHODOLOGY FOR COMPARING LOGISTIC LATENT TRAIT MODELS

Overview of the Design

Item response data for twenty-five tests were obtained from a variety of sources to make comparisons between Rasch and three-parameter model fit to empirical data. Data were from multiple choice tests designed for measurement of achievement or aptitude. The tests covered a broad range of contents, formats, levels, and examinee sample characteristics. Five of the tests were used to explore the effects of sample size and test length on precision of latent trait parameter estimates, while all 25 tests were used to explore questions of model fit.

The tests were scored for conventional (right-wrong) and latent trait (right-wrong-omitted-not reached) analysis, and samples of 1000 examinees were drawn from each test data set. An item analysis and a factor analysis were performed with each test. Conventional item statistics were used to roughly approximate the degree to which each test deviated from the latent trait model assumptions. Then, item and ability parameters were estimated under the conditions of each model. These estimated parameters were substituted for true parameters to make predictions about number-correct score distributions from each model. Predicted distributions were compared with observed raw score distributions by statistical and graphical procedures. Measures of fit were correlated with indices of violation of model assumptions to examine model robustness. Then, precision of parameter estimates from small samples and from short tests were explored with correlational analysis. Finally, computer times and costs for parameter estimation by the two models were compared.

Selection and Sampling of Data

Data Selection

Twenty-five cognitive data sets were used in this study. Response data were obtained from test publishers, local school systems, and statewide testing agencies. The following criteria were utilized to select data:

1. Tests were designed to measure cognitive skills. Both aptitude and achievement tests were used. Most of the tests were normed-referenced, but some criterion-referenced measures were also used.

2. Tests were recognized, quality tests which had been constructed by well-known testing agencies.

3. Test items were multiple choice in format with only one correct response per item. Also, the number of response alternatives per item was consistent for all items in a test.

4. All tests exceeded 40 items in length.

5. The sample size per test was 1000 examinees. Most samples were over 3000, and so 1000 examinees were drawn randomly.

6. Tests were not administered under speeded conditions (although it was assumed that reasonable time limits had been imposed).

Description of Data Sets

Table 1 provides information on the 25 tests analyzed in the study. Additional features are described below:

Stanford Achievement Test (STAN).—The Stanford Achievement Test, published by Psychological Corporation, has ten subtests covering verbal and quantitative skills. The test has been used extensively in Rasch model studies. One thousand examinees were randomly drawn from a nationwide sample of 4000 examinees who had been administered the test in 1973.

Scholastic Aptitude Test (SAT).—Random samples of 1000 were drawn from two samples of 3000 which constituted the 1974 ICC equating samples for the SAT. The SAT, published by Educational Testing Service, is the primary aptitude measure used for college admission. The two samples did not include examinees who had not completed the test.

California Test of Basic Skills (CTBS).—The CTBS is a general achievement test published by McGraw Hill. Two subtests, math comprehension and vocabulary, were used in this study.

California Achievement Test (CAT).—The CAT is another nationally standardized achievement test. Data used here were from the 1974 Anchor Test Study (Rentz & Bashaw, 1975), an equating study of seven well-known reading achievement tests. Verbal and comprehension subtests were used.

Iowa Test of Basic Skills (ITBS).—The ITBS data were also obtained from the 1974 Anchor Test Study. Both comprehension and verbal subtest data were utilized in this study.

Table 1

Features of 25 Data Sets

Test	Number of items (n)	Number of choices per item	Sample[1] Size	Form	Level	Grades
STAN Word Study	50	5	4160	Int.	A	4-6
STAN Reading	71	4	4160	Int.	A	4-6
SAT Verbal #1	85	5	3000	N/A	N/A	11-12
CTBS Math Comp.	48	5	1112	S	4	10
CTBS Vocabulary	40	4	1112	S	4	10
STAN Vocabulary	50	4	4160	Int.	A	4-6
STAN Science	60	4	4160	Int.	A	4-6
STAN Language	80	4	4160	Int.	A	4-6
STAN Soc. Studies	54	4	4160	Int.	A	4-6
STAN Listening	50	4	4160	Int.	A	4-6
STAN Spelling	60	4	4160	Int.	A	4-6
STAN Math Applic.	40	5	4160	Int.	A	4-6
STAN Math Concepts	35	4	4160	Int.	A	4-6
ICRT	16	4	935	269	N/A	4-5
SAT Verbal #2	85	5	3000	N/A	N/A	11-12

[1] 1000 examinees were randomly drawn and used in the study.

Table 1 (continued)

Test	Number of items (n)	Number of choices per item	Sample Size	Form	Level	Grades
Georgia Regents #1	70	4	8026	13	N/A	college
Georgia Regents #2	70	4	7754	14	N/A	college
Georgia Regents #3	70	4	8092	15	N/A	college
CAT Vocabulary	40	4	8522	A	4	6
CTBS Vocabulary	38	4	8561	5	10	4
CAT Comprehension	45	4	8522	A	4	6
ITBS Comprehension	68	4	8561	5	10	4
Rasch-Environment	34	a	486	9	4	12
Rasch-Citizenship	30	a	513	6	5	12
Rasch-Career	39	a	511	3	2	12

[a] Unknown/variable - 2-PAR model estimated.

Georgia Regents.—The Georgia Regents are a minimum competency test required for graduation from Georgia state colleges. Three forms of the verbal subtest were used in the study.

Rasch Tests.—These Rasch constructed tests from the Atlanta Assessment Program had been administered statewide to twelfth graders. The tests were criterion-referenced and covered a dozen goal areas, but only three subtests were selected because too few examinees (only 500) were available on any subtest.

Individualized Criterion-Referenced Test (ICRT).—Criterion-referenced tests in reading and math were obtained from Educational Progress Corporation in Oklahoma. Sixteen items had been matched to each objective. Sufficient data were only available on reading book 269 and so it was the only one used in the study.

Sampling and Scoring Conventions

Samples of 1000 examinees were drawn for each test. The SPSS (Nie et al., 1975) subprogram SAMPLE was used for this purpose. Data were scored as 1=right and 0=wrong for conventional item analysis and factor analysis, and rescored as 1=right, 0=wrong, 10=omitted, and 11=not reached for latent trait parameter estimation. Scoring for latent trait parameter estimation considered that consecutively omitted items at the end of a test had not been attempted by the examinee due to time constraints.

Method for Testing Model Assumptions

Unidimensionality

A unidimensional test is one in which all of the items measure a single underlying trait or ability for all examinee populations of interest. An operational description of unidimensionality arising from the foregoing definition is that only one common factor is obtained from a factor analysis of a test. Although there are numerous ways to evaluate the outcomes of factor analysis, this study used the eigenratio as the criterion of unidimensionality.

For each test the total available set of data were scored in a conventional manner and tetrachoric correlation matrices were obtained with SPSS subprogram TETRACHORIC. For comparative purposes, phi coefficients were also obtained. Since the results of factoring the phi coefficients were essentially identical to those obtained with the teterachorics for five tests, the procedure was eliminated for the remaining 20 tests. The matrices of tetrachorics were factor analyzed with SPSS subprogram FACTOR using a principal components procedure, as an approximation to the common factor solution.

Eigenvalues, first factor variances, and the number of factors with eigenvalues over 1.0 were recorded. Any of these measures might have been used to assess dimensionality, but they all bear relationships to test length. Instead, the eigenratio between the first and second eigenvalues was used to assess dimensionality because it indicates the dominance of the first factor over other factors. High eigenratios indicated unidimensional tests. Tests were rank-ordered based on eigenratios with a rank of 25 signifying the most unidimensional test.

Equality of Item Discrimination Indices

The point-biserial was used as an approximate measure of item discrimination. Point-biserials were calculated with SPSS subprogram RELIABILITY which provides the correlation between the item and the total score adjusted for the item under investigation. Point biserials were considered equal when they were within a .1 confidence band around the mean point-biserial for a test ($\bar{r} \pm .1$). This interval was selected following considerable experimentation, and provided the greatest contrast between tests with homogeneous and heterogeneous item discrimination indices. A FORTRAN program written by the author was used to analyze equality of point-biserial correlations. The point biserials were transformed by a Fisher z so that their sampling distribution would be normal. The mean and variance of the transformed correlations were obtained for each test and these were transformed back into the original metric. A count was made of the number of point-biserials within the "$\bar{r} \pm .1$" confidence region and this number was converted to the percentage of items on a test having equal item discriminations. A high percent indicated a test with nearly equal item discriminations. Tests were rank-ordered based on these percents. High ranks indicated tests with homogeneous item discrimination indices. The standard deviation of point-biserials for a test was recorded as an alternate measure of equality of item discrimination.

Guessing

The guessing parameter of the three-parameter model has no parallel in conventional item statistics. A rough approximation to guessing was found by obtaining the conventional item difficulty for low ability examinees on hard items. A hard item was defined as an item answered incorrectly by more than two-thirds of the sample of 1000. A low ability examinee was defined as an examinee in the lowest decile of the sample. Clearly this rough approximation to guessing lacked the sample-invariant characteristics of latent trait parameters. Item "hardness" was assessed based on difficulty levels in the total sample of examinees. Low ability was judged from number-correct scores. The lowest deciles contained approximately 100 examinees for each test. As a measure of guessing, item difficulties were recomputed on hard items for the lowest ten percent of the sample. These difficulties indicated the percent of low ability examinees who scored correctly on items which were purportedly too difficult. It was assumed that correct answers for these items had been obtained by random guessing. Guessing was only estimated on tests which had more than three hard items. Little variability between tests was observed with this measure of guessing, so rank ordering of tests did not provide very useful information.

Method for Testing Model Fit

Parameter Estimation

For each data set, ability and item parameters were estimated under the assumptions of the Rasch and three-parameter models using an unconditional maximum likelihood approach (UML). In the UML approach, the likelihood function is solved simultaneously for item and ability estimates. Parameter estimation was accomplished with the LOGIST computer program (Wood, Wingersky, & Lord, 1976) with a modification in the likelihood function for handling omitted and "not reached" items (Lord, 1974). For item responses, $u_g=1$ or $u_g=0$, with the probability of a correct response given as: $P_g(\theta_k)$, and where $Q_g=1-P_g$, the likelihood function is given by:

$$L = \prod_{k=1}^{N} \prod_{g=1}^{n} P_g(\theta_k)^{u_g} Q_g(\theta_k)^{1-u_g} \quad , \qquad [14]$$

Log L is differentiated with respect to the unknown parameters θ_k, b_g (and a_g, c_g in the three-parameter case) resulting in "n+N-2" simultaneous equations for the Rasch model and "3n+N-2" simultaneous equations for the three-parameter model. A modified Newton-Raphson procedure is used to solve the equations since a direct solution is impossible due to the number of unknowns. Initial estimates for parameters were computed from conventional item statistics.

On five tests, ability had been restricted to the range -4 to +4 to assure convergence, but Lord (personal communication, 1979) suggested that the limits were unnecessary and inappropriate, and they were removed from subsequent parameter estimations for 20 tests without convergence problems resulting. Overall, .01 percent of abilities were outside of these limits. Item discrimination indices were limited to the range .01 to 2.0 and rarely exceeded these limits. Although no restrictions were placed on magnitudes of item difficulties, the allowable percentage change between stages of the estimation was restricted.

For the Rasch model a convergence criterion of .02 percent was used. In the three-parameter case, the convergence criterion was successively reduced from 200 to .02 percent across stages. The convergence criterion provided accuracy up to the third position after the decimal for both item and ability parameters. For estimation of Rasch abilities and difficulties, item discrimination was set to 1.0 and guessing to 0.0 and held constant.

Estimation Procedure

Examinee samples of 1000 were drawn and scored with LOGIST scoring conventions for the 25 tests. The estimation procedure began by editing data: examinees with zero or perfect scores were removed since abilities cannot be estimated for these examinees (such estimates would be infinity). Examinees who answered less than one-third of the items were also eliminated since it was assumed that the test may have been speeded for these individuals. Items answered correctly or incorrectly by all examinees were then removed since they provide no information for estimation of ability.

The next step of the procedure was to compute initial estimates of item and ability parameters from conventional item statistics. The iterative estimation process for solving the likelihood equations takes the initial estimates as starting values. Table 2 shows the LOGIST control parameters used in this study. Additional control parameters, not shown in the table, retained the default values set in LOGIST.

Procedure for Prediction of Number-Correct Score Distribution

In simulation studies of fit, estimated ability can be directly compared with true ability, but with real data ability is unknown. Instead, models are used to predict some observable characteristic of data by substituting estimated parameters for true values. The predicted data can be compared to the actual data. Number-correct score distributions were predicted in this study. Lord (1980) demonstrated the relationship between ability and the probability density function of number-correct scores, where the number correct score, X, is given by:

$$X = \sum_{g=1}^{n} u_g \quad , \qquad \text{and} \qquad [15]$$

"u_g" equals zero or one and is a binary scored random variable for item g. For a fixed level of ability, θ_k, the frequency distribution of number-correct scores is a generalized binomial. The mean of the conditional distribution of raw scores for fixed ability θ_k, is given by:

Table 2

LOGIST Control Parameters for Estimation Used in the Study

	Model		
Parameter	1-PAR	3-PAR	Description
LITTLN	variable	variable	test length
N	variable	variable	number of examinees in sample
NCH	not set	variable	number of response alternatives
NOPARM	2	2	change default options
MAXST	10	30	maximum stages
NC	1	0	model (1=1-PAR, 2=2-PAR, 0=3-PAR)
IC	2	0	c_g's (2=set c_g's to zero, 0=estimate c_g's)
ITONE	6	6	number of iterations per stage
INTHET	-1	-1	do not limit abilities
MATPD	9	0	process (9=regular, 0=automatic—only for 3-PAR)

$$\mu_{(X|\theta)} = \sum_{g=1}^{n} P_g(\theta_k) \quad , \qquad [16]$$

and the standard deviation is given by:

$$\sigma_{(X|\theta)} = (\sum_{g=1}^{n} P_g(\theta_k)\ Q_g(\theta_k))^{\frac{1}{2}} \quad , \qquad [17]$$

where $P_g(\theta_k)$ is the probability of a correct response and its form is based on the model. The quantity Q_g is equal to "$1\text{-}P_g(\theta_k)$." The mean and standard deviation of the conditional distribution of number-correct scores can be used to generate number-correct score distributions by forming standard deviates for each number-correct score. The total predicted number-correct score distribution is found by summing across the conditional distributions for each level of ability. The procedure was replicated to obtain predicted number-correct score distributions for the Rasch and three-parameter models.

Figure 4 demonstrates the relationship between ability and the conditional distribution of number-correct scores. Seventeen discrete levels of ability were found for each test by dividing the ability distribution into 17 groups each .5 wide and taking the midpoint of the group as the estimate of ability. The number of examinees in each group was determined from abilities estimated by the Rasch or three-parameter model. The lowest ability group was bounded by -4.25 and the highest ability group was bounded by 4.25. Experimental findings from five tests confirmed that the midpoints of the ability groups obtained with this procedure were not significantly different

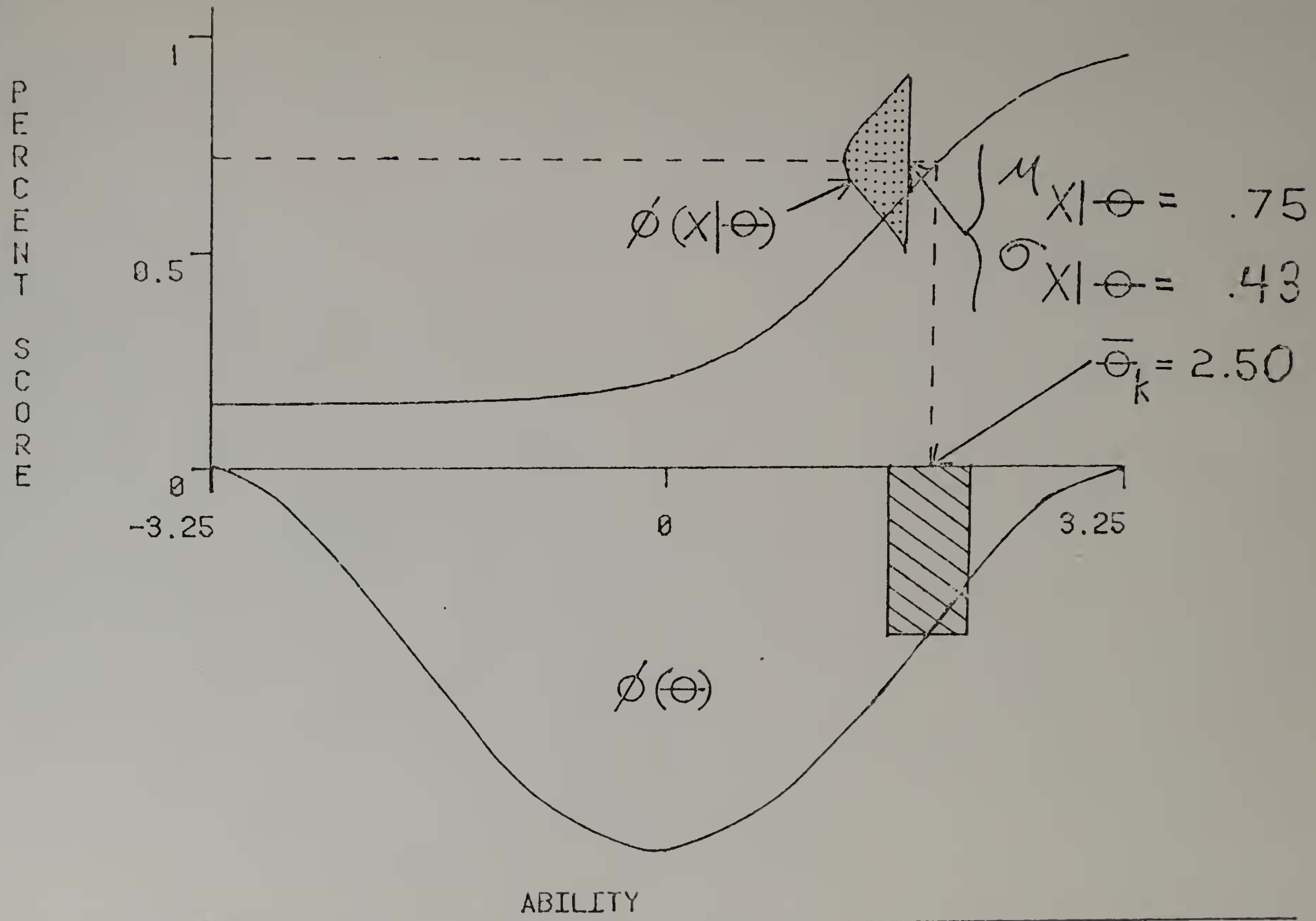


Figure 4. Predicting the Marginal Distribution of Number-Correct Scores from Ability

from the means or medians of the actual ability groups (except at the extremes of the ability distribution).

The mean and standard deviation of the conditional distribution of number-correct scores were found for each of the 17 discrete levels of ability for each test using equations [16] and [17]. Figure 4 shows an ability group with the range 2.25 to 2.75 and mean ability, $\overline{\theta}_k$, 2.50. The mean of the conditional distribution of number-correct scores (in the percent-correct metric) was .75 and the standard deviation was .433. The conditional distribution of raw scores was found by calculating a standard deviate for each raw score:

$$z_g = \frac{X-\mu_{X|\theta}}{\sqrt{\sigma_{X|\theta}{}^2}} \quad . \qquad [18]$$

The normal deviates were transformed to percentage points of the normal distribution. Denoting $\phi(x,|\theta_k)$ as the probability density of number-correct scores, the joint distribution of number-correct scores and ability is found as:

$$\emptyset(X,\theta_k) = \emptyset(X|\theta_k) \quad g^*(\theta_k) \quad , \qquad [19]$$

where $g^*(\theta_k)$ is the number of examinees in ability group k, obtained from the data. The marginal distribution of number-correct scores, $\emptyset(X)$, was found by summing the joint distributions of X and θ_k across the 17 levels of ability:

$$\emptyset(X) = \sum_{k=1}^{17} \emptyset(X,\theta_k) \quad . \qquad [20]$$

Two predicted number-correct score distributions were obtained: one with $P_g(\theta_k)$ estimated from the Rasch model, and one with $P_g(\theta_k)$ estimated from the three-parameter model.

Model Fit

Predicted distributions of number-correct scores were compared to actual distributions of number-correct scores for the 25 tests with non-parametric statistics. Although the generating functions for the predicted distributions resulted in the normal form, there is no assumption in latent trait theory that the distribution of ability is normal. Graphic methods were employed to interpret statistical findings.

Kolmogorov-Smirnov Statistic.—The Kolmorgorov-Smirnov (K-S) statistic was used to compare cumulative distributions of predicted and observed number-correct scores. Cumulative distributions were obtained by accumulating raw score frequencies predicted by the models (f_e) or observed in the data (f_o). Negative, positive, and absolute differences between predicted and observed cumulative distributions were calculated at each raw score level. The K-S statistic is based on the maximum absolute difference, $D=MAX(|f_e-f_o|)$ occurring at any point along the distribution. The difference is used to compute a test statistic Z, which takes into account the number of score levels. Z is compared to tabled values to determine exact probability levels. The probability level for the K-S statistic is not a function of degrees of freedom. A K-S statistic was found for the Rasch model and the three-parameter model for 20 tests.

K-S statistics were computed in a FORTRAN program written by the author.

Chi-Square Statistic. A chi-square test was used as a secondary measure for detecting discrepancies between expected and actual distributions of number-correct scores. Because it is assumed in the chi-square test that sample frequencies are normally distributed about population frequencies, expected cell frequencies below 5 are generally considered insufficient. Because many predicted cell frequencies were zero or negligible, especially at the lower score levels, it was necessary to group across score levels before applying the chi-square test. Depending on test length, three to four raw score levels were grouped together before computing the test statistic. Score groups at the extremes of the distribution contained 5 or 6 raw score levels. Grouping rules were based on expected frequencies for the three-parameter model and were consistently applied to Rasch and observed data frequencies. Grouping of scores resulted in a reduction of degrees of freedom for the test from "n-1" for n items to approximately "1/3 (n-1)" or "1/4 (n-1)".

The chi-square statistic is given as:

$$\chi^2 = \sum_{1=1}^{j} (f_o - f_e)^2 / f_e \quad , \qquad [21]$$

where j is the number of score groups, f_o is the observed frequency of examinees in score group j, and f_e is the expected frequency of persons in group j. The test statistic is compared to tabled values to obtain the exact probability with (j-1) degrees of freedom.

Since test length and, consequently, degrees of freedom were different for each test, a mean square statistic was computed to make comparisons between tests. The mean square was given by:

$$MSQ = \chi^2/\partial f \qquad [22]$$

where ∂f stands for degrees of freedom. Chi-square and mean square statistics were computed for each model for 25 tests in a FORTRAN program written by the author.

Graphic Interpretation

Graphs provided a visual aid for exploring the location of the greatest discrepancies between the predicted and observed score distributions. Frequency plots also provided a means for assessing model fit when score distributions took on different forms. The graphs each pictured the observed number-correct score distribution for a test and two predicted number-correct score distributions based on the one- and three-parameter models. Frequencies in the plots were based on grouped score distributions. The horizontal axis in a graph indicated score group, and the vertical axis depicted relative frequency of examinees. Graphs were produced on a Tektronix 4000 series terminal with the PLOT 10—EASY GRAPHING software.

Comparative Fit.—Model fit was explored with a .01 rejection region, but results were also reported with the more stringent .05 region of rejection. To make comparisons between the two models, the mean K-S statistic across 20 tests was found and the mean square across 25 tests was obtained. Statistical findings were supplmented by graphical evidence for more meaningful interpretation of comparisons between the two models.

Association Between Model Violations and Model Fit.—Correlational methods were used to explore relationships between fit statistics and indices of deviation from model assumptions. Such techniques were only able to detect linear relationships, although associations, if they existed, may have been non-linear. Average K-S statistics across 20 tests, and average mean squares across 25 tests had been obtained by methods described earlier. These measures were correlated with the indices of violation of model assumptions using product moment and rank order methods. Partial correlations were computed to further probe relationships. The IDAP package, an interactive statistical tool written in APL, was employed for these analyses.

Estimation Precision

Precision of Item Parameters Estimated From Small Samples

Five tests were used to explore precision of item parameter estimates from small samples of 250 examinees. Item response data for 250 examinees was drawn from larger samples of 1000 by a spaced sampling plan. Prior to sampling, it was statistically verified for each test that a relationship did not exist between raw score and examinee order. Abilities previously estimated with larger samples of 1000 were assumed to be good approximations of population values and were substituted as true abilities for the 250 examinees. Then, item parameters of the Rasch and three-parameter models were estimated on the small samples. Precision of the small sample item parameter estimates was explored by correlating these estimates with those based on the larger samples of 1000 examinees. Both Pearson and rank order methods were employed. The average absolute differences (AAD's) between item parameters estimated on the small and large samples were also computed. For both the Rasch and three-parameter models, item difficulties ($\hat{b}_g$) estimated from the two sized samples were compared; for the three-parameter model, comparisons were also made between item discrimination and guessing estimates ($\hat{a}_g$ and $\hat{c}_g$) from the two-sized samples. Finally, the Rasch and three-parameter models were compared, using the statistics described above, to determine which model demonstrated more precise small sample estimates of item difficulty.

Precision of Abilities Estimated from Short Tests

Five tests were used to explore precision of ability estimates based on short 20-item tests. Twenty items were randomly sampled from the longer tests. Estimates of item parameters for the 20 items, determined earlier, were assumed to be population values and were substituted as true values for the item parameters. Ability was estimated for 1000 examinees on the short tests under the assumptions of the one- and three-parameter models. Estimation precision for the short tests was examined by correlating short tests ability estimates with those obtained from the full-length tests. Both product moment and rank order correlations were computed. AAD statistics between ability estimates from the 20-item and longer tests were obtained as an additional measure of comparison. The three measures of precision were then evaluated between models to determine whether the Rasch or three-parameter model produced more precise short-test ability estimates.

Cost Method

Little data has been available for making comparisons in cost of estimation of parameters between the Rasch and three-parameter models. Such data were readily available in this study. Both Central Processor Unit (CPU) time and total dollar expense for estimations under each model for each test had been recorded. Average CPU seconds and cost were computed across 25 tests to compare the estimation process between the two models. Cost data were also available from estimations in which either ability or item parameters had been known in advance and only person or item parameters were estimated. These data were recorded for the five tests which had been utilized in the estimation precision analyses. This allowed a cost comparison between the Rasch and three-parameter models when only ability, or only item parameters, needed to be estimated.

The costs and CPU times recorded in this study were based on the batch processing cost of executing the LOGIST computer program and did not include data preparation or time-sharing costs which had been quite substantial in some instances. The data was accumulated on a Control Data Corporation (CDC) CYBER-175, an extremely high speed machine, operating under the NOS 1-3 operating system. Costs cannot be compared to commercial rates because they were charged at academic discounts. The costs for each parameter estimation was based on a weighted average of central processor usage, memory extents, input/output units, and CPU time, and hence cannot even be generalized to similar computer systems in other academic environments. Computation speed on the CYBER-175 for this problem was benchmarked to be approximately twice the speed on CDC 6400 series computers and nearly four times faster than IBM 360/370 series computers. The efficiency of computation on the CYBER-175 for LOGIST was partially attributed to the fact that the machine has a 60-bit word in contrast to the 32-bit word on the IBM computers.

CHAPTER IV

FIT OF LATENT TRAIT MODELS TO EMPIRICAL DATA

Conventional Description of Tests

Standard item statistics for the 25 tests used in the study are presented in Table 3 which includes average item difficulties, average item-total score correlations, and the KR-20 for each test. The measures in the table show that the tests varied considerably in their conventional difficulty levels and average item-total score correlations. The mean item difficulty level across the 25 tests was .59 and ranged from a difficulty level of .711 on an easier test to a difficulty level of .475 on a harder test. Item difficulty levels for the majority of tests were in the range .55 to .65.

Average item-total score correlations ranged from a low of .212 to a high of .538 with an overall average of .380 across tests. Item-total correlations in this study tended to be somewhat lower than values generated in simulation studies of latent trait model fit, although they did reflect values frequently observed with empirical data.

Internal consistency estimates (KR-20) were high for nearly all of the tests: eighty-eight percent of the tests had a KR-20 value over .80 and forty-eight percent of the tests had a KR-20

Table 3

Statistical Descriptors of 25 Tests

Test	Average Item Difficulty	Average Item-Total Correlation	Reliability (KR-20)
Stanford Word Study	.624	.538	.954
Stanford Reading	.544	.444	.947
SAT Verbal #1	.566	.316	.912
CTBS Math Comprehension	.626	.497	.944
CTBS Vocabulary	.539	.479	.927
Stanford Vocabulary	.556	.383	.903
Stanford Science	.510	.389	.921
Stanford Language	.505	.406	.722
Stanford Social Studies	.514	.348	.891
Stanford Listening	.655	.378	.899
Stanford Spelling	.569	.463	.945
Stanford Math Applications	.598	.473	.925
Stanford Math Concepts	.576	.351	.852
ICRT	.572	.415	.808
SAT Verbal #2	.560	.319	.913
Georgia Regents #1	.711	.212	.791
Georgia Regents #2	.710	.248	.842
Georgia Regents #3	.692	.249	.829
CAT Vocabulary	.534	.402	.895
ITBS Vocabulary	.529	.426	.903
CAT Comprehension	.480	.338	.870
ITBS Comprehension	.475	.378	.926
Rasch Test on Environment	.677	.382	.863
Rasch Test on Citizenship	.747	.308	.789
Rasch Test on Career	.739	.395	.882

measure over .90. The least reliable test in the study was the language subtest of the Stanford which had an estimated KR-20 of .722. The most reliable test was the Stanford word study subtest. The reliability estimate was .954.

Measures of Violation of Latent Trait Model Assumptions

Unidimensionality

Results of principal components analyses of 25 tests are presented in Table 4. Test length, the number of factors with eigenvalues greater than one, the percent of variance accounted for by the first factor, the eigenvalue for the first factor, and the eigenratio between the first and second factors are shown in the table. The last column of the table provides information pertaining to a ranking of tests based upon the extent to which they are unidimensional. The highest rank, 25, was assumed to be the most unidimensional of the 25 tests. Although all tests had more than one factor with an eigenvalue over one, inspection of the factor variances, eigenvalues, and factor loading patterns (not shown) suggested that there was only one primary factor on each test. Most secondary factors displayed high factor loadings for only one or two items and were considered to represent unique factors. The factor loading patterns indicated

Table 4

Principal Components Analysis Results for 25 Tests

Test	Test Length	Number of Factors $(\lambda \geq 1)$	First Factor Variance	First Factor Eigenvalue (λ_1)	Eigen-ratio[a] (λ_1/λ_2)	Rank[b]
Stanford Word Study	50	5	49.7%	24.853	9.077	23
Stanford Reading	71	13	37.1	26.309	3.408	2
SAT Verbal #1	85	21	22.7	19.320	4.653	6
CTBS Math Comprehension	48	4	58.6	28.150	11.590	24
CTBS Vocabulary	40	3	51.7	20.660	14.957	25
Stanford Vocabulary	50	6	29.9	14.930	7.225	19
Stanford Science	60	10	28.7	17.230	6.430	17
Stanford Language	80	19	30.2	24.133	4.941	7
Stanford Social Studies	54	15	25.3	13.688	5.805	13
Stanford Listening	50	13	30.6	15.312	8.795	22
Stanford Spelling	60	10	38.6	23.149	7.580	20
Stanford Math Application	40	5	42.3	16.914	6.335	16
Stanford Math Concepts	35	9	27.2	9.528	5.460	12
ICRT	16	3	39.0	6.247	5.437	11
SAT Verbal #2	85	27	23.1	19.669	4.322	5
Georgia Regents #1	70	27	16.8	11.751	3.332	1
Georgia Regents #2	70	26	18.5	12.942	4.065	4
Georgia Regents #2	70	24	20.6	14.392	5.116	8
CAT Vocabulary	40	7	33.8	13.523	5.185	9
ITBS Vocabulary	38	7	34.7	13.186	7.648	21
CAT Comprehension	45	13	25.5	11.477	5.841	14
ITBS Comprehension	68	17	27.9	18.959	5.338	10
Rasch-Envirorment	34	9	34.2	11.630	6.646	18
Rasch-Citizenship	30	9	27.7	8.300	3.860	3
Rasch-Career	39	10	35.8	13.950	5.987	15

[a] Ratio between first and second eigenvalues.

[b] Rank based on eigenratio. High rank indicates the most unidimensional of the tests.

that none of the primary factors seemed to be "difficulty" factors, since all loadings had the same sign.

Variances associated with first factors ranged from a maximum of 58.6 percent on the math comprehension subtest of the CTBS to a minimum of 16.8 percent on one of the forms of the Georgia Regents. Eighty percent of the tests had 25 percent or more of the test variance explained by the first factor.

First factor variance, while indicating the strength of the primary factor, provides little information about unidimensionality since it does not show the dominance of the primary factor over other factors. The ratios between first and second factor variances (the eigenratios) demonstrated the inter-relationship between factors and thus provided a more powerful measure of unidimensionality. Because eigenratios do not vary with test length, they allow comparisons between tests with different numbers of items. Using the eigenratio as a criterion, the most unidimensional test was the CTBS vocabulary subtest which had an eigenratio of 14.957. The CTBS math comprehension subtest was the next most unidimensional test by the criterion and had an eigenratio of 11.59, but it should be noted that this test had the highest percentage of variance associated with its first factor. The remaining 23 tests had eigenratios below ten. The lowest eigenratio obtained in the study was 3.332 on one form of the Georgia Regents. The SAT, the Stanford language subtest, the Rasch citizenship test, and other forms of the Georgia Regents had eigenratios below 4.0 and were considered to be lacking unidimensionality.

The achievement tests tended to be more unidimensional than the aptitude tests in the study.

Equality of Item Discrimination Indices

Data concerning item discrimination indices for the 25 tests are shown in Table 5 which includes average item-total correlations, percents of "equal discrimination" indices, and a ranking of tests based on the percent of equal discrimination indices. The Stanford math applications subtest had the most homogeneous discriminations (87.5 percent). The CAT vocabulary subtest and the Rasch environment test had the most heterogeneous discrimination values since only 50 percent were about equal. Sixty to seventy percent of discrimination indices were approximately equal in value on the remainder of the tests. The data suggest that many of the tests had been designed so that the item-total correlations for most items would be nearly equal. Other results, based on estimated discrimination values ($\hat{a}_g$), collected during the parameter estimation phase of the study, indicated that discrimination values were not so equal.

Descriptive statistics for estimated item discrimination values ($\hat{a}_g$) are shown in Table 6. The tests in the table are rank ordered based on the standard deviation of estimated discrimination values, $\sigma_{\hat{a}_g}$. Tests with the most unequal discrimination values are shown at the top of the list. The mean and standard deviation of estimated discrimination values for each test, along with information on ranges of $\hat{a}_g$ values, are shown in the table. For the test with the most

Table 5

Equality of Item Discrimination Indices
(N=25)

Test	Average point-biserial	Percent of items with equal discrimination	Rank[a]
Stan. Word Study	.538	72.00	14.0
Stan. Reading	.444	60.56	4.0
SAT Verbal #1	.316	77.65	19.5
CTBS Math Comp.	.497	72.92	15.0
CTBS Vocabulary	.379	70.00	10.5
Stan. Vocabulary	.383	82.00	22.0
Stan. Science	.389	81.67	21.0
Stan. Language	.406	71.25	12.0
Stan. Soc. Stud.	.348	66.67	75.0
Stan. Listening	.378	84.00	23.0
Stan. Spelling	.463	71.67	13.0
Stan. Math Applic.	.473	87.50	25.0
Stan. Math Con.	.351	74.29	17.0
ICRT	.415	62.50	6.0
SAT Verbal #2	.319	77.65	19.5
Georgia Regents #1	.212	70.00	10.5
Georgia Regents #2	.248	68.57	9.0
Georgia Regents #3	.249	58.57	3.0
CAT Vocabulary	.402	50.00	1.5
ITBS Vocabulary	.426	73.68	16.0
CAT Comprehension	.338	66.67	7.5
ITBS Comprehension	.378	61.76	5.0
Rasch-Environment	.382	50.00	1.5
Rasch-Citizenship	.308	86.67	24.0
Rasch-Career	.395	76.92	18.0

[a] The highest rank was assigned to the test with the most homogeneous item discrimination values.

Table 6

Descriptive Statistics on Item Discrimination Estimates[1]

	Mean	Deviation	Approximate Range[2]	Endpoints of Approximate Range	Actual Range
Most heterogeneous	.939	.575	1.150	.364-1.514	1.850
	1.263	.532	1.046	.740-1.786	1.977
	.745	.485	.970	.260-1.230	1.076
	1.237	.483	.966	.754-1.720	1.727
	1.274	.468	.936	.806-1.760	1.713
	.890	.443	.886	.447-1.333	1.982
	1.162	.434	.868	.728-1.596	1.570
	.709	.413	.826	.296-1.122	2.002
	1.194	.403	.806	.791-1.597	1.806
	1.192	.401	.802	.791-1.593	1.396
	1.131	.392	.784	.739-1.523	1.671
	1.111	.383	.766	.728-1.494	1.504
	1.025	.382	.764	.643-1.407	1.721
	.755	.376	.752	.379-1.131	1.449
	1.106	.365	.730	.741-1.471	1.538
	.964	.347	.694	.617-1.311	1.519
	.870	.345	.690	.525-1.215	1.767
	.797	.291	.582	.506-1.088	1.674
	.927	.285	.570	.642-1.212	1.607
	.744	.276	.552	.468-1.020	1.304
	.564	.275	.550	.289-0.839	1.221
Most homogeneous	.736	.236	.472	.500-0.972	1.046

[1] Based on N=22 tests. Item parameters were not recorded for three tests.

[2] The approximate range is based on two standard deviations (68 percent).

heterogeneous discriminations, the standard deviation, $\sigma_{\hat{a}_g}$ was .575. The standard deviation of discrimination indices for the test with the most homogeneous discriminations was .236. The average standard deviation of estimated discrimination indices across tests was .399. One or two items on some of the tests had very high item discrimination estimates, which had been set to the maximum value of 2.00 during estimation. When these few items were included in the calculation of the range of discrimination indices for a test, they tended to exaggerate the range, hence an approximate range, based on two standard deviations, was also reported. These data suggest that item discrimination values were more heterogeneous than the item-total correlation evidence had indicated. The correlation between the two measures of equality of item discrimination indices was .414.

Guessing

Results pertaining to estimates of guessing on 25 tests are presented in Table 7. Test length, the number of "hard" items on a test (items answered incorrectly by more than two-thirds of examinees), average test difficulty, average difficulty level for hard items, and average difficulty level for hard items computed for the bottom ten percent of examinees ("guessing") are shown in the table. The next to last column in the table is a rank ordering of tests based on the amount of guessing. The highest rank, 16, was assigned to the test which demonstrated the lowest percent of guessing. The last column in Table 7 shows the chance level parameter ($\bar{\hat{c}}_g$), averaged across hard

Table 7

Estimation of Guessing on 25 Tests

Test	Length	Number of Hard Items[1]	Difficulty Total Test)	Difficulty (Hard Items)	Difficulty (Hard items lowest 10%) Guessing	Rank[2]	$\hat{c}_g$[3]
Stanford Word Study	50	0	.624	a	a	a	N/A
Stanford Reading	71	15	.544	.253	.091	16	.228
SAT Verbal #1	85	25	.566	.239	.104	12	.131
CTBS Math Comprehension	48	1	.626	.306	a	a	N/A
CTBS Vocabulary	40	4	.539	.291	.092	15	.199
Stanford Vocabulary	50	9	.556	.298	.096	14	.157
Stanford Science	60	6	.510	.323	.119	8	.220
Stanford Language	80	12	.505	.301	.127	6	.203
Stanford Social Studies	54	4	.514	.292	.172	2	.203
Stanford Listening	50	2	.655	.315	a	a	N/A
Stanford Spelling	60	6	.569	.289	.100	13	.199
Stanford Math Applications	40	2	.598	.242	a	a	N/A
Stanford Math Concepts	35	3	.576	.303	a	a	N/A
ICRT	16	0	.572	a	a	a	N/A
SAT Verbal #2	85	26	.560	.240	.123	7	.146
Georgia Regents #1	70	7	.711	.277	.136	5	b
Georgia Regents #2	70	6	.710	.245	.110	11	.079
Georgia Regents #3	70	6	.692	.260	.198	1	b
CAT Vocabulary	40	11	.534	.259	.113	10	.216
ITBS Vocabulary	38	5	.529	.321	.150	3	.203
CAT Comprehension	45	12	.480	.258	.118	9	.129
ITBS Comprehension	68	15	.475	.285	.142	4	.227
Rasch-Environment	34	1	.677	.277	a	a	N/A
Rasch-Citizenship	30	0	.747	a	a	a	N/A
Rasch-Career	39	1	.739	.177	a	a	N/A

[1] Items answered incorrectly by 2/3's of examinees.
[2] Rank based on guessing estimate (16 tests). High ranks indicate less guessing.
[3] Guessing parameter, averaged across hard items, estimated by latent trait methods.

[a] Not estimated. Too few hard items.
[b] Not recorded.

items, estimated with latent trait methods. Guessing was estimated on only 16 tests since nine tests had too few difficult items to evaluate guessing. The conventional and the latent trait guessing estimates suggest that there was a considerable amount of random guessing on the 16 tests. On the average, 12 percent of low ability examinees obtained correct responses on hard items, yet only 33 percent of all examinees scored correctly on these items. The percent of low ability examinees who obtained correct answers by chance ranged from nine to 20 percent. The latent trait pseudo-guessing, or chance level parameter, indicated that, on the average, 18 percent of low ability examinees obtained correct answers by random guessing. The chance level parameter, averaged across hard items, ranged from .08 to .23. The correlation between the conventional and latent trait estimates of guessing was .208.

Overall Model Fit

Kolmogorov-Smirnov (K-S) Test

K-S statistics for 20 tests are shown in Table 8, which also includes the rank for each test based on three measures of deviation from latent trait model assumptions. The average K-S statistic across the 20 tests was 1.304 for the Rasch model as compared to 1.289 for the three-parameter model, which indicates that the more general three-parameter model fit data somewhat better, on the average, than the Rasch model. Data fit the three-parameter model better than the Rasch model on 55 percent of the tests. Probability levels associated with

Table 8

Model Fit Based on the K-S Statistic (20 Tests)[1]

Test	K-S 1-Par Model	K-S 3-Par Model	Unidim. Rank[2]	Discrim. Rank[3]	Guessing Rank[4]
Stanford Word Study	1.148	$\underline{1.067}$	23	14.0	a
Stanford Reading	3.384*	$\underline{3.000}$*	2	4.0	16
Stanford Language	1.617	$\underline{1.524}$**	7	12.0	6
Stanford Social Studies	$\underline{1.232}$	1.261	13	7.5	2
Stanford Listening	$\underline{0.899}$	1.117	22	23.0	a
Stanford Spelling	$\underline{1.063}$	1.185	20	13.0	13
Stanford Math Applications	0.942	$\underline{0.866}$	16	25.0	a
Stanford Math Concepts	1.448**	$\underline{1.327}$	12	17.0	a
ICRT	1.617*	$\underline{1.446}$**	11	6.0	a
SAT Verbal #2	1.485*	$\underline{1.477}$**	5	19.5	7
Georgia Regents #1	$\underline{1.140}$	1.150	1	10.5	5
Georgia Regents #2	1.164	$\underline{1.104}$	4	9.0	11
Georgia Regents #3	$\underline{1.138}$	1.338	8	3.0	1
CAT Vocabulary	$\underline{1.247}$	1.333	9	1.5	10
ITBS Vocabulary	$\underline{1.150}$	1.308	21	16.0	3
CAT Comprehension	1.094	$\underline{1.063}$	14	7.5	9
ITBS Comprehension	$\underline{0.984}$	0.995	10	5.0	4
Rasch-Environment	1.007	$\underline{1.005}$	18	1.5	a
Rasch-Environment	$\underline{1.213}$	1.217	3	2.4	a
Rasch-Career	1.113	$\underline{0.997}$	15	18.0	a

[1] The underscored statistic indicates the model with better fit.
[2] Unidimensionality rank based on 25 tests. High ranks indicate the most unidimensional tests.
[3] Equality of item discrimination rank based on 25 tests. High ranks indicate equal discriminations.
[4] Guessing rank based on 16 tests. High ranks indicate the least amount of guessing.

[a] Too few hard items to obtain estimate of guessing.
*Rejected at p=.01.
**Rejected at p=.05.

K-S statistics indicated that the Stanford reading subtest was the only test rejected as not fitting either of the models when a one percent rejection region was the criterion. The 16-item criterion-referenced test did not fit the Rasch model with this criterion. When the region of rejection was increased from one percent to five percent, probability levels showed that the Stanford language subtest and the SAT verbal test did not fit either model very well, the math concepts subtest of the Stanford did not fit the Rasch model, and the 16-item ICRT did not fit the three-parameter model. There was a .977 correlation between K-S statistics for the Rasch and three-parameter models. The rank order correlation between K-S statistics for the two models was .836. These correlations indicated that the pattern of fit of data to the two models was quite similar. It is reasonable to conclude from the K-S test of fit that many standardized achievement and aptitude tests, developed with conventional methods, can be fit by latent trait models. It is also concluded that the Rasch model fits tests nearly as well as the three-parameter model.

Chi-Square Test

Chi-square test results are presented in Table 9, which includes mean squares ($\chi^2/\partial f$) for the Rasch and three-parameter models, rankings based on three deviations from model assumptions, and the degrees of freedom for the chi-square test for the 25 data sets. Although correlations between K-S measures and mean squares were significant ($r=.764$ for the three-parameter model and $r=.776$ for the one-parameter model), the chi-square test proved to be a considerably more rigid

Table 9

Model Fit Based on Mean Square Statistic (25 tests)[1,2]

Test	Mean Sq. 1-Par Model	Mean Sq. 3-Par Model	Unidim Rank[3]	Discrim Rank[4]	Guessing Rank[5]	DF[6]
Stanford Word Study	2.002*	1.827**	23	14.0	a	14
Stanford Reading	6.684*	5.826*	2	4.0	16	19
SAT Verbal #1	1.484	1.632**	6	19.5	12	20
CTBS Math Comprehension	1.197	1.158	24	15.0	a	13
CTBS Vocabulary	1.660	1.144	25	10.5	15	12
Stanford Vocabulary	1.703*	1.877**	19	22.0	14	14
Stanford Science	1.853**	2.116*	17	21.0	8	16
Stanford Language	2.782*	2.860*	7	12.0	6	21
Stanford Social Studies	1.927**	1.608	13	7.5	2	13
Stanford Listening	0.924	0.952	22	23.0	a	13
Stanford Spelling	2.048*	1.983*	20	13.0	13	17
Stanford Math Applications	0.942	0.947	16	25.0	a	12
Stanford Math Concepts	3.659*	2.533*	12	17.0	a	9
ICRT	0.060	0.927	11	6.0	a	4
SAT Verbal #2	0.957	1.006	5	19.5	7	21
Georgia Regents #1	1.237	1.081	1	16.5	5	13
Georgia Regents #2	1.028	0.925	4	9.0	11	13
Georgia Regents #3	0.818	0.698	8	3.0	1	13

Table 9 (continued)

Test	Mean Sq. 1-Par Model	Mean Sq. 3-Par Model	Unidim Rank	Discrim Rank	Guessing Rank	DF
CAT Vocabulary	1.172	3.316*	9	1.5	10	11
ITBS Vocabulary	1.382	0.756	21	16.0	3	10
CAT Comprehension	1.957**	1.645	14	7.5	9	12
ITBS Comprehension	2.221*	2.090	10	5.0	4	13
Rasch-Environment	1.679	1.457	18	1.5	a	9
Rasch-Citizenship	2.390**	2.334**	3	24.0	a	7
Rasch-Career	2.011**	1.959**	15	18.0	a	10

[1] The underscored statistic indicates the model with better fit.
[2] Mean squares were computed as chi-squares divided by degrees of freedom.
[3] Unidimensionality rank based on 25 tests. High ranks indicate the most unidimensional tests.
[4] Equality of item discrimination rank based on 25 tests. High ranks indicate equal discriminations.
[5] Guessing rank based on 16 tests. High ranks indicate the least amount of guessing.
[6] Degrees of freedom for chi-square test.

[a] Too few hard items to obtain estimate of guessing.
*Rejected at p=.01.
**Rejected at p=.05.

test of fit than the K-S test and also displayed some curiously inconsistent results with those from the previous measure. Rank order correlations between K-S statistics and mean square statistics were quite low: .165 for the three-parameter model and .223 for the Rasch model.

The mean square, averaged across 25 tests, was 1.83 for the Rasch model and 1.79 for the three-parameter model indicatinq, as earlier, that the three-parameter model fit the data slightly better than the Rasch model. Lower mean squares were obtained for the three-parameter model on 64 percent of the tests. Exact probability levels associated with chi-square test statistics indicated that four tests were not fit by either model when a .01 rejection region was set. When the rejection region was expanded to five percent, 48 percent of the tests were not fit very well by the Rasch model and 44 percent of tests were not fit very well by the three-parameter model. These chi-square test results are of dubious value because it is unlikely that the test statistics used in the study approximated the chi-square distribution.

The Stanford reading subtest showed the poorest fit to data with both the K-S and mean square statistics. The 16-item ICRT, which had demonstrated very poor fit to both models with the K-S statistics, had low mean squares, indicating reasonably acceptable model fit, with the mean square criterion. The language, spelling, and math concepts subtests of the Stanford did not show very good model fit with the mean square criterion. The Stanford word study subtest, the ITBS comprehension subtest, the Stanford science subtest, and the CAT vocabulary subtest also showed rather poor fit with the mean square criterion, yet demonstrated reasonably adequate fit with the K-S statistic as a criterion. The correlation between mean square fit statistics for the Rasch and three-parameter models was .892,and the rank order correlation between the two mean squares was .812, indicating a systematic relationship in fit for the two models.

There was a small but insignificant association between mean square statistics and test length. Product moment correlations between number of score intervals and mean square statistics were .299 and .287 respectively for the Rasch and three-parameter models suggesting that longer tests had been detected as not fitting the models as well as shorter tests. Test length had no relationship to K-S statistics.

The unfavorable picture of model fit suggested by chi-square statistics may be attributed to a number of shortcomings associated with the approach taken in the study. In much statistical work, sample values provide an acceptable approximation to population values. Lord (1980) noted that because sampling frequency distributions tend to be very irregular, a large amount of error is introduced by substituting sampling frequencies for population frequencies. Lord (1980, p. 239) suggested: "The simplest way to reduce such irregularities is to group the observed scores into class intervals," and Lord provided a set of grouping rules comparable to those used in this study. The resulting test statistics, unfortunately do not have the chi-square distribution. During the course of this study, a substantial amount of experimentation was conducted with re-grouping specifications. It was observed that chi-square values were extremely sensitive to the manner in which scores had been grouped into class intervals.

Test statistics, such as those constructed in this study, only have the chi-square distribution when N is infinitely large. Thus, strictly speaking, it was inappropriate to compare these test statistics to tabled values of the chi-square distribution. There has been some sharp debate concerning just how large N needs to be to permit use of chi-square tables. A rule of thumb offered in many statistical texts is that the chi-square test is appropriate when the expected freqeuncies in all categories are over 5. On the other hand, when N is too large, most data is rejected by the chi-square test (and other statistical tests) despite its practical usefulness.

Since mean square values were used to make inter-model comparisons, they were not subject to the criticisms of the chi-square procedure stated in the previous section. The K-S test required only that data were ordinal so that they could be put in cumulative form, and was used for both significance tests of model fit and inter-model comparisons. The K-S test is considered to be a more powerful test of fit than the chi-square test (Hays & Winkler, 1971) and had none of the limitations of the chi-square or mean square measures. Generally, the mean square results supported the conclusions of the K-S test, namely, that the Rasch model describes cognitive test data nearly as well as the three-parameter model.

Graphic Results

Visual evidence, supplied by graphs, was very consistent with model fit results provided by statistical tests. Frequency polygons for 25 tests used in the study are displayed in Figure 5 through 28. Three frequency distributions are shown in each figure: the observed distribution of number correct scores (solid line); the distribution of number-correct scores predicted with the Rasch model (broken line); and the distribution of number-correct scores predicted with the three-parameter model (mixed line). These distributions show frequencies of scores which had been grouped into class intervals to reduce sampling fluctuations. The horizontal axis in each figure represents the class interval and the vertical axis depicts relative frequency of examinees in each score group.

The Influence of Distribution Form on Model Fit

Three general forms of number-correct score distributions were obtained in this study: normal, uniform, and skewed. No bimodal distributions occurred. Some distributions were quite jagged despite the fact that scores had been grouped, but most were relatively smooth. Normal distributions are shown in Figures 7, 10, 13, 19, 23, and 24. Uniformly distributed scores are found in Figures 8, 11, 12, 15, and 18 and the rest of the figures illustrate skewed distributions of number-correct scores. Tests which had particularly irregular number-correct score distributions are seen in Figures 5, 6, 8, 9, 10, 11, 12, 15, 16, 23, 24, and 26.

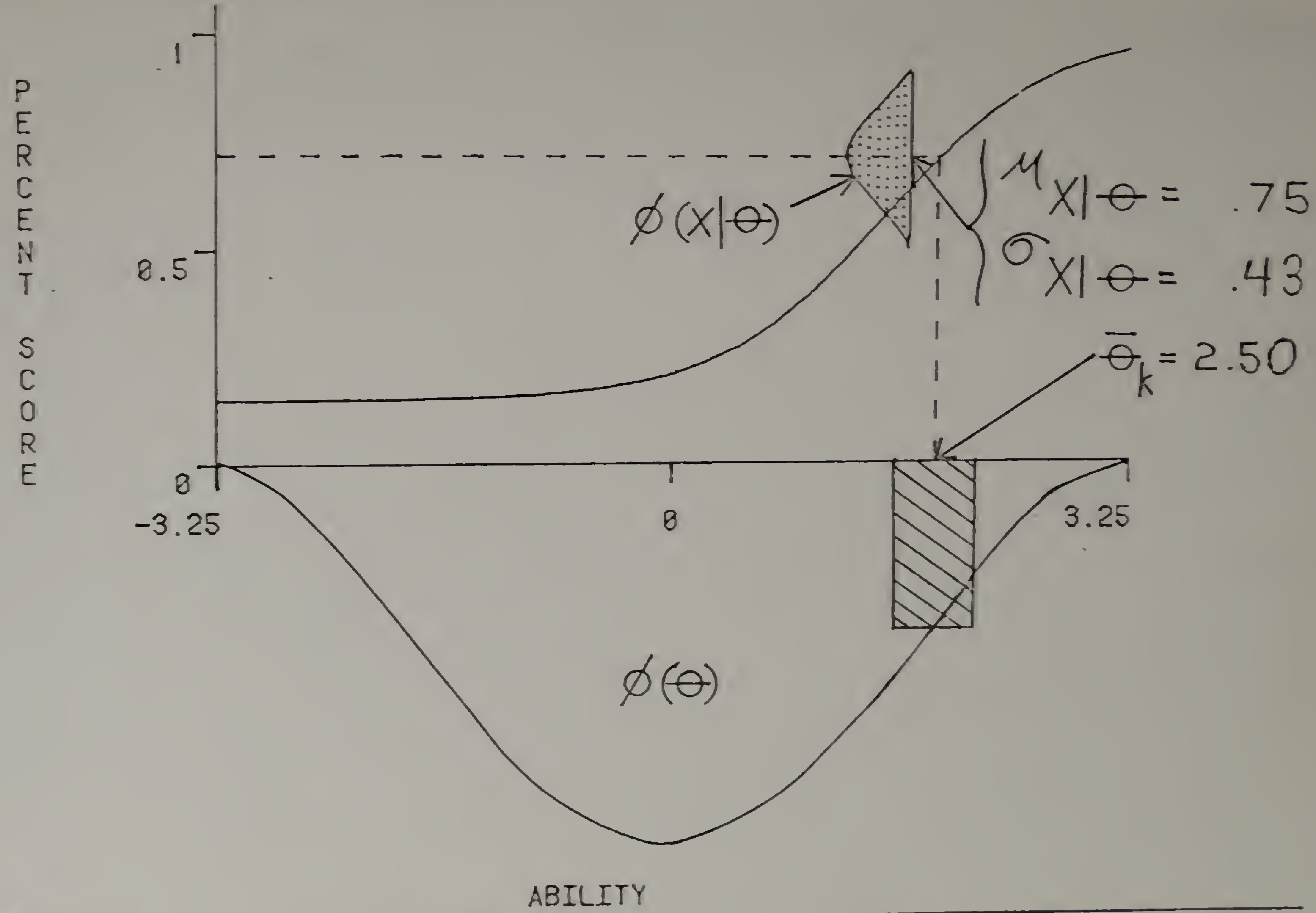


Figure 4. Predicting the Marginal Distribution of Number-Correct Scores from Ability

Figure 5. Observed and Expected Frequencies
Stanford Word Study (50 Items)

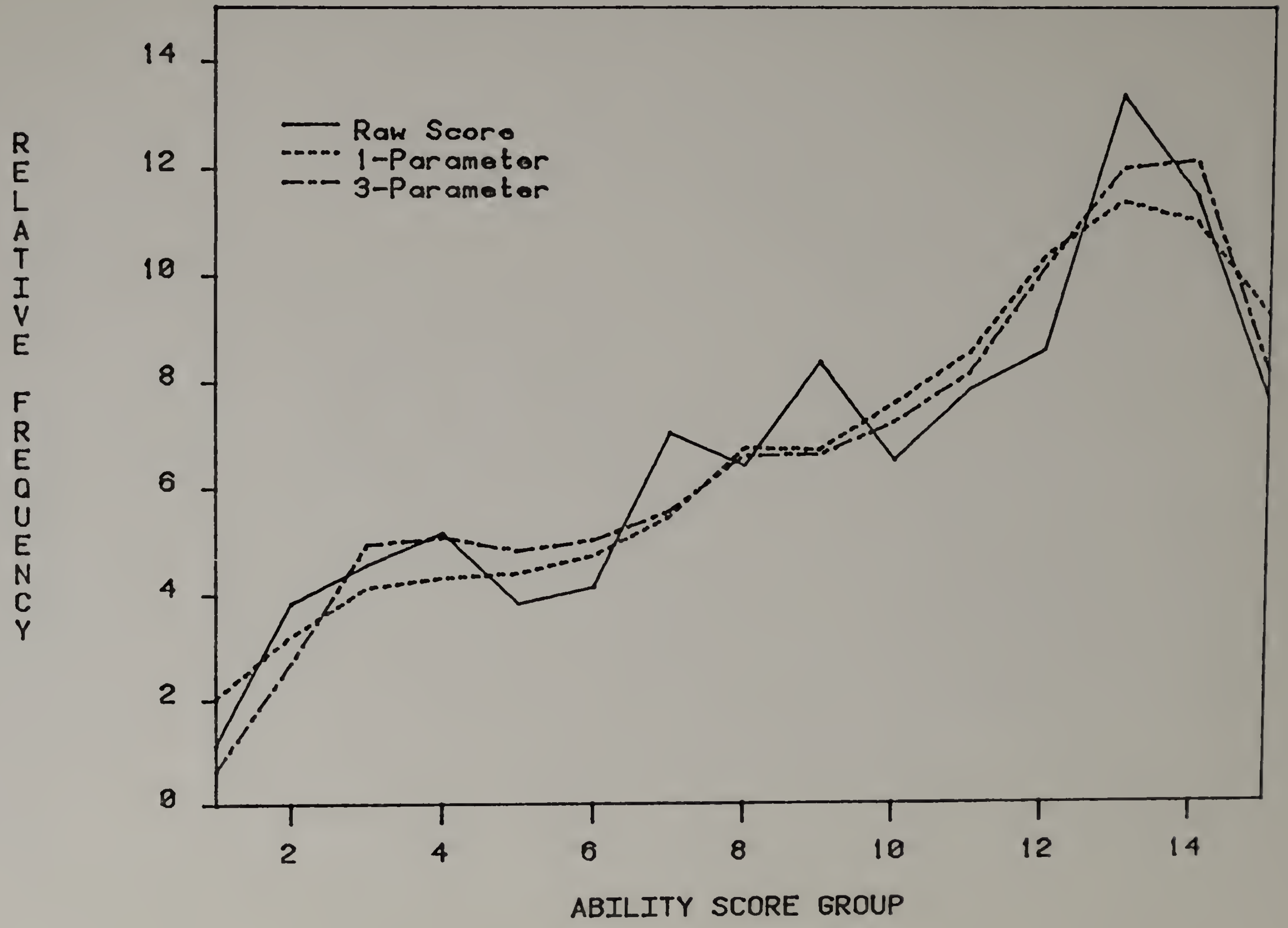


Figure 6. Observed and Expected Frequencies
Stanford Reading (71 Items)

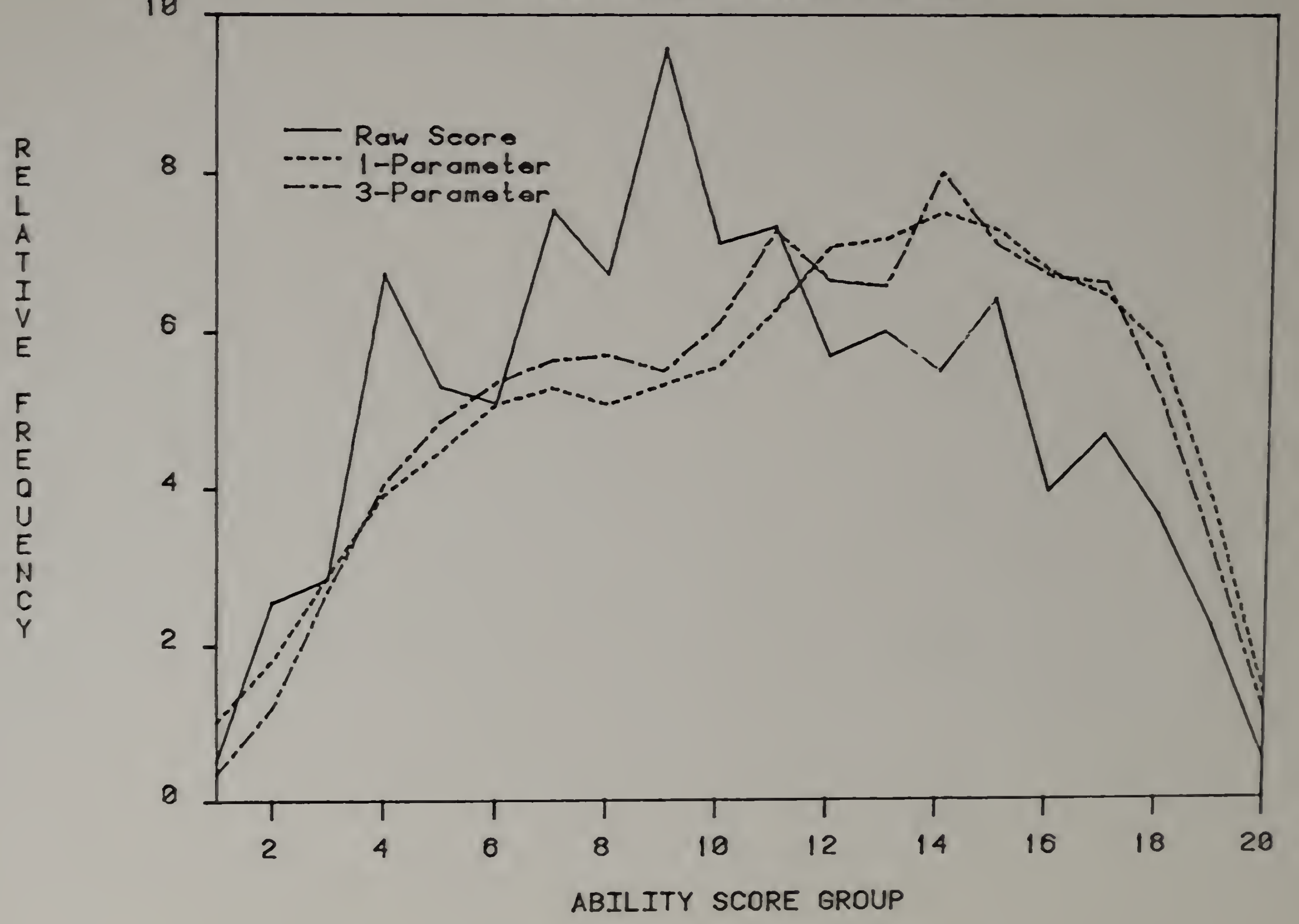


Figure 7. Observed and Expected Frequencies SAT Verbal #1 (85 Items)

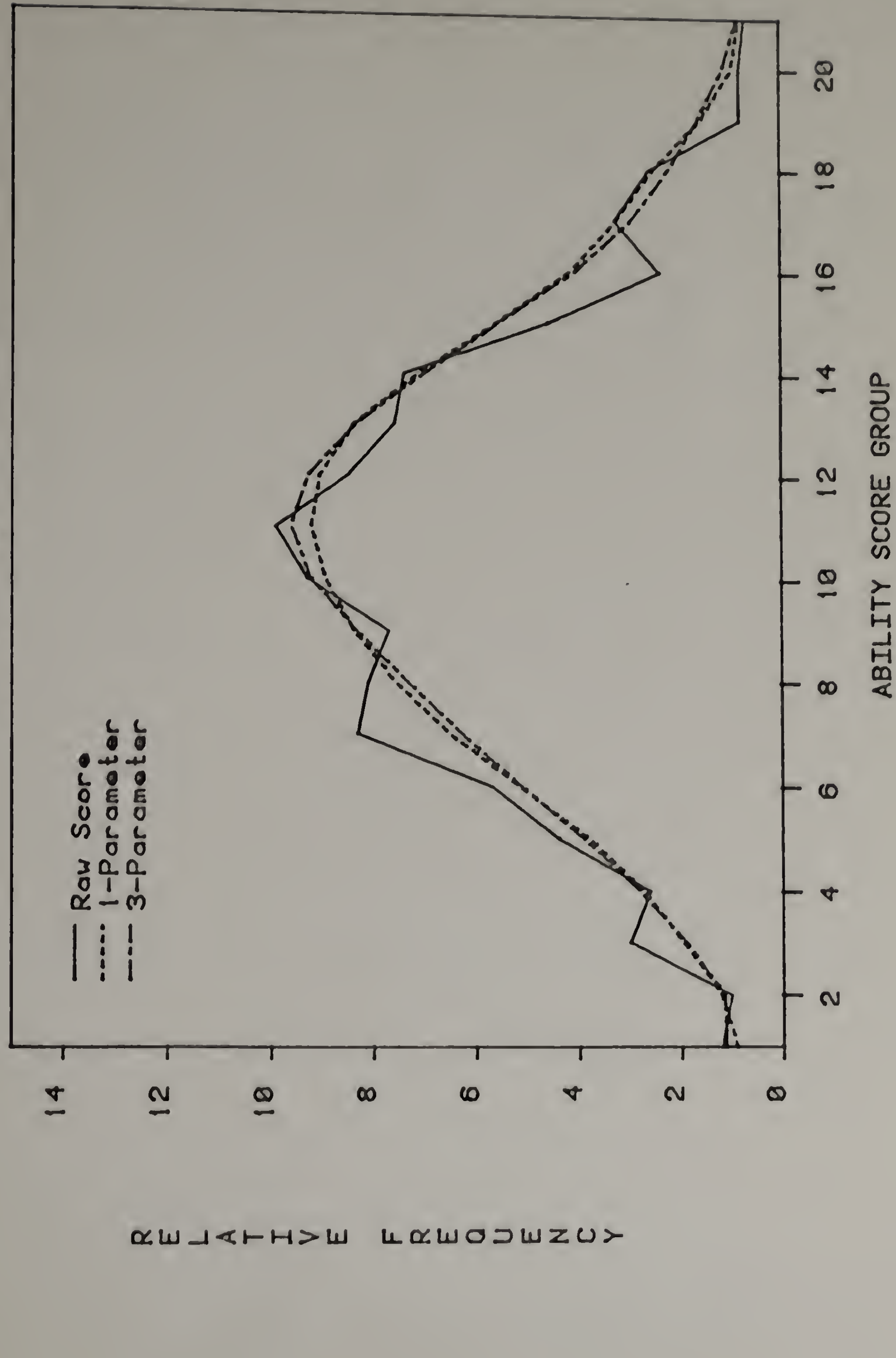


Figure 8. Observed and Expected Frequencies
CTBS Math Comprehension (48 Items)

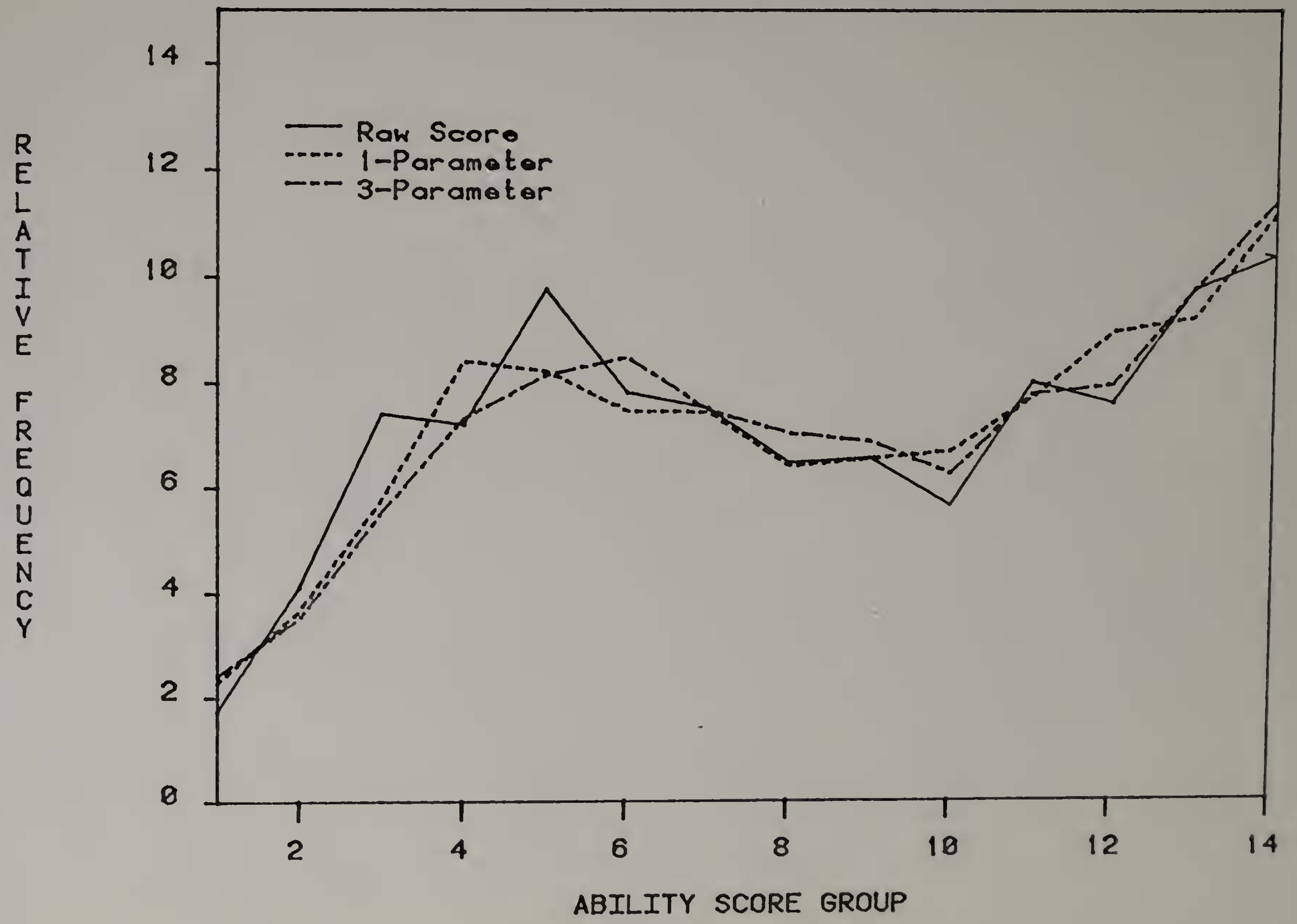


Figure 9. Observed and Expected Frequencies
CTBS Vocabulary (40 Items)

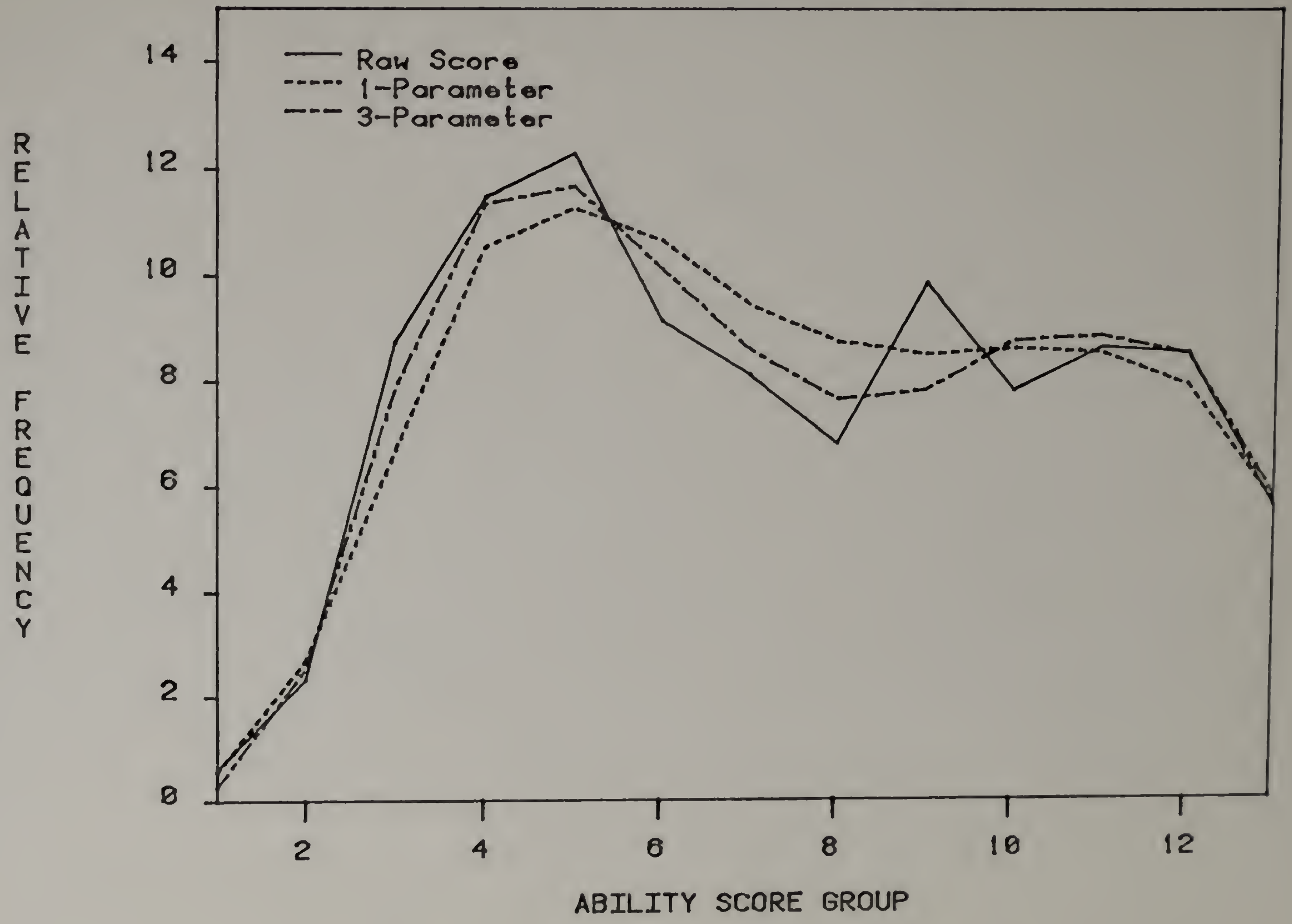


Figure 1Ø. Observed and Expected Frequencies
Stanford Vocabulary (5Ø Items)

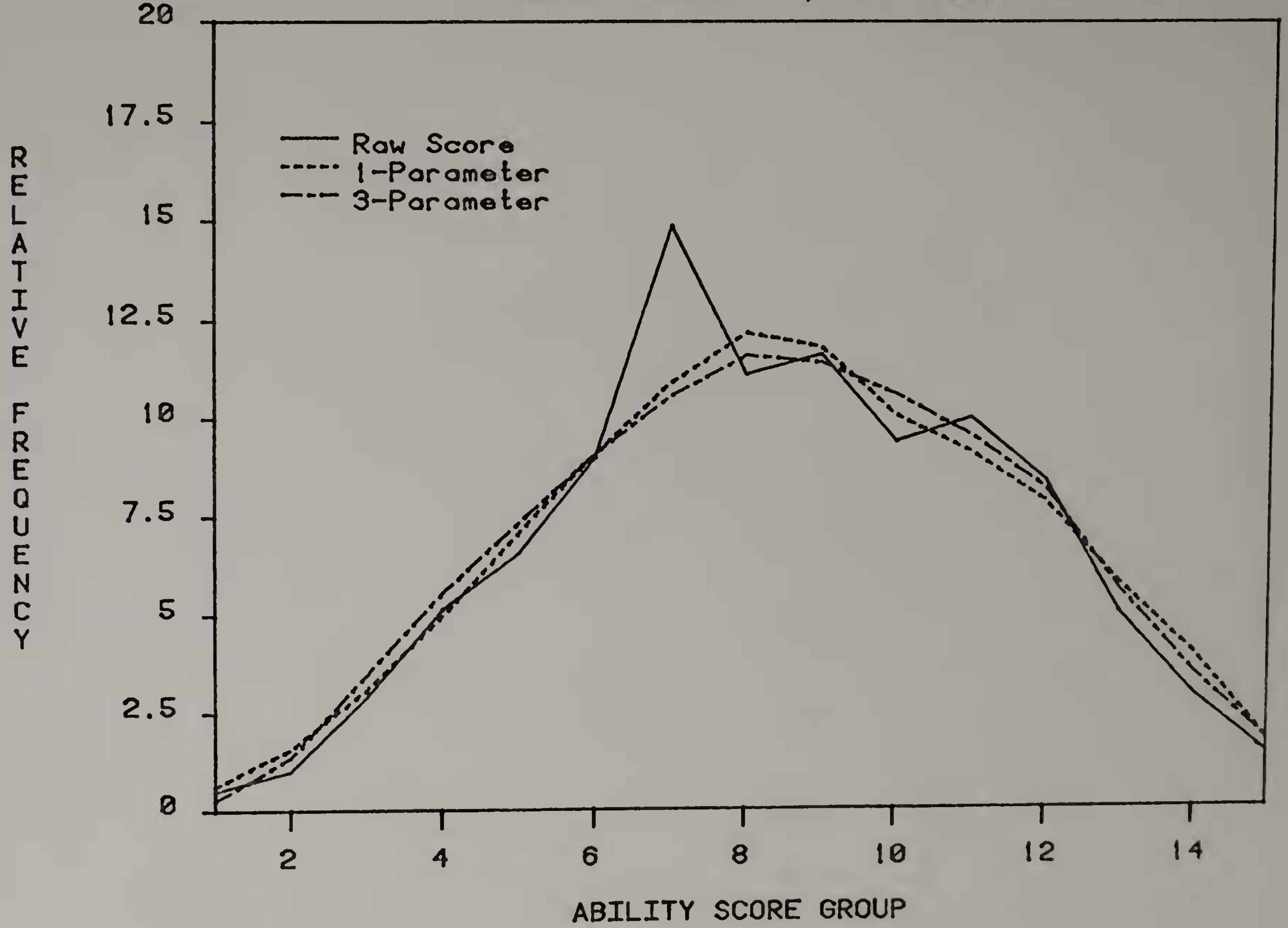


Figure 11. Observed and Expected Frequencues
Stanford Science (60 Items)

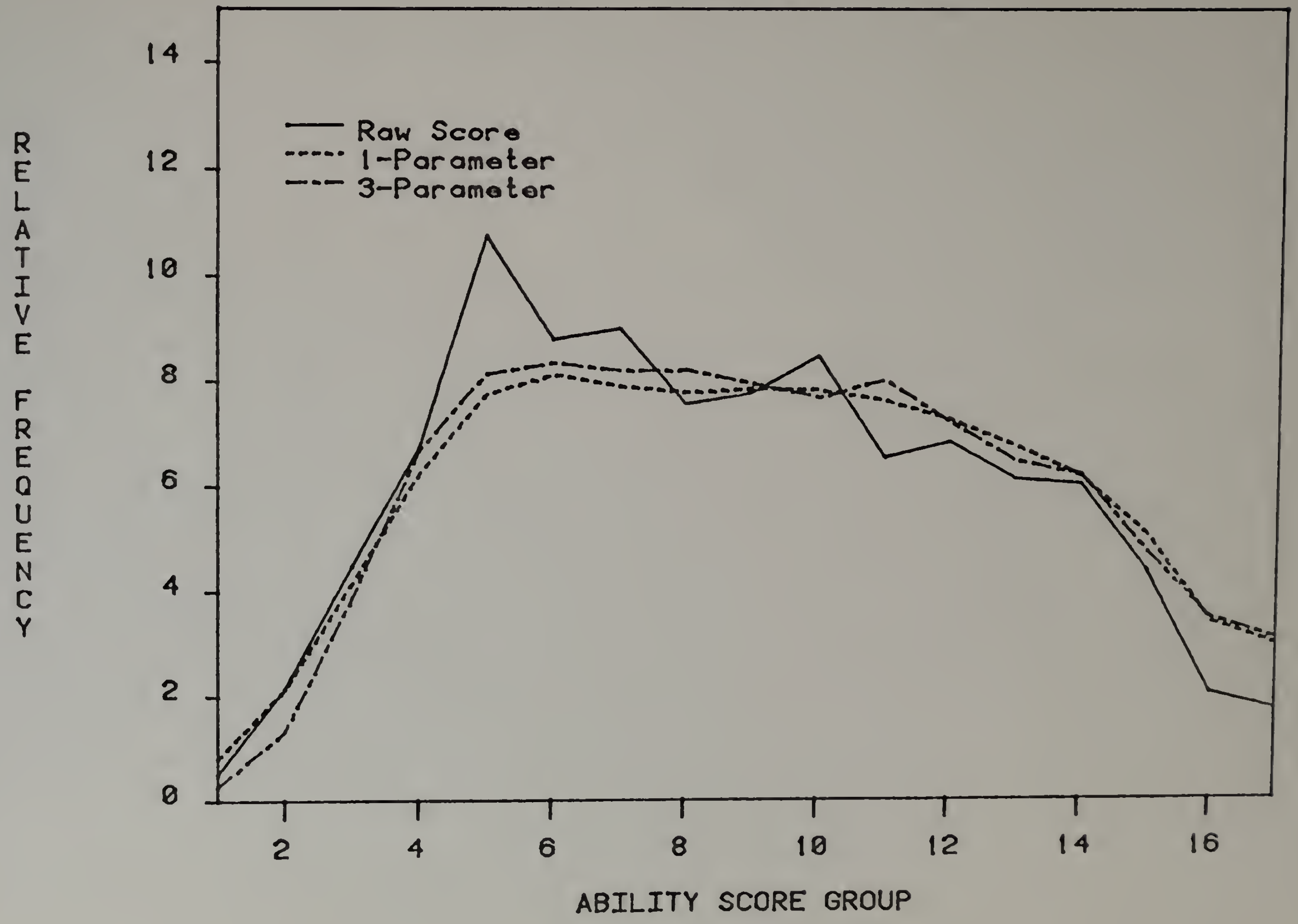


Figure 12. Observed and Expected Frequencies
Stanford Language (80 Items)

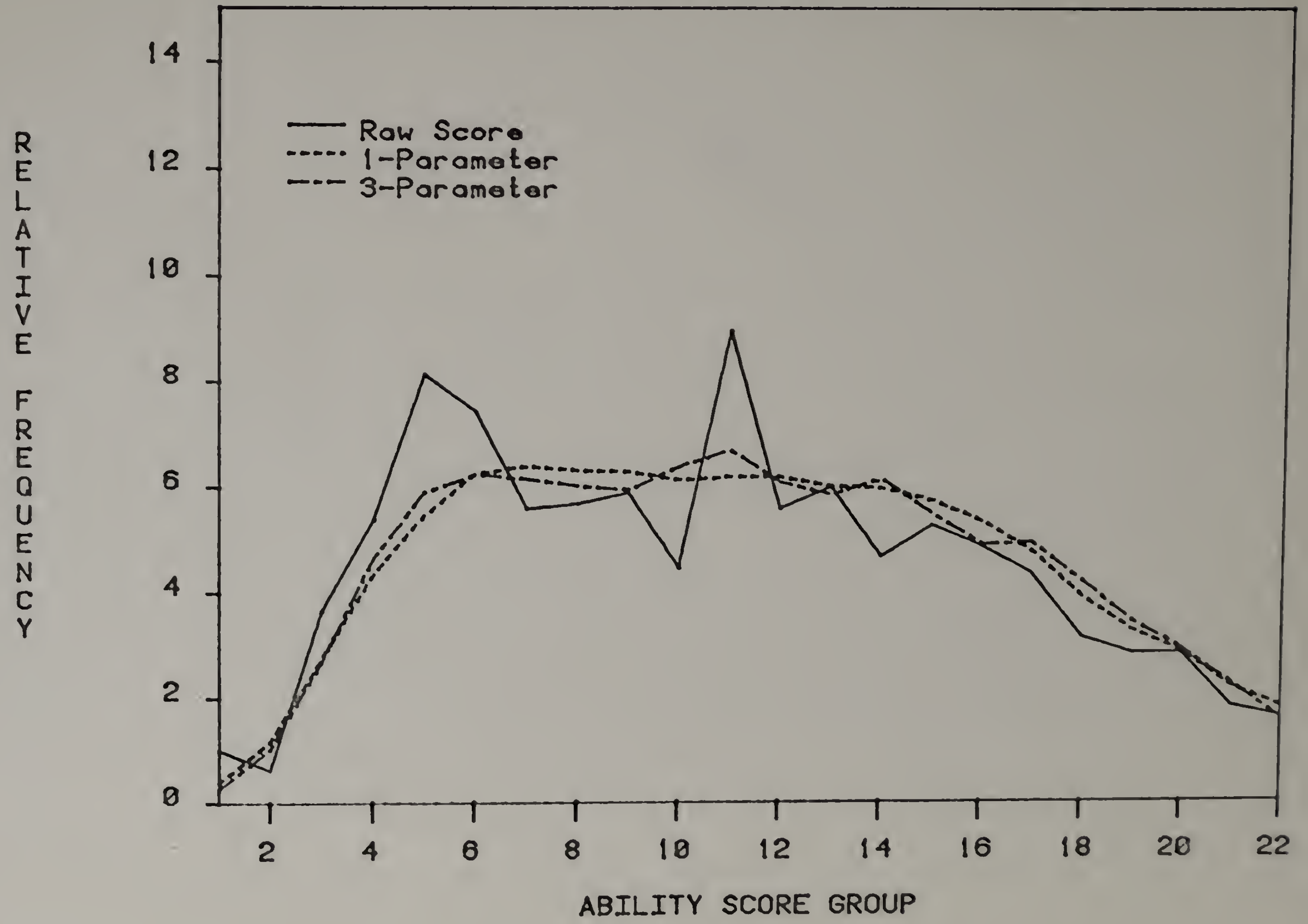


Figure 13. Observed and Expected Frequencies Stanford Social Studies

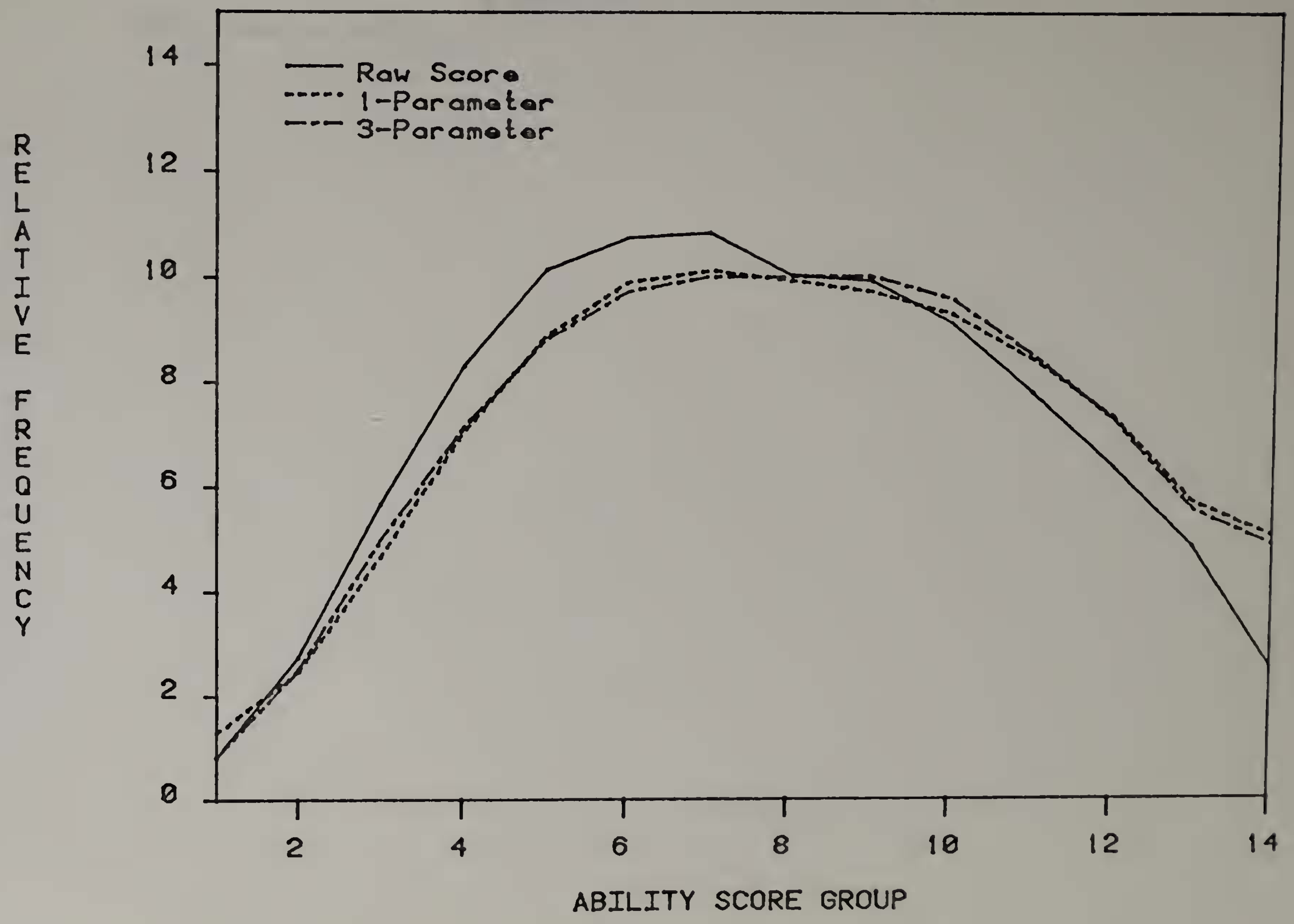


Figure 14. Observed and Expected Frequencies
Stanford Listening (50 Items)

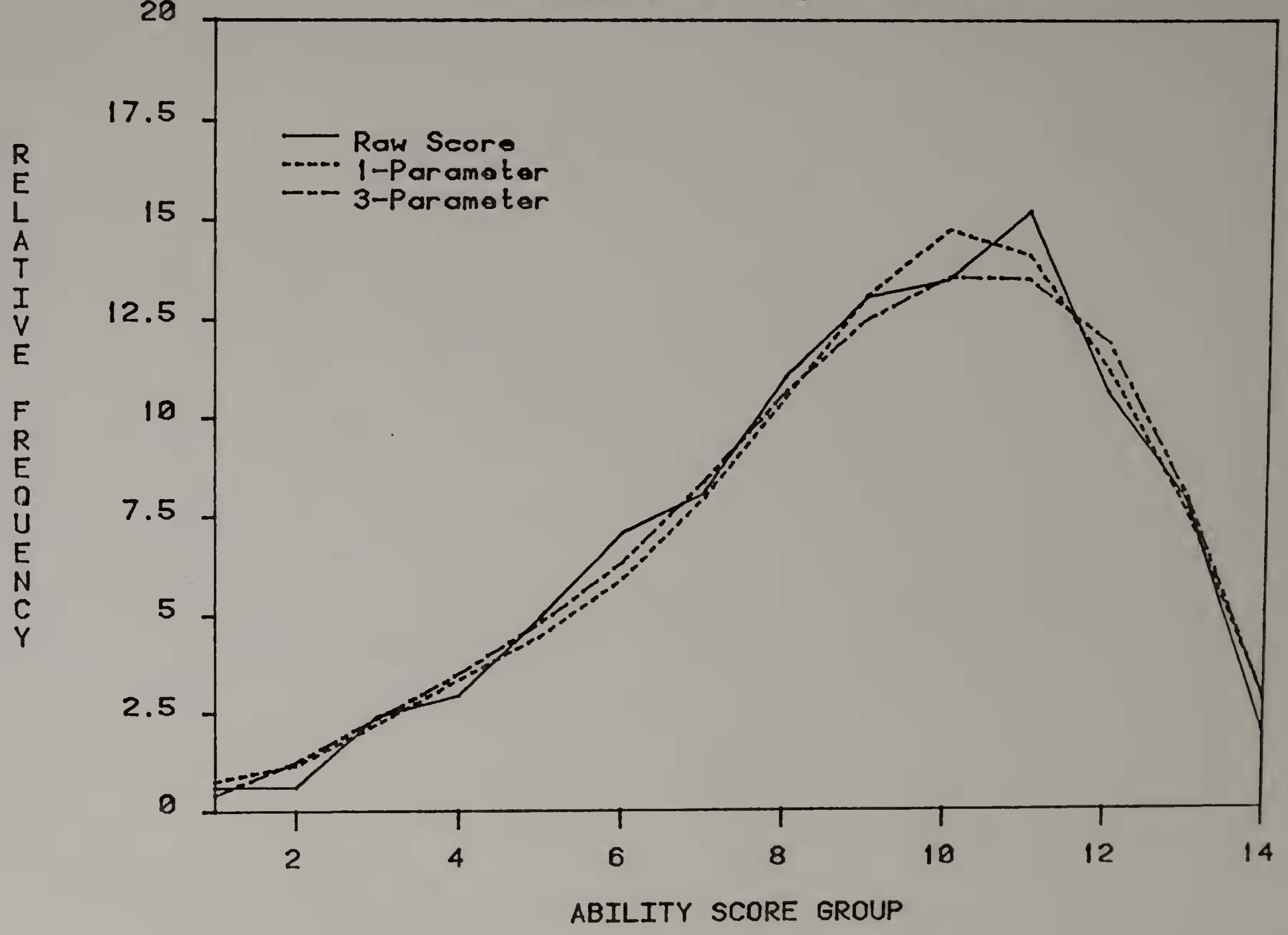


Figure 15. Observed and Expected Frequencies
Stanford Spelling (60 Items)

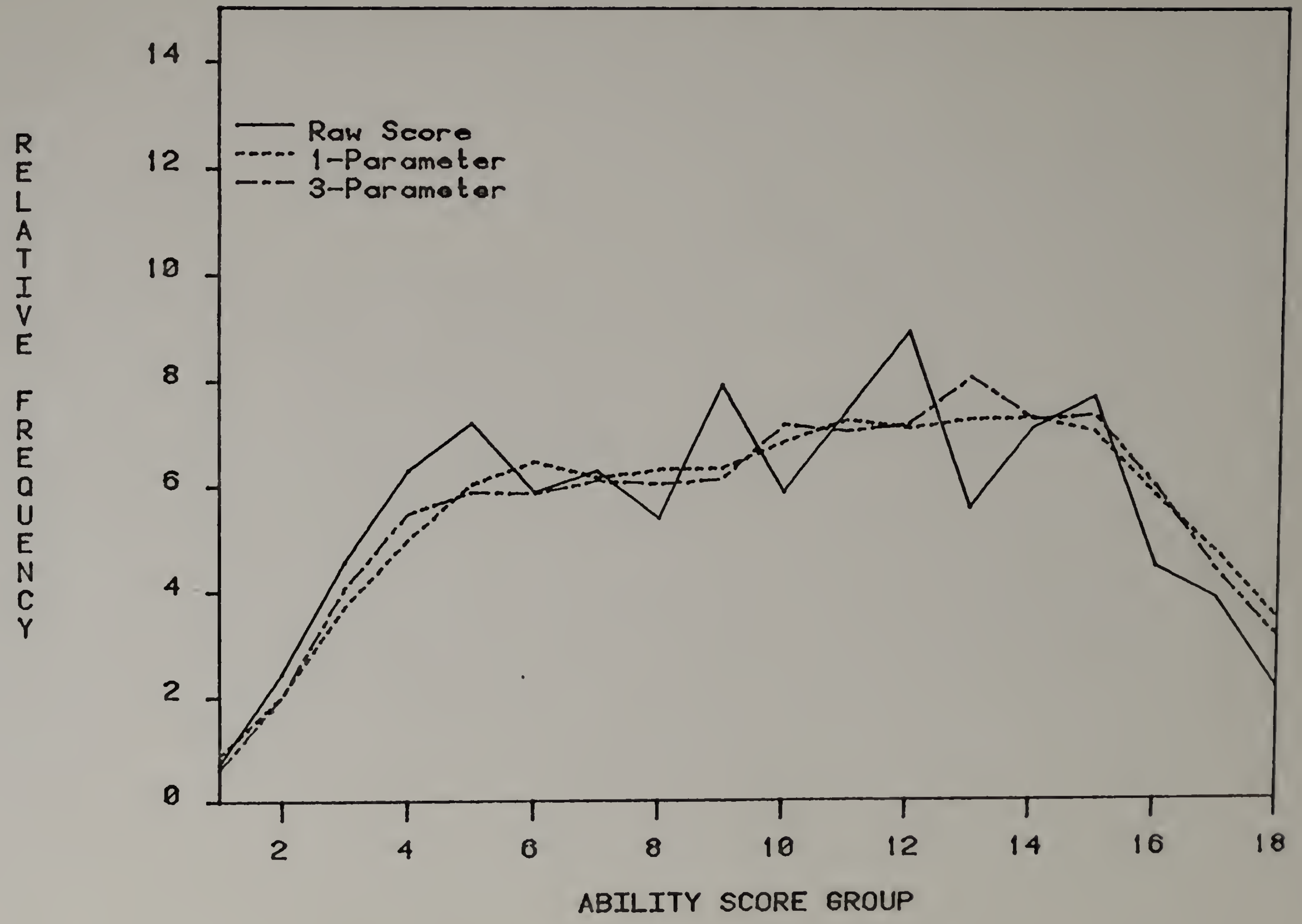


Figure 16. Observed and Expected Frequencies
Stanford Math Applications (40 Items)

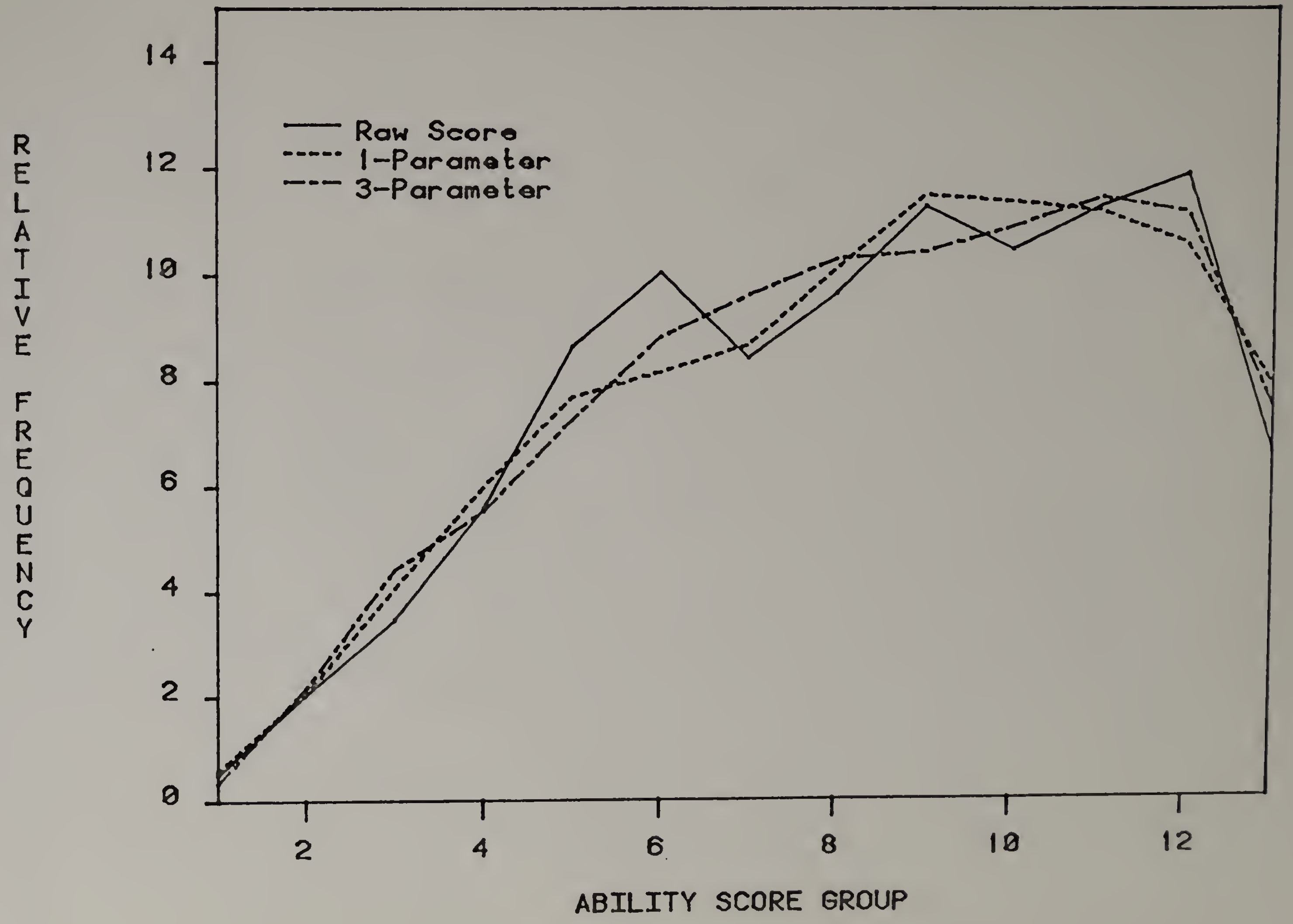


Figure 17. Observed and Expected Frequencies
Stanford Math Concepts (35 Items)

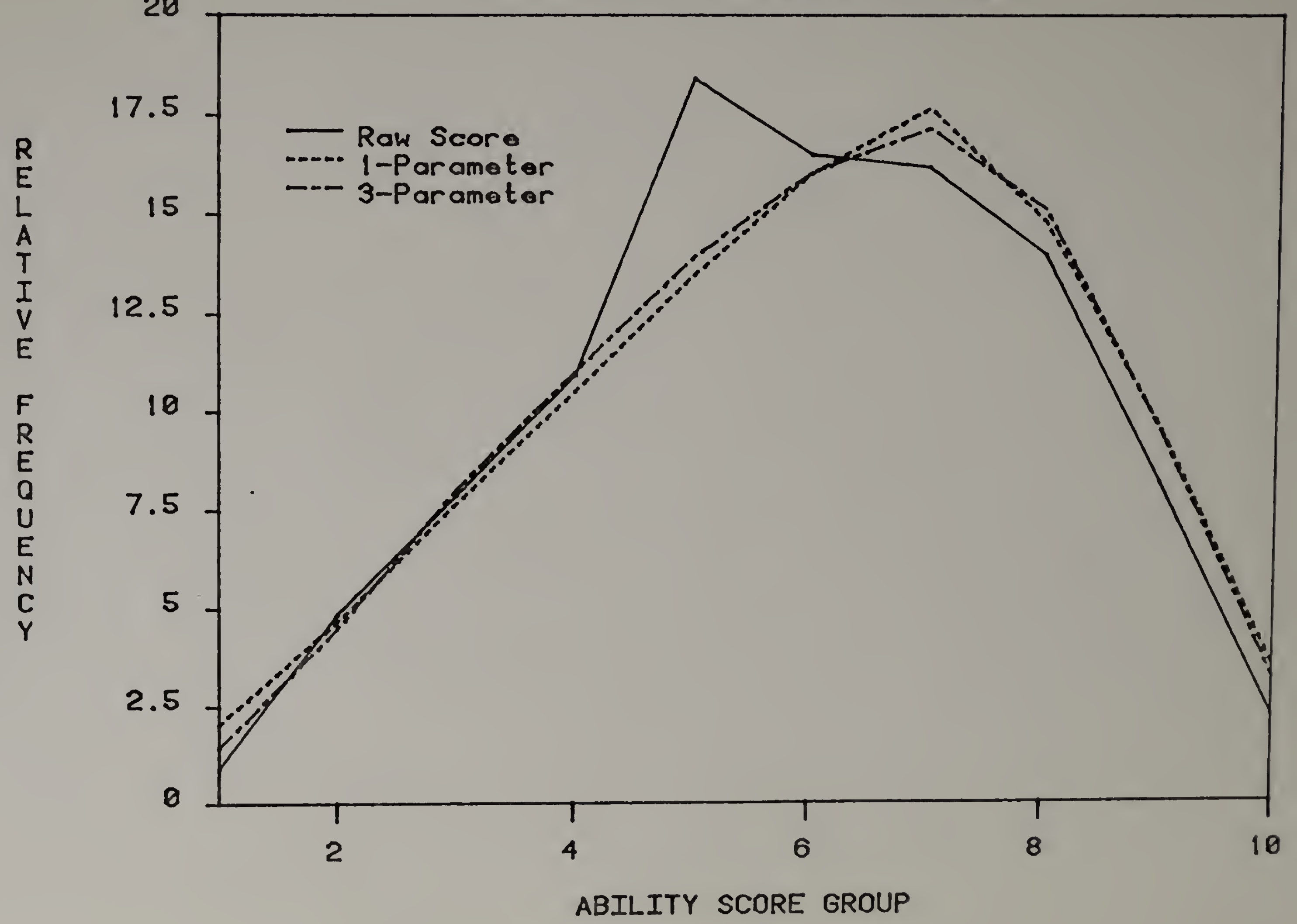


Figure 18. Observed and Expected Frequencies Individual Criterion Referenced Test (16 Items)

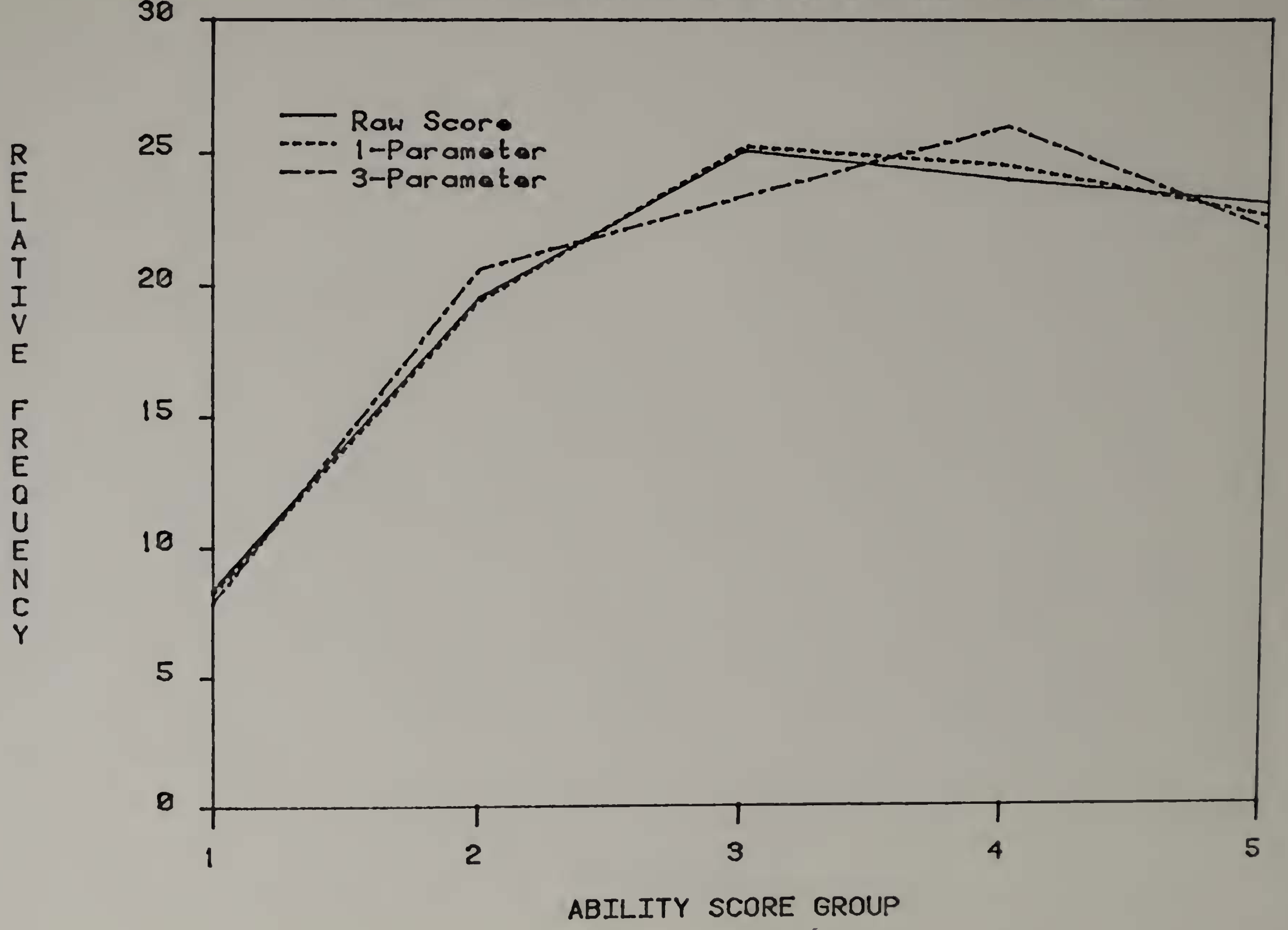


Figure 19. Observed and Expected Frequencies
SAT Verbal #2 (85 Items)

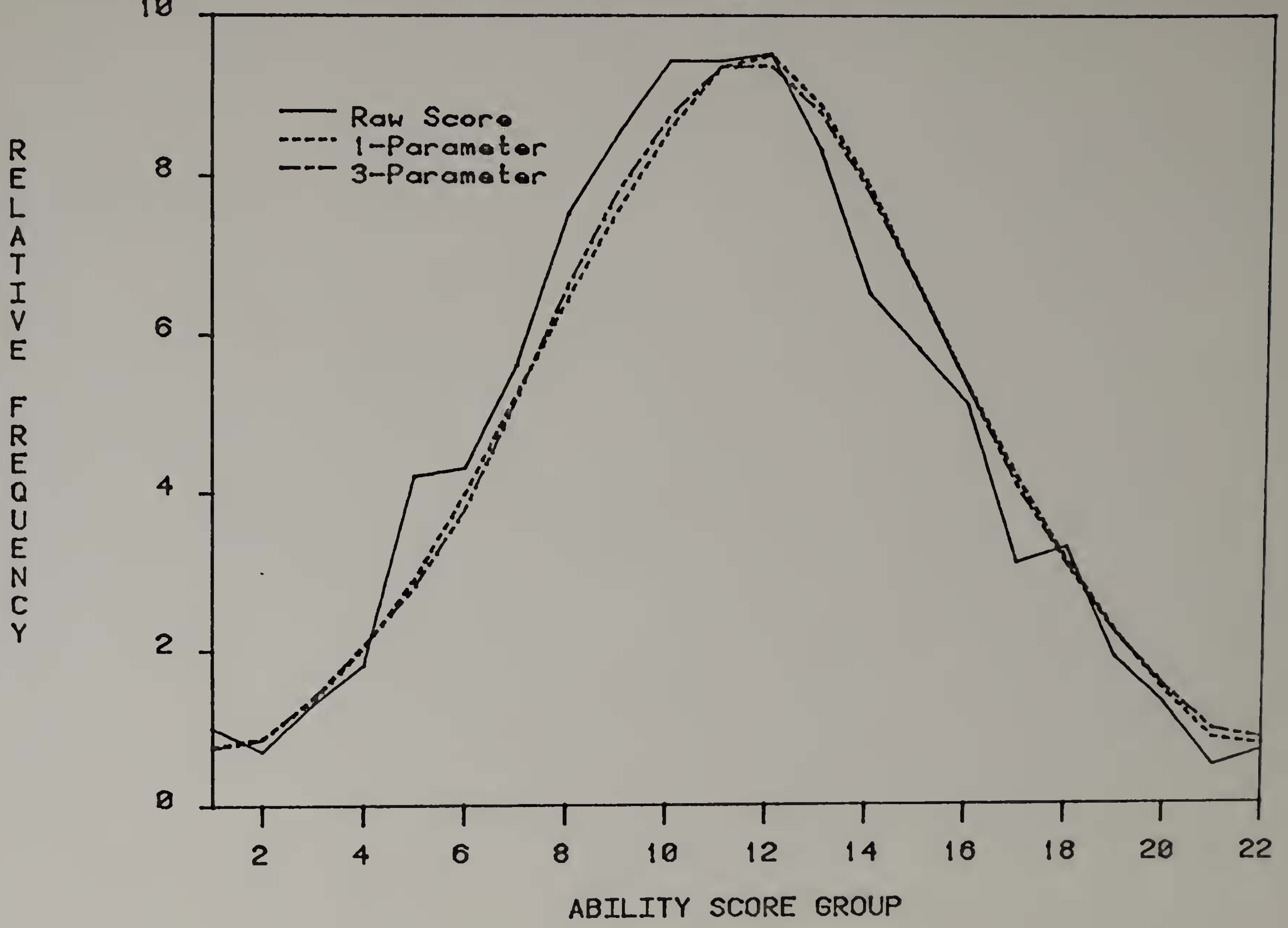


Figure 20. Observed and Expected Frequencies
Georgia Regents #1 (70 Items)

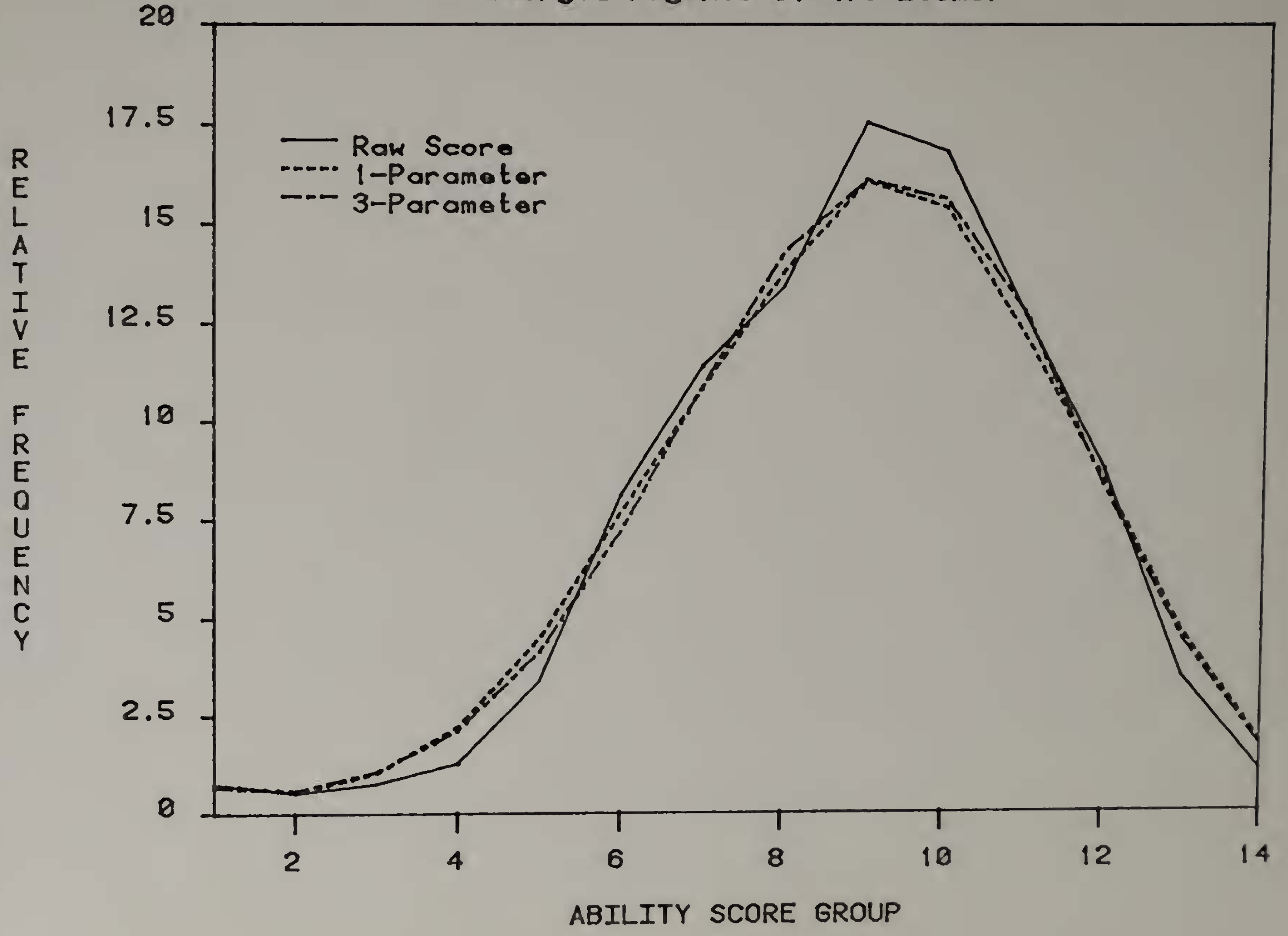


Figure 21. Observed and Expected Score Frequencies
Georgia Regents #2 (70 Items)

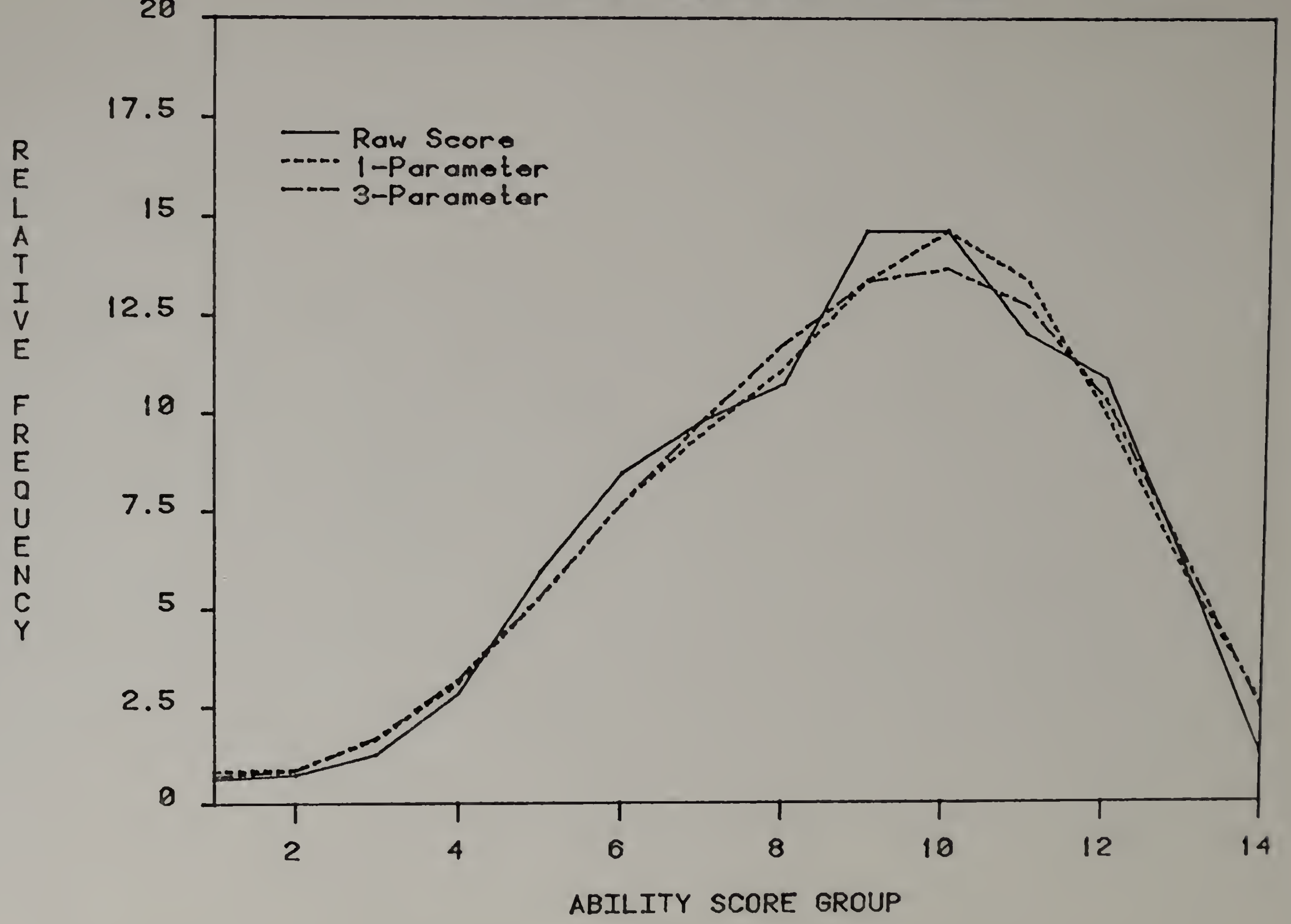


Figure 22. Observed and Expected Score Frequencies
Georgia Regents #3 (70 Items)

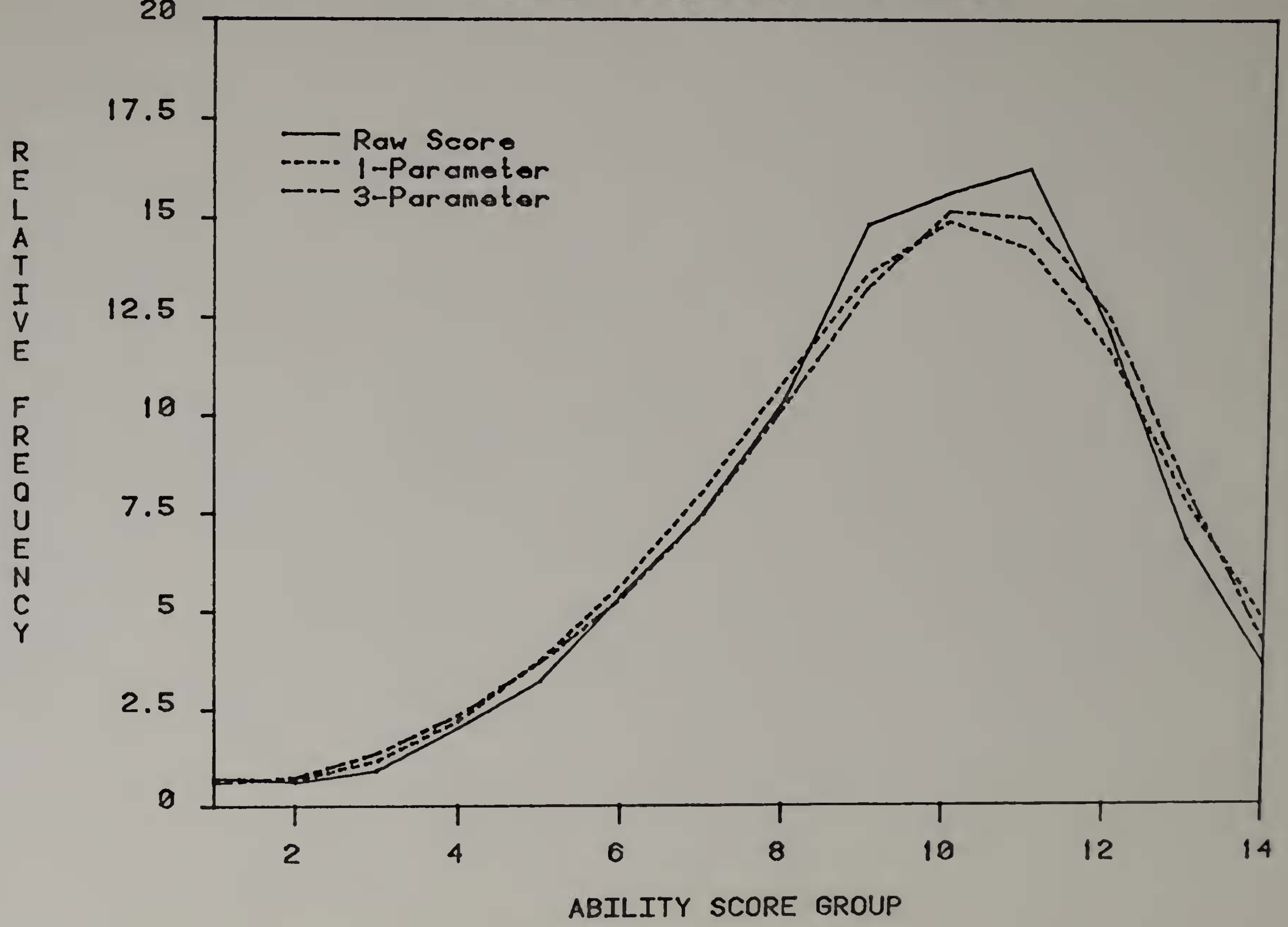


Figure 23. Observed and Expected Frequencies
CAT Vocabulary (40 Items)

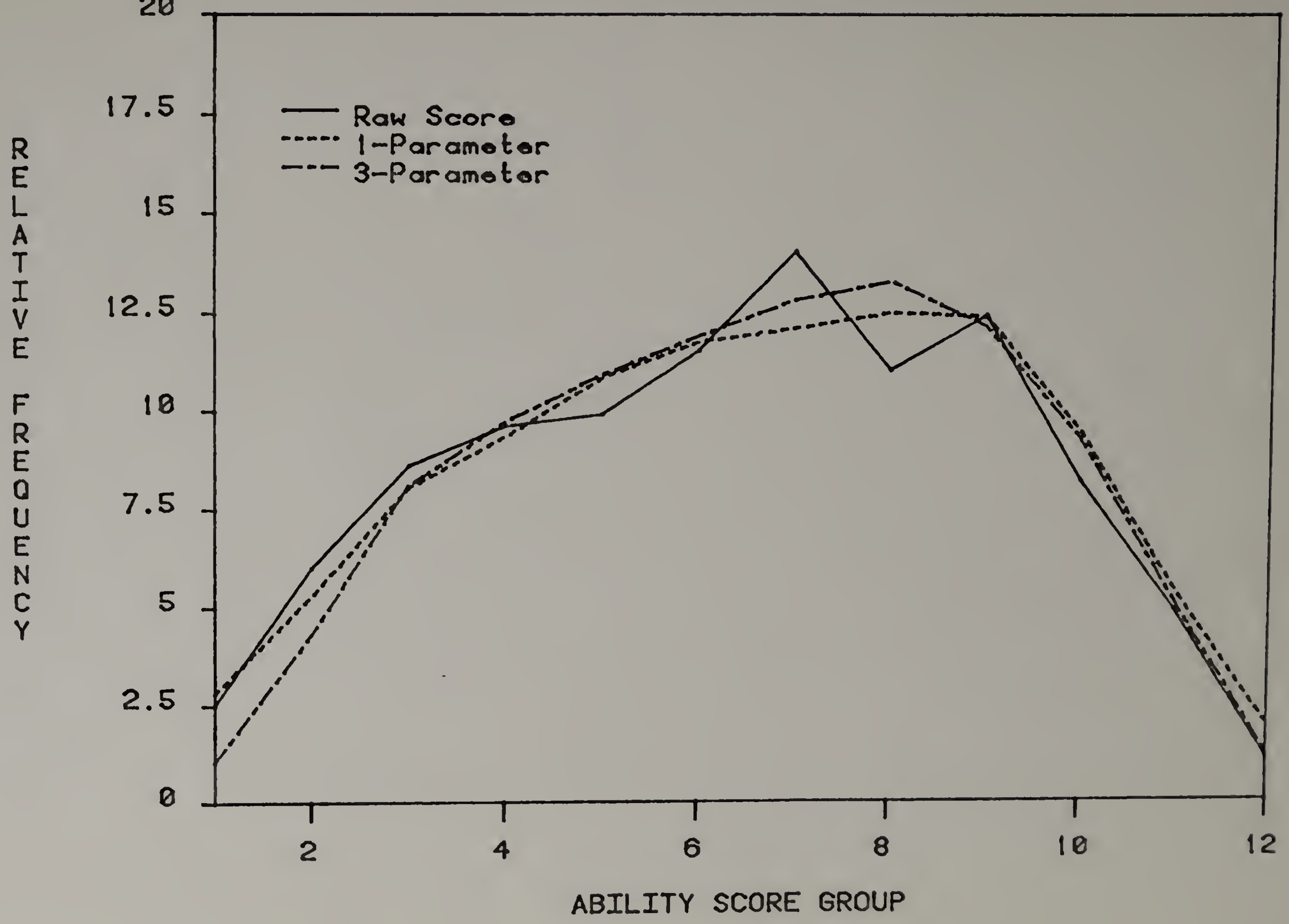


Figure 24. Observed and Expected Frequencies
ITBS Vocabulary (38 Items)

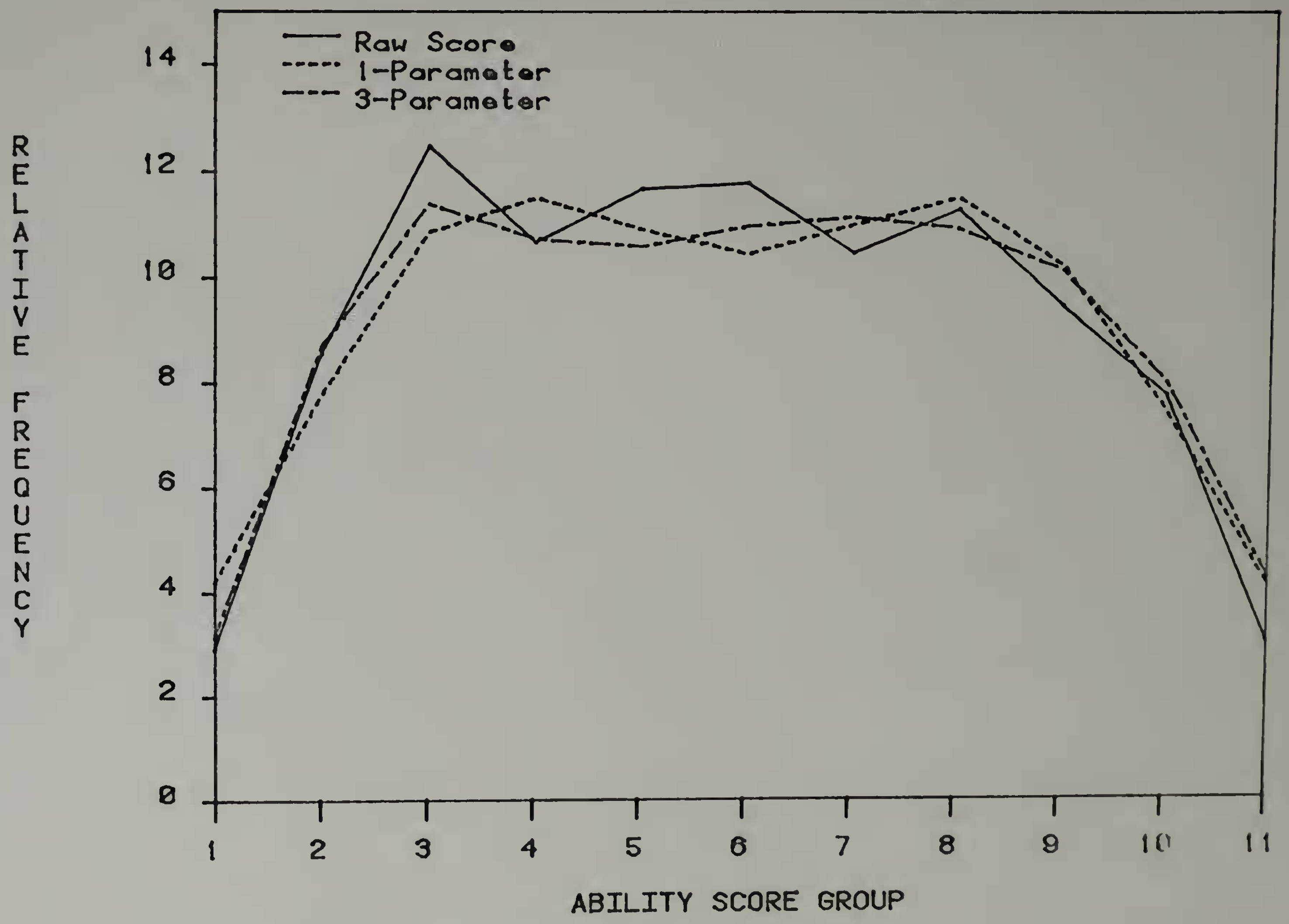


Figure 25. Observed and Expected Frequencies
CAT Comprehension (45 Items)

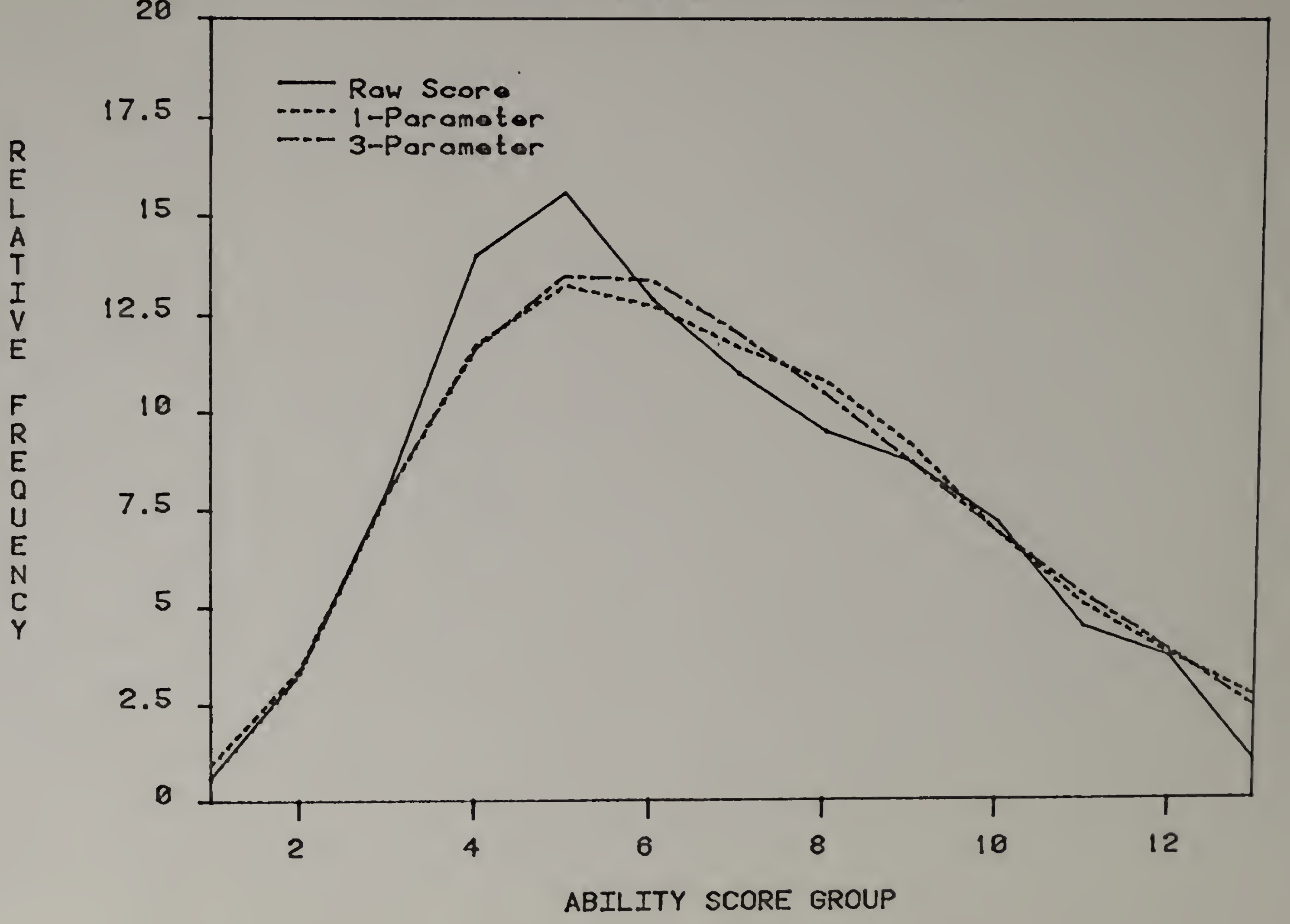


Figure 26. Observed and Expected Frequencies
ITBS Comprehension (68 Items)

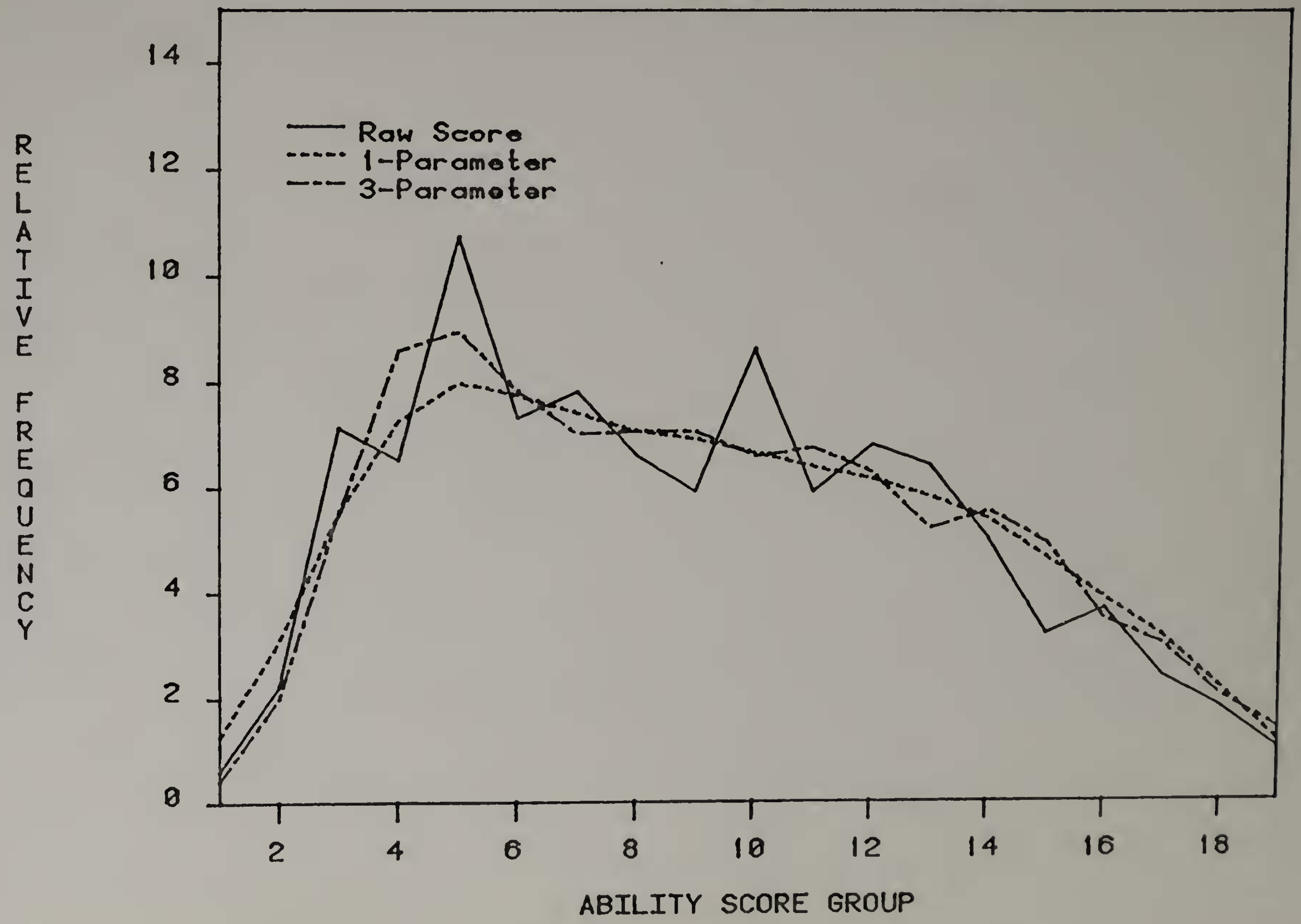


Figure 27. Observed and Expected Score Frequencies Rasch Test on Environment (34 Items)

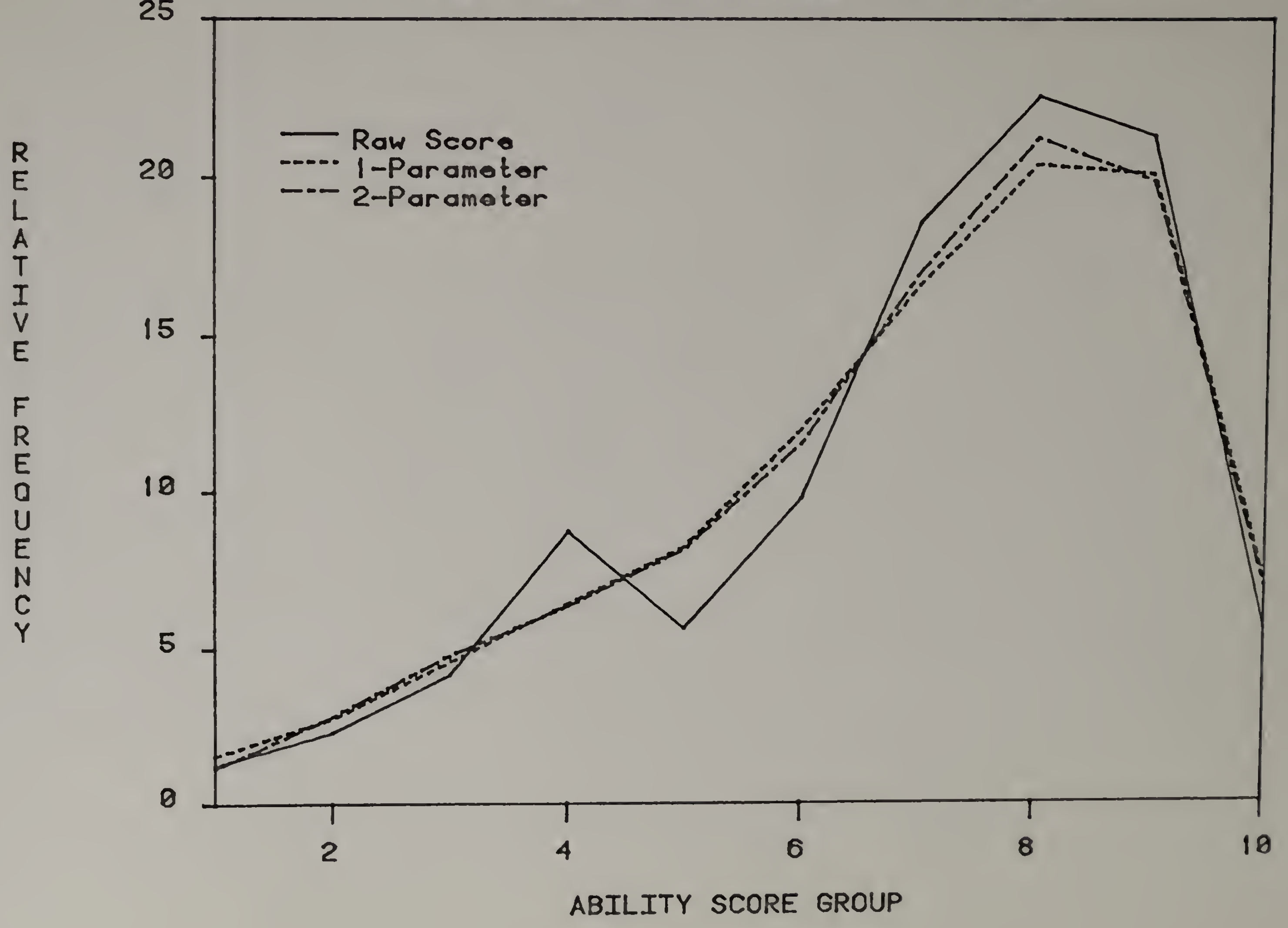
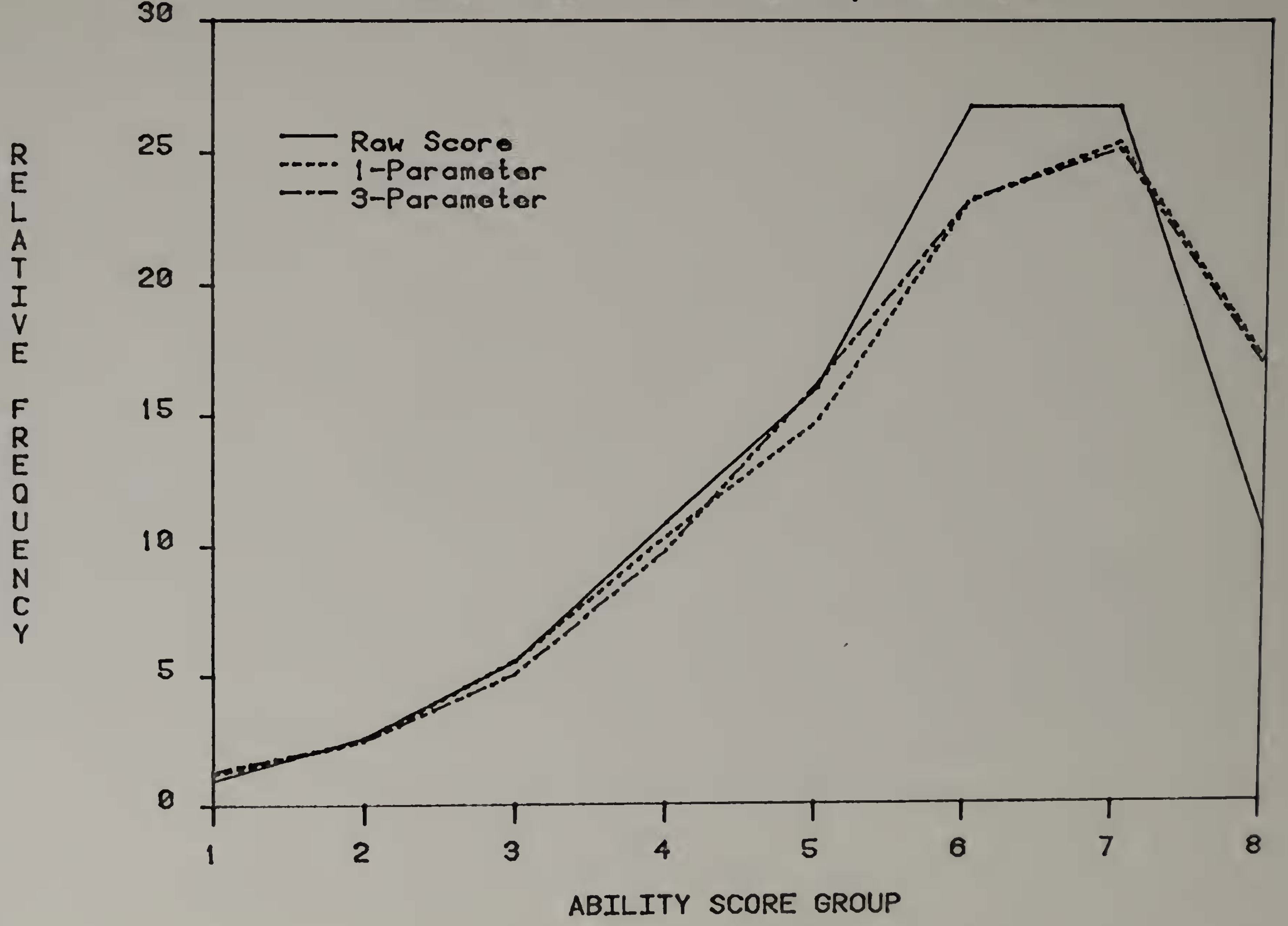


Figure 28. Observed and Expected Frequencies
Rasch Test on Citizenship (30 Items)



Overall, tests having normal score distributions fit both models better than those with uniform or skewed distributions. The chi-square test was more sensitive to departures from normality than the K-S test. Longer tests with uniformly distributed scores were often rejected by the chi-square test, although it is believed that this outcome may be attributed both to test length and to distributional form. Shorter tests, with either positively or negatively skewed number-correct score distributions, were also detected as fitting poorly by chi-square tests. The graphs indicated that latent trait models may not be very adequate for shorter tests. A discussion conderning test length is offered in a later section.

Location of Misfit of Latent Trait Models

The poorest fit of data to the latent trait models occurred in the extremes of the distributions. The graphs indicated that the number-correct score distributions predicted by the three-parameter model tended to be shifted toward the high end of the ability scale. This model predicted fewer low ability examinees and more high ability examinees than the numbers suggested by the observed score distributions. This result was more evident for skewed distributions, but appeared to be evident in other forms as well. On many tests a similar outcome was observed for the Rasch model but, on other tests, the Rasch model predicted too many examinees at both extremes of ability. Since there was considerable opportunity to guess on all of the tests used in the study, it was anticipated that the Rasch model, lacking a guessing parameter, would overestimate abilities, thereby underestimating the frequency of examinees at the low end of the ability scale. Underestimation of the frequency of low ability examinees by the three-parameter model as well, suggests that the estimated chance level parameters were too low. Discrepancies between predicted and observed number-correct score distributions were especially large for both models at the upper end of the ability distribution and tended to be the primary cause for rejection of data by statistical criteria. The results indicate that ability estimates at both extremes of the ability scale for this data were not very precise.

Concluding Remarks on Overall Model Fit

The results of this study indicate that cognitive tests, constructed with conventional test development strategies, can be nicely characterized by latent trait models. Because of the substantial amount of guessing and the lack of equal item discrimination parameters in the data, it was not anticipated that the Rasch model would fit data nearly as well as it did.

Some tests in the study had been rejected as not fitting the models when statistical criteria were used, yet graphic evidence for these same tests implied that fit was moderately good. Some psychometricians have argued that the practical relevance of a model is more important than some arbitrary statistical test of fit. The practical significance of latent trait models can only be verified within the context of specific applications of the models. For estimating abilities, model fit needs to be very good, but for other applications, for example, determining transformations for test equating, more lenient criteria for fit may be warranted.

Because irregularities due to sampling fluctuations in number-correct score distributions contributed to lack of model fit, overall model fit may actually be somewhat better than the results indicated. Nevertheless, model fit was quite poor at the extremes of the distribution, and efforts will be required to improve ability estimates in these locations.

The Relationship Between Model Fit and Violations in Latent Trait Model Assumptions

Unidimensionality

Unidimensionality was an important condition for fit of data to the Rasch and three-parameter models. Data which were multidimensional did not fit the latent trait models very well. Pearson and rank order correlations between K-S fit statistics and unidimensionality measures, shown in Table 10, were significant at the .05 level of probability. Correlations between mean square fit statistics and unidimensionality indices, also shown in Table 10, were not significant, but demonstrated a parallel trend. The negative sign of the correlations meant that high eigenratios (most unidimensional) were associated with good model fit (low fit statistics).

Generally, the tests used in this study had been designed to be unidimensional. When data tended to be multidimensional, problems

Table 10

Association Between Model Fit and Unidimensionality

Criterion	Model			
	1-Parameter		3-Parameter	
	Pearson Corr.	Spearman Corr.	Pearson Corr.	Spearman Corr.
Mean Square[1]	-.26	-.08	-.31	-.17
K-S Statistic[2]	-.46*	-.56*	-.42*	-.47*

[1] N=25 tests.

[2] N=20 tests.

*Significant, $p \leq .05$.

ensued in attempting to solve maximum likelihood equations and the estimation process took considerably longer. Cook, Eignor, and Hutten (1979) observed that when data were characterized by many factors, item parameters fluctuated wildly during estimation which frequently impeded convergence to a solution. Their data consisted of 100-item,multi-objective,criterion-referenced tests in quantitative and verbal areas.

Table 8, presented earlier, demonstrated the correspondence between K-S fit statistics and a ranking of tests based on unidimensionality and other model assumptions. Table 9 presented similar data for mean square statistics. Four of five tests which fit the models best (math applications and listening subtests of the Stanford, and Rasch developed career and environment tests) had been ranked as some of the most unidimensional tests. Based on a ranking of 25 tests, the average rank for the four tests was 18. The CTBS comprehension subtest, which also demonstrated good fit to both latent trait models, had a rank of only 10 on unidimensionality. The tests which showed the poorest fit to the Rasch and three-parameter models (reading and language subtests of the Stanford, the SAT verbal, and the 16-item ICRT) had some of the lowest unidimensionality ranks. The average rank for the four tests was 6. The poorest fitting test in the study, the Stanford reading subtest, was the second most multidimensional test. The next most poorly fitting test, the Stanford language subtest, had a dimensionality rank of 7 and had the lowest internal consistency index (KR-20) of tests in the study.

The results suggest that latent trait models are not appropriate unless data is unidimensional. Principal components analysis provided a reasonable method for determining whether a set of items was unidimensional. For data characterized as multidimensional, multivariate extensions to latent trait theory, suggested by Samejima (1974) may provide an appropriate solution. Alternatively, items can be grouped into unidimensional subsets prior to applying the latent trait methods described in this paper.

Equality of Item Discrimination Indices

A modest relationship was found between fit to the Rasch model and equality of item discrimination indices. Pearson and rank order correlations between Rasch and three-parameter model fit statistics and indices of equality of item discrimination are shown in Table 11. The negative sign of the correlations is interpreted to mean that the more homogeneous the item discriminations, the better the fit of the data to the latent trait models. Correlations between Rasch and three-parameter model fit statistics and measures of homogeneity of item discrimination, alternatively evaluated with latent trait methods, are shown in Table 12. The results in this table are very similar to those found in Table 11.

Three tests which had demonstrated the poorest fit to the latent trait models (Stanford reading, Stanford language, and the ICRT) generally had very heterogeneous item discriminations (ranked 4, 7, and 6 respectively), yet the SAT, which also did not fit the models very well, had quite homogeneous item discrimination values

Table 11

Association Between Model Fit and Equality of Item Discrimination[1]

Criterion	Model			
	1 Parameter		3-Parameter	
	Pearson Corr.	Spearman Corr.	Pearson Corr.	Spearman Corr.
Mean Square[2]	-.10	-.02	-.23	-.03
K-S Statistic[3]	-.22	-.16	-.24	-.19

[1] Based on conventional item-total score correlation.

[2] N=25 tests.

[3] N=20 tests.

Table 12

Association Between Model Fit and Equality of Item Discrimination[1]

Criterion	Model			
	1-Parameter		3-Parameter	
	Pearson Corr.	Spearman Corr.	Pearson Corr.	Spearman Corr.
Mean Square[2]	-.17	-.23	-.31	-.35
K-S Statistic[3]	-.30	-.29	-.23	-.15

[1] Based on the variance of discrimination, $\sigma\hat{a}_{g}^{2}$, estimated with latent trait methods.

[2] N=22 tests.

[3] N=17 tests.

(rank 19.5 out of 25 tests). Except for the ITBS comprehension subtest, tests which fit the latent trait models quite well, demonstrated very homogeneous discrimination values.

Previous results in this area, based on simulation techniques, have shown that the Rasch model does not fit data as well when item discrimination values have been heterogeneous. Since there were interactions between unidimensionality, equality of item discriminations, and guessing in the empirical data used in this study, partial correlation analyses were done to remove the effects of confounding variables. Partial correlations between fit statistics and measures of homogeneity of item discrimination are shown in Table 13. When correlations were controlled for other factors, there was little change in the relationship between model fit and equality of item discriminations.

Two opposite conclusions might be drawn from the results concerning item discrimination and Rasch model fit. The first is that the data used in this study had a sufficiently narrow range of item discrimination values so that, for practical purposes, they did not actually violate the Rasch model assumption of equal item discrimination and, consequently, the Rasch model fit rather well. The second is that the Rasch model fit data despite the fact that item discriminations were not equal, i.e., the Rasch model was robust to violation of the assumption. The results of this study are contrasted to those from other studies to explore which conclusion is appropriate.

Table 13

Partial Correlations Between Model Fit and Equality of Item Discrimination[1],[2]

Correlation	1-parameter model	3-parameter model
Simple Correlation		
Mean Square	-.25	-.40
K-S Statistic	-.29	-.33
1st Order Partial-Unidimensionality Partialled Out		
Mean Square	-.21	-.35
K-S Statistic	-.21	-.26
1st Order Partial-Guessing Partialled Out		
Mean Square	-.26	-.42
K-S Statistic	-.36	-.41
2nd Order Partial-Guessing and Unidimensionality Partialled Out		
Mean Square	-.20	-.38
K-S Statistic	-.28	-.33

[1] Based on N=21 tests (tests with less than 30 items and less than 900 examinees were excluded).

[2] Conventional item-total score correlations were used to estimate equality of item discrimination.

Hambleton and Cook (1978) and Dinero and Haertel (1977) found a weak relationship between homogeneity of item discrimination values and fit of data to the Rasch model. Hambleton and Traub (1971), on the other hand, found that the presence of unequal discrimination values significantly reduced fit of data to the Rasch model. The discrimination parameters in the simulation studies had been generated to span a fairly wide range. Discrimination parameters found in the standardized tests used in this study were thought to be relatively homogeneous.

The ranges of discrimination values for the empirical data in this study are compared to the ranges reported in other studies in Table 14. In this study and the Dinero and Haertel study, the ranges were estimated from the variances of discrimination parameters. In the other studies, discrimination parameters had been generated within the reported ranges. The minimum, maximum, and average ranges of discrimination values are reported for the empirical data. Hambleton and Traub (1971) concluded that when the range of item discriminations was greater than .2, the Rasch model failed to fit simulated data. They observed that a range of discrimination values of .8 is more commonly seen in real data. Hambleton and Cook (1978, p. 8) generated narrow and broad ranges of discrimination values and concluded: "For the values studied in the paper, using discrimination parameters as weights contributed very little to the proper ranking of examinees." Dinero and Haertel (1977, p. 14) examined five variances for item discrimination and concluded: "The present research suggests that the lack of an item discrimination parameter in the Rasch model does not result in poor calibration in the presence of varying discriminations." The comparative results in Table 14 indicate that the ranges of discrimination values in this study were as broad as those explored in the simulation studies. Nevertheless, in the present study, the lack of homogeneity of item discrimination values did not appreciably reduce fit of data to the Rasch model. It is concluded that the Rasch model can tolerate heterogeneity of item discrimination indices. It is not clear why the results of Hambleton and Traub (1971) differ from those in the other studies. One difference between the studies is that the more recent studies utilized sophisticated computer methods for estimation, in contrast to the approximate solutions used by Hambleton and Traub. There is not a consensus on how wide a range of discrimination values can be tolerated by the Rasch model, but if the data in this study are representative of tests used in practice, the question of importance of homogeneous item discriminations for the Rasch model may be moot.

There was also a relationship found between heterogeneity of item discrimination and lack of fit to the three-parameter model. One explanation for this outcome is that item discrimination parameters may not have been estimated very precisely.

Table 14

Summary of Studies which Explored Rasch Model Fit to Data with Varying Item Discrimination Values

Study	Ability Distribution	Number of Examinees	Number of Items	Item Discrimination Distribution	Range of Discrimination	Endpoints of Range
Hambleton & Traub (1971) SIMULATION	Normal	1000	15,30,45	Uniform (mean = .6)	.2 .4 .8	.5- .7 .4- .8 .2-1.0
Dinero & Haertel[1] (1977) SIMULATION	Normal	75	30	Normal and Uniform (mean = 1.0)	.44 .64 .78 .88 1.00	.78-1.22 .68-1.33 .61-1.39 .56-1.44 .50-1.50
Hambleton & Cook (1978) SIMULATION	Normal and Uniform	500	20,40	Uniform (mean = 1.12)	.00 .62 1.24	1.12 .81-1.43 .50-1.74
Hutten[2] (1981) EMPIRICAL	Normal, Uniform, and Skewed	1000	16-85	None assumed (mean = .97)	.47 .78 1.15	.50- .97 .61-1.39 .36-1.51

[1] Variance of item discriminations was reported. Range was approximated as two standard deviations.

[2] Range was calculated from the variance as two standard deviations. Minimum, average, and maximum ranges from the study are shown.

Guessing

Results concerning the effect of guessing on Rasch model fit were not conclusive. Pearson and rank order correlations between estimates of guessing based on conventional item difficulty levels and fit statistics for the Rasch and three-parameter models are presented in Table 15. These data show the surprising result that both the Rasch and three-parameter models fit data better when examinees guessed. A complimentary set of results, based on guessing estimated with latent trait methods, is shown in Table 16. Results in this table suggest that the Rasch and three-parameter models had fit data more poorly when examinees guessed. The rank order correlation between the latent trait estimate of guessing and fit to the three-parameter model was significant at the .05 level of probability.

Some might argue that the latent trait estimates of guessing were not very good (particularly proponents of the Rasch model), but it seemed more plausible to this author that the contridictory results, reported above, can be attributed to poor estimation of guessing with conventional methods. The two approaches to estimating guessing differed significantly.

Table 15

Relationship Between Model Fit and Guessing[1]

Criterion	Model			
	1-Parameter		3-Parameter	
	Pearson[2] Corr.	Spearman[3] Corr.	Pearson Corr.	Spearman Corr.
Mean Square[4]	-.35	.25	-.42	.42
K-S Statistic[5]	-.43	.23	-.36	.08

[1] Based on conventional estimate for guessing.

[2] A negative correlation associate guessing with good model fit.

[3] Since a high rank was assigned to tests with the least guessing, a positive correlation associates guessing with good model fit.

[4] N=16 tests (4 or more hard items per test).

[5] N=12 tests (4 or more hard items per test).

Table 16

Relationship Between Model Fit and Guessing[1]

Criterion	Model			
	1-Parameter		3-Parameter	
	Pearson[2] Corr.	Spearman[3] Corr.	Pearson Corr.	Spearman Corr.
Mean Square[4]	.44	-.24	.51	-.59*
K-S Statistic[5]	.30	-.24	.36	-.36

[1] Based on average chance level parameter, $\bar{c}$, for hard items estimated by latent trait methods.

[2] A positive correlation associates guessing with poor model fit.

[3] A negative correlation associates guessing with poor model fit since high ranks were assigned tests with less guessing.

[4] N=14 tests (4 or more hard items per test).

[5] N=10 tests (4 or more hard items per test).

*Significant, $p \leq .05$.

In latent trait theory, the lower asymptote of the ICC is assumed to provide a measure of guessing. The lower asymptote gives the probability of obtaining a correct answer by chance alone and, as such, is a characteristic of the item. The conventional method gave the percentage of low ability examinees who actually answered difficult itemscorrectly, presumably by chance, and was more a characteristic of the examinees. Conventional item difficulty levels are unfortunately sample dependent and, consequently, no method existed for equating the difficulties for items across tests. Since latent trait parameter estimates are sample-invariant, they allow comparisons between tests.

Another difficulty with the conventional procedure for estimating guessing was that there was no way to equate what "low ability" meant across samples. Conventional difficulties were based on the bottom ten percent of examinees, who were assumed to be of low ability. There was no method for verifying this assumption. The lowest decile group on each test may have had significantly different levels of ability.

Another criticism of the conventional guessing estimates is that they were based on the scores of _all_ examinees regardless of whether they answered an item or not. Conventional test scoring assumes that omitted items are incorrect. Latent trait methods do not base information on omitted items.

Both conventional and latent trait estimates of guessing were based on a small number of examinees, 100 or less, in this study. Neither estimate may be very accurate considering the modest number of low ability examinees represented in the samples in this study.

Given the foregoing problems, guessing results based on conventional item difficulties were suspect. These estimates could represent some other quality of the data, but no relationships with other variables emerged.

In the remainder of this discussion, it is assumed that the latent trait chance level parameter was a more accurate measure of guessing. Given this assumption, the fact that the three-parameter model also fit data rather poorly when there was evidence of guessing needs to be explained.

When ability estimates are not corrected for the chance level parameter they tend to be too high. This phenomena is illustrated with simulated data in Tables 17 to 19. Table 17 provides ability estimates at various probability levels for an item of average difficulty (b_g=0.0) when the estimate of guessing is varied. Table 18 shows ability estimates for a difficult item (b_g=-2.0) given different probability levels and estimates of guessing. Table 19 illustrates the error in ability estimates that results when true values of the lower asymptote are greater than zero by various amounts. The tables show the extent to which ability is overestimated when the chance parameter is assumed to be too low or zero. Overestimation of ability at the low end of the scale would result in

Table 17

The Effect of Underestimating Guessing on Ability Estimates for an Item of Average Difficulty[1]

	Ability Estimates ($\hat{\theta}$) for Estimated Values of $\hat{c}_g$					
Pr(θ)	$\hat{c}_g$=0.0	$\hat{c}_g$=.10	$\hat{c}_g$=.20	$\hat{c}_g$=.25	$\hat{c}_g$=.30	$\hat{c}_g$=.50
0.00	- ∞	X	X	X	X	X
.05	-2.944	X	X	X	X	X
.10	-2.197	- ∞	X	X	X	X
.15	-1.735	-2.833	X	X	X	X
.20	-1.386	-2.079	- ∞	X	X	X
.25	-1.099	-1.609	-2.708	- ∞	X	X
.30	- .847	-1.253	-1.946	-2.639	- ∞	X
.35	- .619	- .955	-1.466	-1.871	-2.565	X
.40	- .405	- .693	-1.099	-1.386	-1.792	X
.45	- .201	- .452	- .788	-1.012	-1.299	X
.50	0.0	- .223	- .511	- .693	- .916	- ∞
.55	.201	0.0	- .251	- .405	- .588	-2.197
.60	.405	.223	0.0	- .133	- .288	-1.386
.65	.619	.452	.251	.133	0.0	- .847
.70	.847	.693	.511	.405	.288	- .405
.75	1.099	.955	.788	.693	.588	0.0
.80	1.386	1.253	1.099	1.012	.916	.405
.85	1.735	1.609	1.466	1.386	1.299	.847
.90	2.197	2.079	1.946	1.871	1.792	1.386
.95	2.944	2.833	2.708	2.639	2.565	2.197
1.00	+ ∞	+ ∞	+ ∞	+ ∞	+ ∞	+ ∞

[1] a_g=1.0; b_g=0.0.

Table 18

The Effect of Underestimating Guessing on Ability Estimates for a Hard Item[1]

Pr (θ)	Ability Estimates (θ) for Estimated Values of $\hat{c}_g$					
	$\hat{c}_g$=0.0	$\hat{c}_g$=.10	$\hat{c}_g$=.20	$\hat{c}_g$=.25	$\hat{c}_g$=.30	$\hat{c}_g$=.50
0.00	- ∞	X	X	X	X	X
.05	-4.944	X	X	X	X	X
.10	-4.197	- ∞	X	X	X	X
.15	-3.735	-4.833	X	X	X	X
.20	-3.386	-4.079	- ∞	X	X	X
.25	-3.099	-3.609	-4.708	- ∞	X	X
.30	-2.847	-3.253	-3.946	-4.639	- ∞	X
.35	-2.619	-2.955	-3.466	-3.871	-4.565	X
.40	-2.405	-2.693	-3.099	-3.386	-3.792	X
.45	-2.201	-2.452	-2.788	-3.012	-3.299	X
.50	-2.000	-2.223	-2.511	-2.693	-2.916	- ∞
.55	-1.799	-2.000	-2.251	-2.405	-2.588	-4.197
.60	-1.595	-1.777	-2.000	-2.133	-2.288	-3.386
.65	-1.381	-1.548	-1.749	-1.867	-2.000	-2.847
.70	-1.153	-1.307	-1.489	-1.595	-1.712	-2.405
.75	- .901	-1.045	-1.212	-1.307	-1.412	-2.000
.80	- .614	- .747	- .901	- .988	-1.084	-1.595
.85	- .265	- .391	- .534	- .614	- .701	-1.153
.90	.197	.079	- .054	- .129	- .208	- .614
.95	.944	.833	.708	.639	.565	.197
1.00	+ ∞	+ ∞	+ ∞	+ ∞	+ ∞	+ ∞

[1] a_g=1.0; b_g=-2.00.

Table 19

Error in Estimating Ability ($\hat{\theta}$) when the Rasch Model (No Guessing) is Assumed and There is Actually Guessing

	Error in Rasch Ability Estimates ($\hat{\theta}$) for Various True Values of c_g				
Pr(θ)	c_g=.10	c_g=.20	c_g=.25	c_g=.30	c_g=.50
0.00	a	a	a	a	a
.05	a	a	a	a	a
.10	a	a	a	a	a
.15	1.098	a	a	a	a
.20	.693	a	a	a	a
.25	.511	1.609	a	a	a
.30	.406	1.099	1.792	a	a
.35	.336	.847	1.252	1.946	a
.40	.288	.694	.981	1.387	a
.45	.251	.587	.811	1.098	a
.50	.223	.511	.693	.916	a
.55	.201	.452	.606	.789	2.398
.60	.182	.405	.538	.693	1.791
.65	.167	.368	.486	.619	1.466
.70	.154	.336	.442	.559	1.252
.75	.143	.310	.405	.510	1.098
.80	.133	.287	.374	.470	.981
.85	.126	.269	.349	.436	.888
.90	.118	.251	.326	.405	.811
.95	.111	.236	.305	.379	.747
1.00	b	b	b	b	b

[a] Error approaches infinity.

[b] Error approaches zero.

an underestimation of the number of examinees in the lowest score groups.

Other results in the study showed that the numbers of examinees predicted in the lower score groups by both the Rasch and three-parameter models were too low. These results could have been obtained either because chance level parameter estimates were too low (or zero in the case of the Rasch model), or because other item parameters had been poorly estimated. Swaminathan and Gifford (1979) demonstrated that maximum likelihood estimates of difficulty and discrimination with the LOGIST procedure were rather good in contrast to estimates of guessing. This suggests that in this study guessing estimates had been underdetermined.

Overall, guessing estimates were obtained for only 25 percent of items; the remainder of items retained their initial values throughout the estimation process. This undesirable situation was attributed to the fact that too few low ability examinees were represented in the samples. Other parameter estimates were based on 1000 examinees.

Although the relationship between guessing and model fit was not significant, the results suggest that the guessing parameter may be quite important for estimating ability. Unfortunately, the maximum likelihood estimate for guessing was not well determined on samples of 1000. These difficulties suggest that other methods for dealing with guessing, such as those suggested by Waller (1974a, 1974b, 1976) might provide more useful approaches.

The Effects of Sample Size and Test Lengh on Precision of Latent Trait Parameter Estimates

Ability Estimation on Short Tests

Three measures of association between abilities estimated on short (20-item) and longer tests are shown in Table 20. The high Pearson and rank order correlations between abilities on short and long tests indicate that ability estimates on short tests were very good. A third measure, the average absolute difference (AAD) between the two sets of estimates, was approximately one-third of a standard deviation and supported the information provided by the correlations. Abilities estimated on the short test with the Rasch model were somewhat better than those estimated with the three-parameter model. Although the Rasch ability estimates were more consistent, it cannot be concluded that they were also more valid. The AAD statistic can be viewed as an approximate measure of error in the short test ability estimates. Based on a practical range in ability from -4 to +4, or 8 units, a .3 difference in ability estimates represents an error of 3.7 percent for the Rasch model and the .37 difference represents an error of 4.6 percent for the three-parameter model. These results suggest that ability estimates from short tests were quite precise with both models. This result is particularly important for tailored testing, but means, in general, that test administration times need not be substantial when latent trait methods are used to score data.

Table 20

Three Measures of Precision of Ability Estimates from Short and Long Tests[1]

Statistic	3-Parameter Model	1-Parameter Model
Pearson Correlation	.866	.923
Spearman Correlation	.918	.926
Av. Abs. Difference	.372	.300

[1] Based on n=5 tests. Short tests were 20 items. Long tests were over 40 items.

Item Parameter Estimation on Small Samples

Item parameters estimated on small samples (N=250) are compared with those estimated on larger samples (N=1000) in Table 21. Correlations between difficulties estimated from the small and large samples are quite high for both the Rasch and three-parameter models. The difficulties estimated with the Rasch model are slightly more precise than those estimated with the three-parameter model. Based on a practical range of difficulty from -2 to +2, an AAD of .153 represents a 3.8 percent difference in difficulties estimated on the two different size samples for the three-parameter model. The corresponding percentage error, based on an AAD of .22, was 3.0 percent for the Rasch model.

Item discrimination estimates from the small sample were not as good as difficulty estimates. Since item discrimination had a practical range of 0 to +2, only two units, an AAD statistic of .407 represented a 20 percent difference in item discrimination estimates from the small and large samples. Correlation between the two sets of discrimination estimates were reasonably high, although the magnitudes of the discriminations were quite disparate in the smaller and larger samples. The results suggest that samples of 250 examinees were too small for estimating item discrimination.

Estimates of guessing from the small and larger samples had only a modest correlation, but were very close in magnitude. Since the range of guessing parameters is rather narrow (approximately 0 to .20 for this data), the values of the parameters in the two samples differed by approximately 15 percent. This error was somewhat less

Table 21

Three Measures of Precision of Item Parameter Estimates from Small and Large Samples[1]

Statistic	3-Parameter Model			1-Parameter Model
	$\hat{a}_g$	$\hat{a}_g$	$\hat{a}_g$	$\hat{b}_g$
Pearson Correlation	.833	.974	.413	.987
Spearman Correlation	.830	.975	.478	.983
Av. Abs. Diff.	.407	.513	.030	.122

[1] Based on n=5 tests. Small samples were 250 examinees. Large samples were 1000 examinees.

than the error in estimating item discrimination values. This result is attributed to the fact that 75 percent of guessing parameters in the study were inestimable and kept at their initial values. The initial values were the same regardless of sample size since they were computed as one divided by the number of choices on an item. The low correlation between guessing parameters estimated on the different size samples indicates that guessing estimates on the small sample were very poor. Considering that 75 percent of the guessing estimates were identical in the two different size samples, the remainder had been extremely inconsistent in value, resulting in very low correlations.

The results on guessing clearly show that samples greater than 250 examinees are required to obtain good estimates of guessing. It is not clear whether 1000 examinees are a sufficient number of examinees for estimating guessing.

Generally, the results of this section indicate that the Rasch model can be used effectively with samples of only 250. More examinees are required to obtain good item parameter estimates with the three-parameter model, but just how many examinees are needed isn't known.

Cost of Latent Trait Parameter Estimation

Test length, computer time (CPU seconds), and batch processing cost for estimating parameters of the Rasch and three-parameter models for 25 tests are shown in Table 22. It can be seen that the CPU time and cost for estimation were directly proportional to the number of items in a test. Doubling test length had the effect

Table 22

Computer Time (CPU Seconds) and Cost for Parameter Estimation by the Rasch and Three-Parameter Models for 25 Tests[1,2]

Test	Number of Items	1-Parameter Model		3-Parameter Model	
		Time	Cost	Time	Cost
Stan. Word Study	50	34"	$12.91	92"	$34.87
Stan. Reading	71	46	17.57	116	43.98
SAT Verbal #1	85	68	23.38	179	69.96
CTBS Math Comp.	48	35	13.66	81	31.56
CTBS Vocabulary	40	27	10.62	99	38.46
Stan. Vocabulary	50	34	13.00	101	38.64
Stan. Science	60	39	15.01	116	44.37
Stan. Language	80	49	18.63	155	58.25
Stan. Soc. Stud.	54	32	12.11	94	35.39
Stan. Listening	50	35	13.11	78	29.54
Stan. Spelling	60	41	15.69	98	36.91
Stan. Math Applic.	40	27	10.09	67	25.35
Stan. Math Con.	35	22	8.37	54	20.49
ICRT	16	9	3.85	32	12.43
SAT Verbal #2	85	60	22.91	134	50.51
Georgia Regents #1	70	45	17.21	118	44.61
Georgia Regents #2	70	46	17.30	116	43.79
Georgia Regents #3	70	47	17.99	116	43.92
CAT Vocabulary	40	26	9.93	84	31.67
ITBS Vocabulary	38	21	8.24	61	23.33
CAT Comprehension	45	26	9.96	98	37.00
ITBS Comprehension	68	38	14.46	139	52.43
Rasch-Environment[3]	34	12	4.69	29	11.19
Rasch-Citizenship[3]	30	11	4.42	18	6.89
Rasch-Career[3]	30	12	4.78	35	13.55

[1] Based on a CDC CYBER-175 computer.

[2] N=1000 examinees.

[3] N-500 examinees.

of approximately doubling the time and cost of an estimation. Although data are not shown, a similar relationship existed between the number of examinees and the time and cost of an estimation.

Cost results, across tests, are summarized in Table 23 which shows the minimum, maximum, and average CPU time and cost for estimation by each model. When abilities and item parameters were both simultaneously estimated, costs were three times more for the three-parameter model than for the Rasch model. When item parameters were known, and only abilities were estimated, the time and cost of estimation was identical for the two models. This result is shown in Table 24 which gives the average cost for parameter estimation by each model with known item parameter values. This result suggests that when items are drawn from item banks, regardless of the number of parameters (one, two, or three), the cost for estimating ability remains the same. The average cost for estimating the parameters of the Rasch and three-parameter models, when abilities are already known, is shown in Table 25. The cost of estimating item parameters with the three-parameter model was less than twice that of the Rasch model.

The costs reported in this study were significantly less than costs normally associated with latent trait parameter estimation. It should be emphasized that these costs were based on academic discounts and cannot be generalized to commercial computer establishments. Additional cautions regarding the interpretation of costs reported in this study were given in Chapter III and should be reviewed at this time to avoid confusion.

Table 23

Mean Computer Time (CPU Seconds) and Cost for Parameter Estimation by the Rasch and Three-Parameter Models[1]

	1-Parameter Model		3-Parameter Model	
	Time	Cost	Time	Cost
Minimum	9"	$ 3.85	18"	$ 6.89
Maximum	68	23.40	179	69.00
Average	33.68	12.80	92.4	35.12

[1] N=25 tests.

Table 24

Average Cost for Estimating Ability for 250 Examinees When Item Parameters are Known[1]

Model	Average Cost
1-Parameter Model	$3.24
3-Parameter Model	$3.28

[1] Based on N=5, 20-item tests.

Table 25

Average Cost for Estimating Item Parameters for 20 Items When Ability is Known[1]

Model	Average Cost
1-Parameter Model	\$3.27
3-Parameter Model	\$5.46

[1] Based on N=5 tests with 250 examinees.

Summary of the Results

Several comparisons between the Rasch and three-parameter models were presented in this chapter. The important findings are summarized as follows:

1. Standardized cognitive tests, constructed by conventional methods, were accurately described by the Rasch and three-parameter logistic latent trait models;
2. The Rasch model characterized data nearly as well as the three-parameter model;
3. Unidimensionality was an important consideration in latent model fit;
4. Lack of conformity to the Rasch model assumptions of equal item discrimination and no guessing had a small impact on Rasch model fit;
5. The presence of unequal item discriminations and guessing also affected three-parameter model fit;
6. Precise ability estimates were found with both the Rasch and three-parameter models on 20-item tests;
7. Precise estimates of item difficulty were obtained with samples of 250 examinees for both the Rasch and three-parameter models;
8. Estimates of item discrimination and guessing were not very good on samples of 250 examinees;
9. The cost of estimating item parameters and abilities simultaneously with three times more for the three-parameter model; and,
10. When item parameters were known, the cost of obtaining ability estimates was the same for the Rasch and three-parameter models.

CHAPTER V

CONCLUSIONS AND FUTURE DIRECTIONS

Review of the Design

The study explored the fit of the Rasch and three-parameter models to 25 empirical data sets. The degree to which data met the assumptions of the models was the primary variable investigated. Estimation precision based on test length and sample size was also examined. Item and ability parameters were estimated under the assumptions of the Rasch and three-parameter models. These were substituted for true parameters to make predictions about number-correct score distributions. Goodness of fit to observed score distributions was assessed with mean squares, Kolmogorov-Smirnov statistics, and graphic procedures. Correlation techniques were applied to evaluate model fit in relation to degree of unidimensionality, equality of item discrimination, amount of guessing, sample size, and test length. In addition, parameter estimation costs for the Rasch and three-parameter models were compared.

Results and Conclusions

The results of this study demonstrated that latent trait theory provided at least adequate models for describing high quality standardized tests in a number of subject areas. Aptitude and achievement tests, developed for norm and criterion-referenced measurement, displayed good fit to the Rasch and three-parameter models. It is concluded that latent trait theory is appropriate for analyzing standardized tests used to measure ability.

The Rasch model compared favorably with the three-parameter model in this study. On fifty percent of the tests, the Rasch model fit data as well as the three-parameter model. Overall, for each test, there was little difference between number-correct score distributions predicted by the Rasch and three-parameter models. The close results were due in part to the fact that the number-correct score is the sufficient statistic for the Rasch model, but not for the three-parameter model. Fit of data to the three-parameter model might have appeared better had a less biased criterion been used. Nevertheless, the results for the two models were so similar, that it is concluded that ability estimates from the Rasch model are nearly as acceptable as those from the three-parameter model.

The study illustrated the importance of the assumption of unidimensionality for latent trait model fit. As item sets tended to be more multidimensional, fit of data to the Rasch and three-parameter models was reduced. It is concluded that the latent trait models described in this study can only be applied to unidimensional tests. One approach to handling multidimensional data is to apply the models only to unidimensional subsets of the items. Factor analysis was suggested as a method for assessing dimensionality. The principal components solution used in this study offered a reasonable approximation to the more computationally bound principal axis method of factoring. Factor analytic solutions, suggested by Christofferson (1976) and Muthen (1978), are claimed to be more appropriate than conventional factoring techniques for dichotomous variables and might be applied when computer programs for the methods become available.

Rasch model fit was slightly impaired when data were characterized by heterogeneous item discrimination indices. A similar outcome was obtained with the three-parameter model. The results suggested that the presence of unequal item discrimination values may have been undesirable for both models. Because the results were not significant, it can be assumed that the Rasch model was fairly robust to departures from the assumption of equal item discrimination. The similar outcome for the three-parameter model may indicate that estimation of discrimination parameters is less accurate than desired, but the result can probably be better attributed to the potential unfairness of the number-correct score criterion for assessing three-parameter model fit.

The analysis of item-total score correlations for the purpose of assessing equality of item discrimination produced consistent results with those based on an analysis of estimated item discrimination indices. Lord and Novick (1968) gave an approximation to item discrimination from the biserial correlation. The item-total score correlation, or point-biserial, tends to fluctuate between samples because of varying difficulty levels and the presence of guessing. Consequently, the point-biserial cannot be translated into item discrimination. Despite this shortcoming, in this study, the values of point-biserials provided an adequate technique for evaluating homogeneity of item discrimination. The result was attributed to the high correspondence between biserial and point-biserial correlations in the data.

Both the Rasch and three-parameter models were affected by the presence of guessing. Both models fit data less well when examinees guessed. The Rasch model result was attributed to the lack of a guessing parameter in the model. In the three-parameter case, the result was attributed to underestimation of the chance level parameter. Estimates of chance level parameters in this study were not very acceptable,which may have been due to the modest numbers of low ability examinees represented in the samples.

Previous results (Hambleton & Traub, 1971) have shown that the Rasch model was quite sensitive to departures from the no guessing assumption. Since the results concerning guessing in this study were not significant, the previous conclusion could neither be supported or refuted. A method relying on item difficulty levels for estimating guessing failed to produce any meaningful results. Because the outcomes concerning guessing in this study were so confusing, it is thought that simulation techniques may be required for exploring the effects of guessing on model fit. Simulation studies can be designed to simultaneously vary guessing and heterogeneity of item discrimination values so that unique and mutual effects on model fit can be studied.

The results of the study demonstrated that good estimates of ability and item difficulty can be obtained with the Rasch model from tests with only 20 items and from samples with only 250 examinees. Although good ability estimates were obtained using the three-parameter model when tests had only 20 items, parameter estimates for guessing and item discrimination from samples with only 250 examinees were not adequate. The results of the study supported Lord's (1980) contention that 1000 examinees are required for obtaining good estimates of item discrimination. It would seem that larger samples may be needed to estimate the guessing parameter so that a moderate number of examinees are represented at the low end of the ability scale.

The study was not able to pinpoint the minimum number of examinees required in a sample for estimating item discrimination or guessing. More research, using a variety of sample sizes, is required in this area. It was also not determined whether tests shorter than 20 items can be used for estimating ability. Research is needed in this area as well.

It was shown that when ability was estimated from items with known values for item parameters, computer expenses for the Rasch and three-parameter models were about the same. When ability and item parameters were estimated simultaneously, the cost of estimation for the three-parameter model was three times more than for the Rasch model, but both costs were not high ($70.00 maximum for 40 items and 1000 examinees). In most practical work, item parameters are estimated at the onset, with subsequent estimations consisting only of abilities. Thus, in the long run, differences in parameter estimation costs for the two models would seem to be negligible.

Methodological Issues

Assessment of Model Fit

There have been many procedures developed for evaluating latent trait model fit. Since some of these methods are appropriate for only one of the latent trait models, the method used in this study was chosen because it could be applied to both the Rasch and three-parameter models.

The procedure used for testing model fit was based on steps outlined in Lord and Novick (1968). Parameters were estimated from the models and then substituted for true values to make predictions about some observable quality of the data, in this case, number-correct score distributions. The predicted distributions were compared using statistical and graphical methods to assess deviations from observed score distributions. The rationale for using number-correct score as a criterion was based on Lord's (1980, p. 51) specification of the relationship between fixed ability and the conditional distribution of number-correct scores. Using this relationship, the distribution of number-correct scores could be generated across ability levels. In conventional measurement, the number-correct score provides an estimate of true-score. True-score, which can be described as $\sum_{g=1}^{n} P_g(\theta)$ with latent trait parameters, provides a common basis for comparing latent trait models.

Hambleton and Traub (1973) used a significant modification of this method to compare fit of data to the Rasch and two-paramter models. They also used estimated values of parameters to make predictions about observed score distributions, but their score distributions were obtained by weighting number-correct scores by the optimal scoring weights derived from latent trait theory. The justification for using weighted number-correct scores was that the raw score is only a sufficient statistic for Rasch ability, but does not contain sufficient information to describe ability estimated with the two-parameter model. The weight which provided the sufficient statistic for ability for the two-parameter model was item discrimination. In assessing fit of data to the two-parameter model, Hambleton and Traub compared the predicted weighted score distribution to the observed weighted score distribution. Since the raw score was the sufficient statistic for the Rasch model ability estimates, all weights were one and the method reduced to the one used in this study in the Rasch case.

The method employed by Hambleton and Traub, although preferable in some ways to the one employed in this study, is lacking a theoretical basis since latent trait theory does not provide a relationship between distributions of weighted number-correct scores and ability. Nevertheless, this method would have been used in this study if there was a sufficient statistic for ability estimated with the three-parameter model.

Because of the complexity introduced into the three-parameter model by the guessing parameter, a sufficient statistic cannot be found. Birnbaum (1968) provided an optimal scoring weight for the three-parameter model which did not have the desirable properties of a sufficient statistic since it was not independent of level of ability. It is not possible to decompose the maximum likelihood function for the three-parameter model into two terms, one independent of ability. There is also not a sufficient statistic for the two-parameter model when the item response function is based on the normal ogive. For this reason, the logistic model is seen as a more desirable model.

Because predicted number-correct score distributions in this study were not adjusted by some optimal weight, the results are thought to have been biased in favor of Rasch model fit. Results for direct comparisons of the two models should be viewed with this caution in mind.

An Alternative for Assessing Model Fit

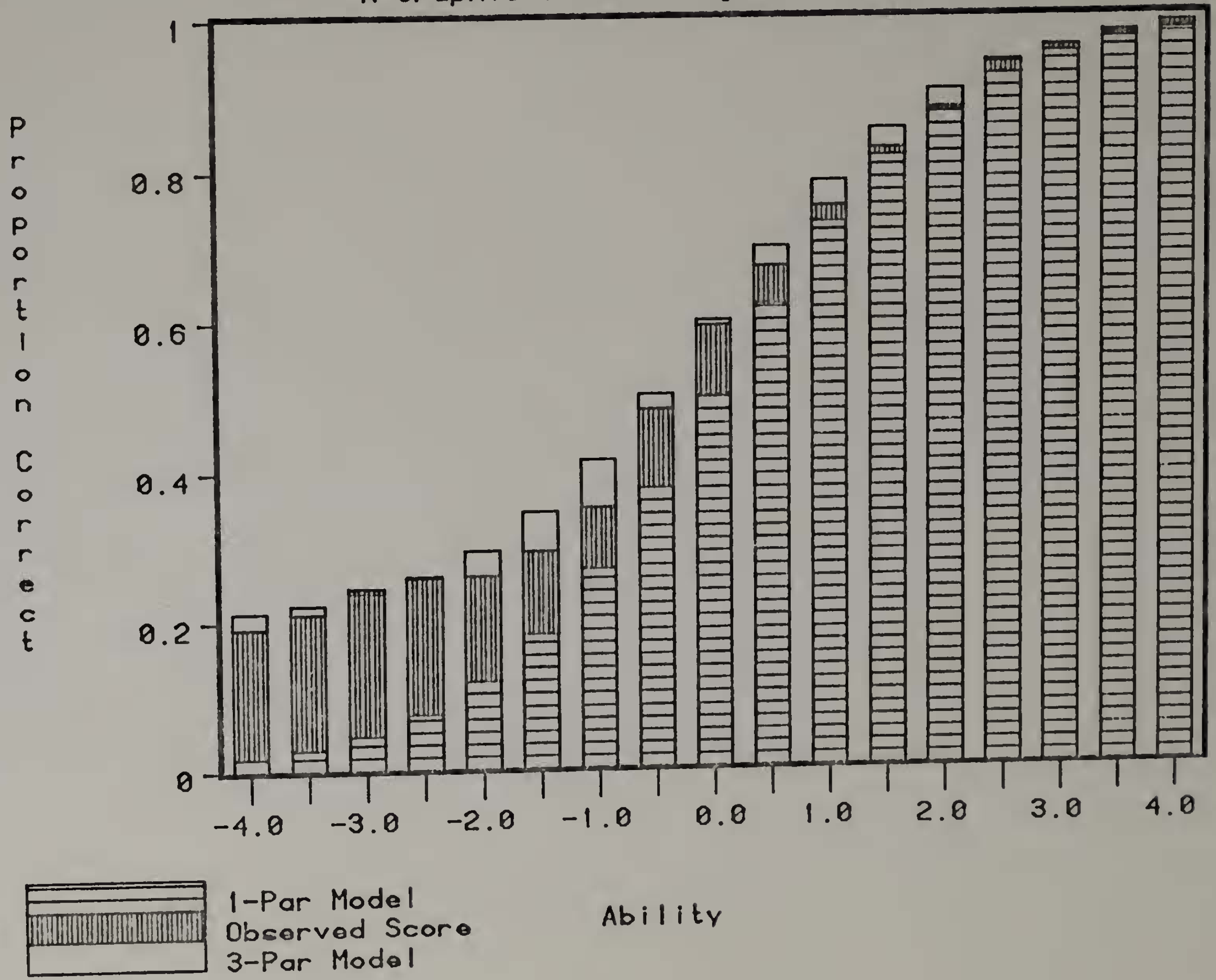
The mean square fit statistic, developed by Wright and Panchapakesan (1969), was designed for testing fit of data to the Rasch model, but could have been applied to the three-parameter model. This statistic has received criticism from George (1979) and others because of its frequent application to data in which sampling assumptions (large N) have not been fulfilled. In these cases the statistic is only approximately chi-square. Experience with chi-square values in this study demonstrated how sensitive the statistic is to sample size and to re-grouping of data into arbitrary class intervals. For these reasons approximate chi-square statistics are best avoided in testing model fit. The K-S statistic and graphic procedures applied in this study fortunately were not plagued by the same difficulties.

Although the mean square statistic has been inappropriate in many instances for computing chi-square values, the method suggests a graphic technique for comparing models that seems quite promising. The numerator of the mean square statistic is computed as the differential between the frequency of examinees who obtained a correct answer on an item and the expected frequency of examinees who got the item correct. The method, to be described here, is applied on an item basis, although there is some justification for using the method with test characteristic functions.

The ability distribution is divided into i class intervals in a manner similar to the one used in this study. For each ability level, i, the item characteristic function for item g, $P_{ig}(\theta_i)$, is found for each model using estimated item parameters. $P_{ig}(\theta_i)$ is the probability that examinees in ability group i will get item g correct and can be taken as an estimate of the proportion of examinees in ability group i who obtained correct responses to the item. This proportion can be compared graphically with the observed sample proportion, P_{ig}, of examinees in score group i who actually got item g correct. Figure 30 demonstrates the graphical method for a hypothetical item based on simulated data. Since the sum of the item probabilities provides the test characteristic function, $E(Z) = \sum_{g=1}^{n} P_{ig}(\theta_i)$, it can be argued that E(Z) is an estimate of the proportion correct score on the test. This could be compared to the observed proportion correct score for each ability group i.

Figure 30.

A Graphic Test of Logistic Model Fit



Evaluation of Departures from Model Assumptions

Latent trait theory is considered a superior method for measuring ability because of its sample-invariant properties. In this study, three departures from latent trait model assumptions had been evaluated with conventional methods. The conventional estimate of guessing failed to produce any meaningful data. The conventional estimate of equality of item discrimination was comparable to a method based on latent trait theory. Although unidimensionality measurement could not be directly compared to an assessment based on latent trait theory, the conventional estimate appeared to be accurate since multidimensional tests could not be fit by the latent trait models. A criticism of all conventional approaches used in this study to measure departures from latent trait model assumptions was that each was based on some sample-dependent quantity. Factor analysis, for example, only appraised whether a set of items had been unidimensional for a specific sample of examinees. Little could have been concluded about samples composed of different examinees. Sample-invariant measures of departures from latent trait model assumptions are needed. This is an area in which considerable research is warranted.

Assumption of Linearity in Correlation Methods

Inferences regarding the impact of departures from latent trait model assumptions on model fit were based on correlations between fit statistics and various measures of departure from model assumptions.

The Pearson product moment correlation is used to investigate linear relationships. The assumption of linearity may not have been warranted in this study. There was no reason to anticipate that mean square or K-S statistics would have been linearly related to contrived measures of departure from model assumptions. In future research in this area, techniques which do not assume linearity would seem to be more appropriate.

Methods for Estimation of Latent Trait Parameters

The LOGIST method for parameter estimation used in this study employed an unconditional maximum likelihood approach designed to estimate abilities and item parameters of the three-parameter model. By assuming that all item discriminations were equal and by fixing the lower asymptotes to zero, the method was applied to the Rasch model. The estimation costs reported in the study were based on the LOGIST method and consequently were more expensive for estimating Rasch model parameters than they would have been if another method, for example, BICAL (Wright & Mead, 1976), had been used. The reason for this is that LOGIST estimated "N" abilities, where N was the number of examinees, rather than the "n-1" ability estimates required by the Rasch model, where n was the number of items. Since raw score is the sufficient statistic for Rasch ability, "n-1" raw scores correspond to "n-1" ability estimates. LOGIST had been used to estimate parameters for both the three-parameter and Rasch models to avoid introduction of variation from other unanticipated factors. It would be desirable to compare the costs of estimation of Rasch model parameters between LOGIST and BICAL or some other Rasch model estimation routine.

Limitations of the Study

Drawbacks of Empirical Data

Because empirical data was used in the study, results could be generalized to the real world, but some information could not be obtained from empirical data. Because of the interaction of variables in real data, it was not possible to evaluate the unique importance of various factors on model fit. When data is generated in simulation studies, there is absolute control over dimensionality, equality of item discrimination, and guessing in a data set. With real data, when a test did not fit one of the latent trait models, it was impossible to conclude whether misfit was due to a single factor or to a mutual contribution of factors. Too few tests had been available for partial correlation analyses to have been useful. Some of the information sought in the study would have been more easily gathered from simulation techniques. It is important to emphasize that results based on simulated data can not substitute for those based on real data.

Restricted Number of Models Studied

Interest in latent trait models has primarily focused on the Rasch and three-parameter models. The two-parameter model has also been a topic of interest. Because of the substantial number of analyses dictated by the study design, time constraints, and limited resources prohibited exploration of the two-parameter model. The nominal response model and the graded response model are interesting generalizations of the logistic models. These models require special estimation techniques and that data be scored in special ways. Neither computer programs or appropriately scored data were available for this study. Future research on the fit of latent trait models to empirical data should include some of these interesting modifications to the general logistic model.

Data Limitations

The number of tests analyzed in this study was twice the number evaluated in previous comparative research on latent trait models. The complexities of data management in the study prohibited inclusion of more tests, although this may have been desirable. Data were limited in a number of ways. Perfect data can only be obtained through simulation. Real data is naturally imperfect. Data were scored differently; some data were from timed test administrations; others were not. Tests differed in length, content area, and composition of examinee samples. Texts for question sets were not available, so interpretation was limited to statistical qualities. While lack of knowledge about data limited interpretation, it also assured non-biased treatment of data.

Future Directions

A number of topics for future research have been mentioned in this chapter. The study concludes with a review of areas which need additional clarification:

1. Better tests for departures from latent trait model assumptions. This study applied sample-dependent item statistics to evaluate departures from model assumptions. Sample-invariant methods are needed to replace these techniques. Bejar (1980) suggested a method for assessing dimensionality. New approaches are needed for assessing equality of item discrimination and guessing.

2. Additional criteria for model fit. The number-correct score had too many shortcomings as a criterion for model fit. A graphic method which used predicted item probabilities was suggested as one alternative. Little research has examined person fit to the latent trait models. Computer-based graphing programs provide a means for exploring thousands of items or persons. More research is needed to determine causes of misfit of persons as well as items and to assess the impact of poorly fitting persons/items on overall model fit.

3. Need for simulation studies. Confounding of variables in empirical data prohibited evaluation of unique contributions of various factors on model fit. Simulation designs, which offer experimenter control, are needed to explore the unique and mutual effects of guessing and heterogeneity of item discriminations on model fit.

4. Minimum sample size and test length needed for latent trait parameter estimation. The study explored estimation precision for 20-item tests and for examinee samples of 250. Additional research is needed to determine if latent trait parameters can be estimated on shorter tests and even smaller samples.

5. Fit of additional latent trait models. This study focused on fit of data to the one- and three-parameter logistic latent trait models. More empirical research is needed to evaluate the appropriateness of the two-parameter, nominal, and graded response models for real data.

6. Comparison to other parameter estimation methods. The costs reported in this study could not be generalized to other computing establishments. Benchmark data is needed for LOGIST on other academic and commercial computing facilities. In addition, data is needed which compares the costs, and other aspects of LOGIST, to other estimation programs.

7. Other ways of evaluating the impact of departure from model assumptions on model fit. Fit statistics in this study were correlated with measures of departure from latent trait model assumptions. Since these relationships cannot be assumed to be linear, other methods are needed which do not make the assumption of linearity. A decision theoretic approach might be used in this area. This would be based on a set of classification rules for fit and conformity to assumptions.

REFERENCES

Anderson, E. B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, *38*, 123-140.

Anderson, E. B., & Masden, M. Estimating the parameters of the latent population distribution. *Psychometrika*, 1977, *42*, 357-374.

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 231-238.

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

Bejar, I. I. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement*, 1980, *17*, 283-296.

Bejar, I. I. Assessing the unidimensionality of achievement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement. Research Report 77-5, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1977.

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 tests*. Reading, MA: Addison-Wesley, 1968.

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

Bock, D. R., & Lieberman, M. Fitting a response model for n dichotomously scored items. *Psychometrika*, 1970, *35*, 179-197.

Christofferson, A. Factor analysis of dichotomized variables. *Psychometrika*, 1975, *40*, 5-32.

Connally, A. J., Nachtman, W., & Pritchett, E. M. *Key Math: Diagnostic Arithmetic Test*. Circle Pines, MN: American Guidance Service, 1971.

Cook. L. L., Eignor, D. R., & Hutten, L. R. Considerations in the application of latent trait theory to objectives-based criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Cowell, W. R. ICC preequating in the TOEFL testing program. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Dinero, T. E., & Haertel, E. Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement*, 1977, *1*, 581-592.

Divgi, D. R. A nonparametric test for comparing goodness of fit in latent trait theory. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Donlon, T. F. An exploratory study of the implications of test speededness. Princeton, NJ: Educational Testing Service, 1978. (mimeographed)

Douglas, J. B. A comparison of item characteristic curve models for a classroom examination system. Unpublished doctoral dissertation, Michigan State University, 1980.

George, A. A. Theoretical and practical consequences of the use of standardized residuals as Rasch model fit statistics. Paper presented at the annual meeting of the American Educational Association, San Francisco, 1979.

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

Gustaffson, J. Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 1980, *33*, 205-233.

Gustaffson, J. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement*, 1979, *16*, 153-158.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R. K. Latent ability scales: Interpretations and uses. In S. Mayo (Ed.), *Interpreting test performance: New directions for testing and measurement* (No. 6). San Francisco: Jossey-Bass, 1980.

Hambleton, R. K. Latent trait models and their applications. In R. Traub (Ed.), *Methodological developments: New directions for testing and measurement* (No. 4). San Francisco: Jossey-Bass, 1979.

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

Hambleton, R. K., & Cook, L. L. Some results on the robustness of latent trait models. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

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

Hambleton, R. K., & Traub, R. E. An analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 195-211.

Hambleton, R. K., & Traub, R. E. Some empirical results on the robustness of the Rasch test theory model. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. *Journal of Educational Measurement*, 1978, *15*, 43-47.

Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 1965, *30*, 179-185.

Jensema, C. J. An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 29-48.

Koch, B. R., & Reckase, M. D. A live tailored testing comparison of the one- and three-parameter logistic models. Paper presented at the National Council on Measurement in Education annual meeting, Toronto, 1978.

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

Lord, F. M. A theory of test scores. *Psychometric Monograph No. 7*, Psychometric Society, 1952.

Lord, F. M. Estimating item characteristic curves without knowledge of their mathematical forms—a confrontation with Birnbaum's logistic model. *Psychometrika*, 1970, *35*, 43-50.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 1974, *39*, 247-265.

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117-138.

Lord, F. M., & Novick, M. R. *Statistical theories of mental tests.* Reading, MA: Addison-Wesley, 1968.

Lumsden, J. Person reliability. *Applied Psychological Measurement*, 1977, *1*, 477-482.

Lumsden, J. The construction of unidimensional tests. *Psychological Bulletin*, 1961, *58*, 122-131.

Marco, G. L., Peterson, N. S., & Stewart, E. E. An evaluation of linear and equipercentile equating methods. Paper presented at the ETS Research Conference on Test Equating, Princeton, NJ, 1980.

Martin-Löf, P. Statistiska modellar. Anteckningar frán Seminarier läsáret 1969-70 utarbetade av Rolf Sundberg.2: a uppl. Institute för Försäkringmatematik och matematisic Statistik vid Stockholms Universitet, 1973.

Mead, R. Assessing the fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Muthén, B. Contributions to factor analysis of dichotomous variables. *Psychometrika*, 1978, *43*, 551-560.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *Statistical Package for the Social Sciences.* NY: McGraw Hill, 1975.

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Pine, S. M. Applications of item response theory to test bias. Unpublished manuscript, Department of Psychology, University of Minnesota, Minneapolis, 1976.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Paedagoishe Institute, 1960.

Reckase, M. D. A comparison of the one- and three-parameter logistic models for item calibration. Paper presented at the annual meeting of the American Educational Research Institute, Toronto, 1978.

Ree, M. J., & Jensen, H. E. Item characteristic curve parameters: Effects of sample size on vertical equating. Unpublished manuscript, Personnel Research Division, Brooks Air Force Base, Texas, 1980.

Rentz, R. R., & Bashaw, W. L. Equating reading tests with the Rasch model: Final report. Athens, GA: Educational Research Laboratory, College of Education, University of Georgia, 1975.

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work: A synthesis of the literature for practitioners. Mimeographed, 1978.

Ross, J. An empirical study of a logistic mental test model. *Psychometrika*, 1966, *31*, 325-340.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional space. *Psychometrika*, 1974, *39*, 111-121.

Slinde, J. A., & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, 1978, *15*, 23-35.

Slinde, J. A., & Linn, R. L. The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement,* 1979, *3*, 437-452.

Swaminathan, H., & Gifford, J. Estimation of parameters in the three-parameter latent trait model. Presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, *34*, 253-269.

Waller, M. I. An objective procedure for comparing the one-, two-, and three-parameter logistic latent trait models. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Waller, M. I. Estimating Guessing Tendency. Research Bulletin 74-33, Princeton: Educational Testing Service, 1974. (a)

Waller, M. I. Estimating parameters in the Rasch model: Removing the effects of random guessing. Research Bulletin 76-8, Princeton: Educational Testing Service, 1976.

Waller, M. I. Removing the effects of random guessing from latent trait ability estimates. Research Bulletin 74-32, Princeton: Educational Testing Service, 1974. (b)

Whitely, S. E. Models, meanings, and misunderstandings: Some issues in applying Rasch's theory. *Journal of Educational Measurement*, 1977, *14*, 227-235.

Whitely, S. E., & Davis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, 1974, *11*, 163-178.

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

Woodcock, R. *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service, 1974.

Wright, B. D. Misunderstanding the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 219-225.

Wright, B. D., & Mead, R. J. BICAL: Calibrating rating scales with the Rasch model. Research Memorandum No. 23. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-48.

Wright, B. D., & Stone, M. H. *Best test design*. Chicago: MESA Press, 1979.