THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# A METHOD OF DISCRIMINATING PARTIAL KNOWLEDGE

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

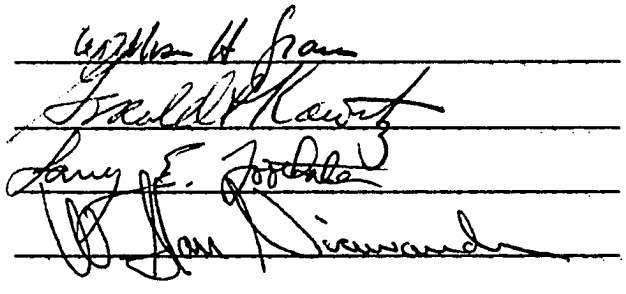
ROBERT HENRY BLACK

Norman, Oklahoma

1974

A METHOD OF DISCRIMINATING PARTIAL KNOWLEDGE

APPROVED BY



DISSERTATION COMMITTEE

## ACKNOWLEDGEMENTS

This writer would like to express his gratitude and appreciation to the following individuals.

Dr. William Graves for his unselfish and flexible support of opportunities for the development of my teaching skills and research interests. Dr. Allen Nicewander who, in addition to introducing me to test theory, with patience has been instrumental in the development of this dissertation, has provided very insightful critism, and helped me gain enthusiasm for an analytical approach to problems. Dr. Larry Toothaker who taught me the basics of computer simulation and provided insights as to its value and application. Dr. Gerald Kowitz who humbled me when I needed it, encouraged me when I needed it, and was always interested in my personal and professional welfare. Rebecca Gragg for her steady assistance in typing, correcting and compiling this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

## LIST OF FIGURES

# A METHOD OF DISCRIMINATING PARTIAL KNOWLEDGE

## CHAPTER I

## INTRODUCTION

### Testing in the Social Sciences

In a general sense, tests in the social sciences are used to measure the nature and extent of differences among individuals. Thus a test is defined as a systematic procedure for measuring a sample of an individual's behavior. In a strict sense, the response an examinee makes to a test item is the only behavior a test measures. Even this behavior is only a sample of possible behaviors within a given domain.

The necessity for sampling gives rise to two questions. First, would the examinee obtain the same score if he were to respond to a different sample of items from the same behavior domain? This question concerns the reliability of a test. Second, are the items chosen for inclusion in a test a representative sample of the universe of possible behaviors in the area of interest? This is the question of validity.

Test constructors and test users find themselves in a special situation. Because their tests are never perfectly valid or reliable, test scores contain rather sizable errors of measurement. In addition, the characteristics or differences

among individuals which are of greatest interest for study are usually not directly measurable but rather must be studied indirectly, through the measurement of other quantities e.g., the responses of examinees to test items.

The investigator comes to grips with these problems by constructing theories of mental testing and formulating models that provide a framework which permits logical deductions concerning general and specific relationships which have yet to be empirically demonstrated.

These models allow the investigator to make measurements because they provide procedures for the assignment of numbers to specific characteristics of the experimental units in a way that preserves the specific relationships in the behavioral domain of interest. Thus, test scores become indicants from which an investigator may make inferences about the characteristics of an unobservable variable.

In psychological testing, these characteristics are often referred to as traits. A trait is a hypothetical construct referring, in an operational sense, to a cluster of empirically interrelated behaviors. The trait name (e.g., intelligence, self concept) is a descriptive label applied to the group of behaviors.

Through the years psychology as a science has become organized and unified by the development of theories which have served to describe, explain, and predict some aspects of individual differences. In the course of this development, mental tests have distinguished themselves in the areas of vocational placement, diagnosis, hypo-

thesis testing and hypothesis building in research settings, and in many areas of evaluation.

Although nothing in the definition of a test requires that one specific format be used, much of psychology uses the responses of an examinee to questions on paper and pencil tests as an inferential numerical index of the strength of a psychological trait. The test item, then, represents the experimental stimulus the psychologist deems sufficient to elicit behavior characteristic of a specific latent psychological trait.

The test constructor generally wants to determine as reliably as possible the rank order of a group of examinees on a given psychological trait as measured by a set of stimulus test items. If the test constructor is dissatisfied with the test reliability or validity, or both, several alternatives for improving these characteristics present themselves. Among other strategies, he may replace or revise some of the test items, he may improve the criterion measure, he may lengthen the test, or he may score the test in a manner which may yield more information from the test items. It is with scoring formulas that investigations of partial knowledge have been concerned.

## Multiple-Choice Test Item (conventional scoring)

A multiple-choice item scored in the conventional manner asks the examinee to choose the correct alternative for one point credit and gives no credit when an incorrect alternative (distractor) is chosen.

Several authors (Garvin, 1972; Hambleton, 1970; Rippey, 1971) indicate that an examinee's ability to choose the correct alternative to a given item is not particularly informative about the state of knowledge of the examinee with respect to the item. No matter how or why they were selected, all correct answers look alike. A single, unqualified choice does not separate the confident examinee from the timid one. Nor does it distinguish between the lucky guesser and the expert. It is not difficult to imagine situations in which the selection of alternatives based on grossly disparate levels of relevant knowledge receive the same credit.

Hambleton (1970) suggests further that the multiple-choice testing format poses a problem when an incorrect alternative or an omit is given because nothing of great value is learned about the examinee except that he has failed to identify the correct alternative.

Dressel and Schmid (1953) put forth the argument:

> there are meaningful distinctions in the ability of students which are not disclosed by the selection or non-selection of the keyed response to the usual multiple-choice item. It is apparent that these distinctions are particularly significant in the case wherein the responses themselves help to set the situation to which the student must respond.
>
> There is a tendency to assume that such a difference in the student certainty about the correctness of his response will be accounted for over the entire test. To put it the opposite way, the student whose response contains an element of guessing will tend to miss enough items over the entire test to differentiate him from the student who responds with complete certainty. This hypothesis needs more careful investigation rather than ready acceptance. Particularly this is true if assurance about what one knows and does not know is a desired educational outcome. (p 576)

So, as Coombs, Milholland, and Womer (1956) have suggested, although the multiple-choice testing format enjoys a great deal of popularity, its merits are not necessarily optimal psychometrically. When multiple-choice items are scored in the conventional manner, Hambleton et al. (1970), Coombs et al. (1956), and others have pointed out several disadvantages:

First, the accuracy of estimating the degree to which an examinee is in possession of a psychological trait is reduced because of the inability to discriminate between partial and complete knowledge. Second, is the encouragement of guessing, which is only compensated for, not penalized by, the conventional right-minus-wrong correction formulas (Hamilton, 1950). Third, guessing operates to truncate scores at the lowest ability levels while dichotomous scoring operates to truncate scores systematically at the highest ability levels. The result is a reduction in the range of scores and the introduction of a chance variable. Both of these effects combine to reduce the reliability of the test and the test item (Frary, 1969a, 1969b; Garvin, 1972; Grier and Ditrichs, 1968).

## Multiple-Choice Test Item (partial knowledge)

The concept of partial knowledge has grown out of the belief that multiple-choice tests have been used inefficiently because the only score obtained is the number right score.

As Powell (1968) expressed it:

> Much time is spent by the examiner in the preparation of foils for multiple-choice tests. A proportionally large time is spent by the examinee

> in making his selection decision among the alternatives. In spite of the time thus spent, the foils are generally treated as a mask to the right answer and are lumped together in a general wrong category. The rating of the examinee is usually entirely dependent on his total number of correct items on any given test or subtest. On the other hand, if a multiple-choice test has been well prepared, particular wrong answers may have nearly equivalent discriminating power as do the right answers. (p 403)

The concern here is placed on the scoring formula and the ability to extract more information from each test item rather than with the multiple-choice item itself.

Nedelsky (1954) pointed out that examiners using conventional scoring method were making the assumption that with respect to the ability tested by given questions all students who choose any one of several wrong alternatives form a fairly homogeneous group. He noted further that this assumption is demonstrably false for most tests because neither the degree nor the kind of falseness is the same for all wrong alternatives. Nedelsky (1954) presented the results of a study of examinee scores based on the frequency with which they chose a particular kind of incorrect alternative. The conclusion was that although the poor examinees exhibited no reliably measurable differences in their ability to select correct alternatives, they did show considerable differences in their ability to reject grossly incorrect alternatives.

From another point of view, we might argue that while an examinee may not know the correct alternative to an item, he may know some of the things which are incorrect. The idea of correct discrimination among distractors in multiple-choice tests was used by Coombs,

Milholland, and Womer (1956) to conceptualize partial knowledge. In formulating a basis upon which to test for evidence of partial knowledge, Coombs et at. (1956) considered the conventional scoring formula for correcting for guessing. This formula assumes that an examinee either knows the correct alternative or guesses randomly. If there were no partial knowledge and there were a way of telling, on those items an examinee missed, what his second choice for the correct alternative would be, he would be expected to get 1/(K - 1) of them correct by chance, where K is the number of alternatives. However, if partial knowledge exists there would be a disproportionate number of the examinees getting more than 1/(K - 1) of these items correct on their second choice. This line of reasoning could be extended to an examinee's third, fourth, and fifth choices. Coombs et al. (1956) devised an investigation to test this hypothesis. Their

> results indicated that examinees with less than complete information on a given subject may have considerable partial information and that this may be used as a valid basis for discriminating among them. (p. 22)

Davis and Fifer (1956) carried out a study designed to find out whether the source of variance associated with distractors was of any practical value. Their method was to compare the gain in reliability and validity of an experimental scoring formula over the conventional scoring formula. They concluded that

> the increase in reliability arises from the differential weighting of responses to incorrect choices in items. Variance arising from selection by examinees among distractors of unequal merit is obtained; this variance is excluded

> from measurement when all incorrect choices are weighted equally. (p 165)

Other investigators (Hambleton et al. 1970; Jacobs, 1962; Jacobs and Vandeventer, 1970; Sigel, 1963) have approached partial knowledge from the point of view that the choice of a distractor reflects a non-chance influence of some importance. The results of these studies indicate that good multiple-choice test items stimulate a rather involved and extended thought process on the part of the examinee. Although each of these studies have made attempts to recover this information, Shuford, Albert, and Massengill (1966) argue:

> ...upon reflection it is quite apparent that all techniques in current use for assessing the present state of a student's knowledge fail to extract all of the potentially available information. In the case of objective testing.....the response methods upon which they are based extract only a very small fraction of the information (partial knowledge) potentially available from each query. (p 126)

## The Problem

Methods devised to incorporate this basic idea of differential examinee knowledge into mathematical models which make theoretical and practical sense in the context of test theory have taken several forms. These forms fall into the basic category of differential weighting of item alternatives.

There have been many investigations of partial knowledge over the past 50 years. Although the standards for evaluating the different models have not been consistent there seem to be two conclusions which can be reached. First, there is ample evidence (intuitive,

analytical, and experimental) that partial knowledge exists in an amount worth recovering. Second, given that the quality of item writing is high, formula scoring methods provide a valid tool for recovering partial knowledge.

## Latent Ability Test Model

In 1952, Lord presented a latent ability test model, adapted from the works of Lawley (1943) and Lazarsfeld (1950), for use with binary scored aptitude and achievement tests. This model specifies a function which relates the probability of success on an item to the underlying latent traits or abilities which the test measures. When a single latent trait is assumed to underlie test performance, the function is termed an item characteristic curve.

The item characteristic curve approach specifies the interrelation of underlying examinee ability, item discrimination, and item difficulty in a way that provides a logical framework for describing precisely how an item functions. To date there have been no studies of partial knowledge using the mathematical model proposed by Lord (1952).

The purpose of this study will be the construction and evaluation of the properties of a partial knowledge extension of Lord's (1952) basic latent trait model. A three parameter binary scoring formula will be contrasted with a three parameter rank order scoring formula (the third parameter being a guessing parameter) in terms of item reliability and validity varied across levels of item difficulty and discrimination.

## CHAPTER II

## REVIEW OF RELATED LITERATURE

The review of the literature is organized to point out the major developments in the area of partial knowledge investigation. There have been three main directions of study to date. First, differential weighting of item alternatives; second, confidence testing; and third, probabilistic scoring. Each of these categories have analytical, experimental, and intuitive arguments supporting them.

### Differential Weighting of Item Responses

There are two general methods of weighting item options in tests. One involves weights chosen empirically to maximize the relationship of the testing instrument to some internal or external criteria (Stanley and Wang, 1968). The other involves the use of *a priori* weights.

Keying option weights to some internal or external criterion stems from the work of Strong (1943) in the area of interest and personality inventories. Strong weighted the options of his interest items so as to maximally differentiate among various occupational groupings of people. Strong used the percentage of response to each option as a basis for keying each option to each group of people. Kuder (1957) also utilized this approach.

Both Strong and Kuder found positive empirical evidence to support the value of differential option weighting in interest and personality inventories.

Staffelbach (1930) obtained regression coefficients for three scores on a 60 item true-false test. The three scores were number correct, number incorrect, and number omitted. Since the test was made up of true-false items, the weighting was for incorrect responses as opposed to omitted responses.

Kelly (1934) developed a weighting procedure for use with dichotomous variables. His procedure took into account the item-criterion correlation.

One of the earliest investigations of the effects of differential option weighting on test reliability and predictive validity was done by Guilford, Lovell, and Williams (1942). They used the first 100 items of a 308 item general psychology test as those for which response weights were to be chosen. From 300 answer sheets 2 samples of 100 were chosen. The first was from those making the highest scores, the second from those making the lowest scores. Percentages of response for each item were then calculated and used as response weights.

An additional sample of 100 was drawn from the original 300 students who took the test. Each of these 100 answer sheets were scored using the conventional and weighted procedures. Scores on the odd and even items were used to calculate the reliability coefficients.

A very serious limitation involves the fact that the 100 test papers used to calculate reliability for the weighted scores were sampled from the same sample on which the weights were initially established. This may have produced spuriously high reliability coefficients.

A study by Dressel and Schmid (1953) was among the earliest to attempt to increase the discriminating power of multiple-choice items by varying the formula scoring procedures. There were five groups of examinees, each taking a 44-item test under a different set of instructions. The first group was scored by the number of correct responses. The second group was asked to indicate the certainty of their responses on a 4-point scale. The third group was to mark all alternatives they thought correct. Group four had a test modified so that more than one correct response was possible. The fifth group took a test having exactly two correct answers per item. Dressel and Schmid did not report any significant gains in reliability among the five methods.

Coombs, Milholland, and Womer (1956) devised a study in which the task presented to examinees was that of selecting and marking the distractors rather than the answer to multiple-choice questions. One point credit is gained for each distractor correctly identified and three points credit lost if the answer is incorrectly marked as a distractor. Coombs, et al. (1956) postulated that this seven-point item score scale would produce greater item and test variance than the conventional two-point item score scale. They also suggested that their experimental method would penalize random guessing associated with partial knowledge. To test these hypotheses they administered a 40-item, 4-choice multiple choice test. Increases in reliability were noted in terms of Kuder Richardson 20 formula (KR - 20).

The specific examinee response to difficult and easy items provided evidence that the reliability of a test composed of difficult items is more likely to be increased by the use of response weights

than the reliability of a test made up of easy items. This result has also been expressed by Lord (1963).

Nedelsky (1954) presented a study of examinee scores based on the frequency with which they chose a particular kind of wrong response, specifically a response which, if mistaken for a right response, showed gross ignorance on the part of the examinee. In Nedelsky's system, instructors classified the distractors to each multiple-choice item of the test as:

R response or right answer

F response or responses which are so obviously wrong that they would have little appeal except to the poorest examinees.

W responses other than F or R responses

A composite C-score was proposed.

$$C = R - F/f$$

where: f is the average number of F responses per item in the test.

Nedelsky's data were obtained from the administration of a 113 item physical science test to 306 examinees. Nedelsky then computed KR-20 reliability coefficients for R, F, and C scores for examinees who were graded A, B, C, D, F on the test. The R score was found to have negative reliability for D and F graded examinees. The F-score reliability was highest for this group of examinees.

The C-score was considered to be the most reliable of the three scores, possibly because only 70 of the 113 items contained F-responses. However, it was noted that the F-score furnishes evidence that, although the poorer students exhibit no reliably measurable differences

in their ability to select correct answers, they did show considerable differences in their ability to reject grossly wrong answers.

Merwin (1959) studied six methods of scoring three-choice multiple-choice items while varying the item parameters. He used correct answer only, a set of integer weights, and weights based upon the mean criterion score for examinees choosing a particular response pattern. Merwin concluded that scoring methods used in connection with the latter weights will yield an item validity as high as any other method. He also noted that the gains in item reliability and validity were relatively small and would be even smaller after cross-validation.

Davis and Fifer (1959) investigated the effects of item option weighting of multiple-choice items on the reliability and validity of a high school arithmetic reasoning test. From a pool of 300 items, two parallel forms were constructed, each containing 45 items. Two mathematicians, working independently, assigned weights to each alternative in the two tests. These weights were on a seven-point scale ranging from -3 to +3. These a priori weights were then used for all choices in the two tests. A sample of 370 examinees were scored, using the weights and the conventional right-only method. Parallel-forms reliability was computed and a gain from .68 (the conventional method) to .76 (the weighted response method) was noted. This increase in test reliability was equivalent to that obtained by lengthening the test one and one-half times. Davis and Fifer did not, however, find a significant increase in test validity using the option weighting. They did conclude that a significant increase in test reliability can be gained without reducing the validity, altering test length, testing

time, or scoring time if the option weighting is used on a well-constructed test.

Sabers and White (1969) reported an empirical study of the scoring procedure used by Davis and Fifer (1959). Methodologically speaking, their study was weaker than that of Davis and Fifer, and therefore were unable to replicate the findings. Sabers and White endeavored to increase validity but obtained an improvement of not more than .03. This small improvement was due in part to the mismatching of cross-validation groups.

Hambleton, Roberts, and Traub (1970) made a comparison of the reliability and validity of two methods of assessing partial knowledge on multiple-choice tests. They administered the midterm exam in an educational measurement course under three different procedures. The first was the conventional right-only method, the second was a method using differential weighting of responses, and the third was a confidence-testing format. To arrive at differential response weights, 22 experts rank ordered for correctness the five responses for each of the 40 multiple-choice items in the midterm exam. These rankings were scaled using a technique devised by Brock (1960). This technique assigns values to ranks so as to discriminate optimally among the objects being scaled. The confidence testing was scored using the procedure suggested by Shuford and Massengill (1967). Hambleton, et al. estimated the reliability from the odd-even split halves and validity from the correlation between scores on the midterm exam and scores on the final exam. Coefficients of effective length of .692 and .711 were noted for the reliability increase and 4.1 and 2.05 for

validity. These seem to be rather substantial increases. They should be noted with great caution for several reasons. First, the sample representativeness and size are in serious question. Second, the testing time was unequal in each of the three procedures. Third, the test employed in the study was easy for the group being tested. In situations like this it is doubtful that partial knowledge is being tested.

Bayuk (1973) conducted an investigation to determine the effects of response-alternative weighting and item weighting on reliability and predictive validity. Weights were assigned which were proportional to the mean criterion score of examinees selecting that alternative. Weights were derived for each alternative including omit and not read. Item weights were computed by maximizing the relationship between the composite of item scores and a criterion using multiple regression. Results indicated that scores resulting from response-alternative weighting were significantly more reliable than scores corrected for chance success. Scores significantly less reliable than scores corrected for chance were obtained when item weighting and response weighting were used together. There were no gains in predictive validity reported.

## Confidence Rating

Multiple-choice items scored in the conventional manner seem to imply that knowledge is a dichotomous or trichotomous entity. The majority of the advocates of confidence testing view knowledge as a continuous variable in the sense that there are varying degrees of it. Some

authors contend that confidence testing discourages guessing since the score systems for some methods are derived in such a manner that an examinee will maximize his expected score only if he reveals his true degree of certainty in responding.

Much of the subject of confidence testing is concerned with the manner in which the examinee is asked to respond to the items and the scoring formula that is used for each item.

In general terms, the examinee is asked to indicate not only what he believes to be the correct response to an item, but also how certain he is of his response. When his response is scored, the examinee receives more credit for a correct response given confidently than he receives for one given diffidently. But the penalty for an incorrect response given confidently is heavy enough to discourage unwarranted pretense of confidence (Ebel, 1965).

Hevner (1932) reported one of the first uses of confidence testing for minimizing the effect of guessing in true and false testing. She set out to study the degree of improvement in reliability between the conventional and confidence testing formula scoring systems on tests of music appreciation. Subjects in her study were to choose the more musical of two pieces and then indicate their degree of confidence in their choice on a three point scale.

Hevner compared the reliability of four different scoring formulas. The first was the number of correct responses; the second was the number correct minus an incorrect score using the weights mentioned in the weighted correct procedure. The weighted correct score

showed the most improvement in reliability.

Of the three methods compared to the conventional scoring formula, the weighted correct showed the greatest gain in reliability. Since there was no penalty for misplaced confidence, Hevner found it necessary to keep the scoring formula a secret so that the dishonest subjects could not raise their score artificially.

Soderquist (1936) reported a study similar to that of Hevner. His scoring formula used a weighted-correct minus a weighted-incorrect score; the weights for the incorrect responses were double the amount of credit claimed by the student on the item. The weighted-correct minus the weighted-incorrect score was compared with the conventional right minus wrong score and reliabilities were computed on random split-halves. Soderquist found substantial gains in reliability using the scoring formula weighted for student confidence. Soderquist found coefficients of effective length of 2.2 using the scoring formula weighted for student confidence.

Several authors reviewed the studies by Hevner and Soderquist and postulated the existence of personality traits which might influence confidence testing procedures. Wiley and Trimble (1936) performed a study which seemed to confirm this. Although they concluded that personality factors were present and that confidence testing could be used to study personality, they did not indicate specifically which personality variables were operating in their study. In an attempt to isolate personality factors more specifically, Swineford (1938) administered several true-false tests using Soderquist's confidence testing method. Swineford

identified what was termed a gambling score. She concluded that even though a coefficient of effective length of 1.42 was obtained using Soderquist's system, confidence testing confounds the measurement of achievement with an irrelevant personality trait, the willingness to gamble in a competitive academic situation. In 1941, Swineford replicated the earlier study using other tests and further concluded that boys tended to gamble to a significantly greater extent than did girls, both sexes gambled more on unfamiliar material, and that gambling scores were independent of achievement test scores.

Jacobs (1968) repeated Swineford's study and found the same results. In 1971 Jacobs formally questioned the use of confidence testing on the grounds that the scoring procedure is contaminated to a very large extent by individual differences in examinee personality. Two students of equal true ability but indicating different degrees of confidence would look like students of differing ability under most confidence testing procedures.

In an effort to improve the discrimination of multiple-choice items without increasing testing time, Dressel and Schmid (1953) experimented with four modifications of the conventional multiple-choice item.

They termed the first modification a free choice test. Under this test condition the examinees could choose as many alternatives as they thought correct. The second modification was termed the degree-of-certainty test. Under this testing condition the examinees were to indicate on a four point scale how certain they were with respect to a single response. The other two modifications are de-

scribed under response weighting procedures. Unlike earlier studies of this nature, the examinees in each testing condition were made aware of the scoring formula being used. Under the free choice testing condition, superior students were found to mark fewer alternatives across test items than did average and poor students. The degree-of-certainty testing condition was found to differentiate among superior, average, and poor students quite well. There was an improvement in reliability using the degree-of-certainty method as indicated by a coefficient of effective length of 1.16.

Ebel (1965) described what is basically a modification of Soderquist's scoring formula, and adapted it for use with true-false test items. Like the early experimenters in confidence testing, Ebel's intent was to reduce the error component due to guessing in test scores. Ebel's formula scoring system combines the basic features of confidence testing and both forms (additive and subtractive) of the correction for guessing.

Ebel (1965) reported reliability data from three different classroom tests using the Kuder-Richardson 20 formula. He found the confidence testing formula scoring procedure to yield coefficients of effective length of 1.84, 1.48, and 1.72.

Ebel (1965) concluded:

> The results of these hypothetical studies suggests that confidence weighting can be effective if the more capable students are also more discriminating than less capable students in choosing which responses to give confidently. But the results of recent experimental studies suggest that sometimes the more capable students are not much more successful than their less capable classmates in deciding when to answer confidently and when to answer cautiously. (p 56)

Ebel also found the attitude toward gambling, as did Swineford (1938), to affect the test score and to be uncorrelated with achievement. To neutralize the irrevelant influence of the gambling trait, Ebel suggested that the proportion of answers that must be given confidently be specified in advance for all students.

There have been two basic approaches to confidence testing described thus far. One method, the examinee may indicate any number of answers to be correct or incorrect. This approach is typified by Dressel and Schmid (1953). The other approach asks the examinee to first indicate his response and then to indicate his confidence in that response. Ebel (1965), Hevner (1932), Jacobs (1968), Soderquist (1936), and Swineford (1941, 1938) have used this approach to formula scoring. Each of these two methods gives a correct response given confidently more credit than a correct response given without confidence.

## Probabilistic Scoring

In 1965 the statistician de Finetti brought a high degree of mathematical sophistication to confidence testing by deriving formula scoring methods based on assumptions of examinee behavior, elements of decision theory, and personal probability. He posed the question of how an examinee should behave when he is required to choose one among k alternatives to a test item. The majority of earlier confidence-testing scoring formulas were quite arbitrary in their makeup. De Finetti's method, based on a mathematical model, presented a continuous scoring method which seemed very powerful. It was assumed that for each k-choice item, the degree of examinee partial knowledge

relevant to the item could be expressed in a complete and unique way by a set of values $p_j$, $j = 1, 2, \ldots, k$ such that

$$p_j \geq 0, \text{ and } \sum_{j=1}^{k} p_j = 1.$$

The $p_j$ values are the examinee's personal probabilities that the jth choice is the correct alternative. The item score takes the form

$$0 \leq S_h = 2p_h - \sum_{j=1}^{k} p_j^2 \leq 1,$$

where h is the correct alternative. In all cases the minimal value is attained when the total probability is concentrated on a single incorrect alternative and the maximum value is attained when it is all on the correct alternative. Since the penalty is the square of the distance from that point representing the examinee's opinion to the correct alternative, the examinee must indicate his true personal probability if he is to maximize his expected score.

Recognizing that the assignment of exact probabilities to each item alternative was a very difficult task, de Finetti experimented with several other simpler approaches to the problem. These alternate methods were designed to estimate an examinee's personal probability. The most notable of de Finetti's methods is the five-star scoring formula. This method restricts examinees to a finite set of probability responses in multiples of .2. Like the continuous method, examinees must place the five .2 stars on the item alternatives so as to indicate his relative strength of belief about the alternatives.

The distribution of stars for each item is referred to in tables provided by de Finetti to produce the item score.

Even though the theoretical work of de Finetti (1965) is promising, several psychological and operational factors were never considered. No studies of possible score contaminating factors (such as Swineford's gambling trait) have been done. Nothing is mentioned about the difficulty of the directions, time necessary for hand scoring, increase or decrease of testing time, or improvement in test reliability.

Other authors approached the confidence testing problem using scoring formulas with reproducing properties; that is, an examinee could maximize his expected score with respect to his personal probability distribution only if he honestly indicates his personal probabilities. Early work in this field was done by Toda (1963). Toda experimented with logarithmic and quadratic schemes. Roby (1965) reported a spherical scoring formula.

Shuford, Albert, and Massengill (1966) in an important paper suggested that a larger amount of information can be extracted from objective test items than is accomplished by a standard scoring method. They further suggested that the additional information about ability is contained in an examinee's personal probabilities for various item alternatives. Their scoring formula is termed admissible probability measurement and has reproducing properties.

Although their formula scoring procedures went through some evolution, a single truncated logarithmic scoring function was developed, and is given below

$$f(R_k) = \left\{ \begin{array}{ll} 1 + \log_{10} R_k & \text{for } .01 < R_k \leq 1 \\ -1 & \text{for } 0 \leq R_k \leq .01 \end{array} \right\}$$

where $R_k$ is the probability given to the correct response. Shuford, et al. have shown that a payoff function is necessary if the examinee is to be expected to indicate his true level of certainty. They further demonstrated for conditions with more than two alternatives that the logarithmic is the only valid method to use. Shuford et al. have marketed their scoring technique in a kit form.

Ebel (1968) acknowledged the logic of their method but criticized the kit because the administration time was nearly double that of a conventional test, and the kit itself was too complex. He further cited the lack of evidence of increases in validity and reliability. Echternacht (1971) criticized the work of Shuford, et al. for lack of control groups and very small sample sizes. He further concludes that confidence test scores (using the truncated logarithmic scoring function) could be higher than conventional right-only scores in part because of the scoring scheme. Hansen (1971) found that examinees displayed a tendency to either be confident or not. This confidence characteristic was found to be stable from test to test and only slightly correlated with the examinee's knowledge. Hansen concluded that training in the use of confidence testing methods does not reduce the error in the scoring system. To ease the understanding of directions and difficulty in scoring, Michael (1968) experimented with a simpler modification of personal probability. Her scoring

formula required examinees to give 10 points to the various item alternatives. Each item was scored by the proportion of points given to the correct response. Michael found higher reliabilities and lower standard errors using this method. Ripply (1970) used Michael's method in a study and recommended its use because of the scoring ease and high reliabiltiy.

Regardless of the specific formula scoring used, the primary purpose of confidence weighting and subjective probability has been to increase ability-related variance while reducing error variation. It is in this light that it must be evaluated (Lord, 1968).

Several studies have shown these scoring formulas to be complex and difficult for subjects to understand. Other studies have pointed out the existence of a general "gambling" factor that may actually increase error variation in the test.

Ripply (1971) and Ebel (1965) suggested male and female differences on the gambling trait and that examinees don't handle their confidence well.

As Stanley and Wang (1970) stated:

> The derivation of optimum response strategies in multiple choice testing represents an application of mathematical decision theory which underscores the decision process inherent in such tests. The success of testing procedures which attempt to control the decision process will be critically dependent on the ability of the subjects to effectively use optimal strategies. It is not certain that all subjects are equally capable of learning to use such strategies.

There have been improved reliability coefficients and other evidence of the usefulness of the above procedures, but Garvin (1972) points out that

> widely disparate situational factors--test length, format, difficulty, and content, and respondent motivation--and most important, disparate experimental methodologies, make it difficult to abstract generalizations from these studies. (p 4)

The rank order scoring procedure ( to be defined in Chapter III) offers some relief at this point. It is an obvious alternative to probabilistic scoring (mentioned by de Finetti, 1965) which makes explicit the probability distributions for items having varied characteristics. This will allow the assessment of model capabilities independent of the determination of the fit of the model to empirical data. The question of model capabilities is more basic since for models showing insufficient promise, tests of empirical fit would be superfluous.

However, from an empirical view point, ranking procedures should be very easy to teach examinees and should make it difficult for examinees to adopt a strategy, other than to respond honestly, that would maximize their expected score.

# CHAPTER III

# METHODOLOGY

This study was designed to compare the reliabilty and validity of a binary and a rank order latent trait test model over a range of situations. This was to be accomplished by the computer simulation of the conditional, joint, and marginal probability distributions of test score for each of the two models. The variance and covariances necessary for the computation of item reliability and validity followed from these probability distributions.

## Basic Assumptions of the Binary Model

It is assumed that the trait or ability under consideration can be thought of as an ordered variable represented numerically in a single dimension. This means that the examinees are considered as existing on a continuum in a way that implies that the amount of ability an examinee possesses is represented quantitatively by his position on the continuum.

The following are also assumed.

1) The proportion of correct responses made by examinees of very low ability will be close to 1/k, where k is the number of alternatives. The proportion of correct responses made by examinees of very high ability will be close to 1.0.

2) The proportion of correct responses increases as the ability level of the examinees increases.

3) All examinees will answer each test item.

4) Examinee ability is normally distributed in the population.

5) The number of examinees at any specified level of ability is assumed to be so large that sampling fluctuations may be ignored.

Three Parameter Normal Ogive Model for Binary Score

In this model, the item characteristic curve takes the form

$$P_g(\theta_i) = c_g + (1 - c_g)\Phi(a_g(\theta_i - b_g)) \qquad (3.1)$$

Where $P_g(\theta_i)$ is the probability that an examinee with ability $\theta_i$ answers item g correctly. The parameter $a_g$ is the item discrimination index and is proportional to the slope of $P_g(\theta_i)$ at the point $\theta_i = b_g$. This parameter indicates the quality of an item in the basic sense of the amount of information the item provides about $\theta$. The parameter $b_g$ is the item difficulty index and represents the point on the ability scale at which the slope of the item characteristic curve is a maximum. The parameter $c_g$ is the guessing parameter or the lower asymptote of the item characteristic curve. The symbol $\Phi$ indicates the cumulative normal distribution function. It can be seen from (3.1) that an item will only be useful if the probability of a correct answer increases as $\theta$ increases. It is for this reason that consideration will be restricted to items having the properity $0 < a_g \leq \infty$. It is assumed that $-\infty < b_g < \infty$ and $c_g = 1/k$, where k is the number of item alternatives.

Three Parameter Normal Ogive Model for Rank-Order Score

A test administered under the rank order model requires the examinee to rank order, using the ranks 1 to k, the alternatives he believes to be most, second,....., and least correct. In addition to

to the assumptions of the binary model, the rank order model assumes that the ranks (1, 2,....., k) an examinee places on the correct alternative may be used as an index of his partial knowledge of the trait or ability being measured. Limiting consideration to the rank placed on the correct alternative has the effect of reducing the number of possible item scores from k! to k. The value $X_h$ is the item score if the correct alternative is given the rank h (h = 1, 2,....., k) with $X_h$ decreasing ($X_1 > X_2 > ..... X_k$). An examination of equation (3.1) reveals the probability of successfully identifying the correct alternative in the normal ogive model for binary scores to be equivalent to placing the rank of one on the correct alternative in the normal ogive model for rank order scores. The model considered here states that the probability of placing the rank of one ($P(R_1)$) on the correct alternative of item $X_g$ given ability $\Theta_i$ takes the form

$$P(R_1) = P_g(\Theta_i) = P(X_g=1|\Theta_i)$$

$$= c_g + (1 - c_g)\Phi(a_g(\Theta_i - b_g) \qquad (3.2)$$

Equation (3.2) indicates that the examinee of ability $\Theta_i$ has assigned the first rank with a probability $P(R_1)$ that he assigned it to the correct alternative. Consideration now turns to the probability $P(R_2)$ that the examinee will place the rank of two on the correct alternative. Let $P_a\Theta_i = \Phi(a_g(\Theta_i - b_g))$. Also $\Sigma P(R_k) = 1.0$. With k-1 ranks remaining to be assigned, the probability of the examinee assigning the rank of two to the correct alternative is hypothesized to take the form

$$P(R_2) = P(X_g=2|\theta_i) =$$

$$\{1 - P(R_1)\}\{P_a\theta_i + c_{g2}(1 - P_a\theta_i)\} , \qquad (3.3)$$

where $c_{g2} = 1/(k - 1)$

Following a similar line of reasoning, the probability that the examinee assigns the rank of three to the correct alternative takes the form

$$P(R_3) = P(X_g=3|\theta_i) =$$

$$\{1 - P(R_1) - P(R_2)\}\{P_a\theta_i + c_{g3}(1 - P_a\theta_i)\} , \qquad (3.4)$$

where $c_{g3} = 1/(k - 2)$

The remaining two ranks in a five choice item follow the same pattern and are:

$$P(R_4) = P(X_g=4|\theta_i) = \qquad (3.5)$$

$$\{1 - P(R_1) - P(R_2) - P(R_3)\}\{P_a\theta_i + c_{g4}(1 - P_a\theta_i)\} ,$$

where $c_{g4} = 1/(k - 3)$

$$P(R_5) = P(X_g=5|\theta_i) = 1 - P(R_1) - P(R_2) - P(R_3) - P(R_4) \qquad (3.6)$$

The normal ogive model for rank ordered alternatives takes the general form

$$P(X_g=h|\theta_i) = \begin{cases} \{c_g + (1 - c_g)\Phi(a_g(\theta_i - b_g)\} \text{ if } h=1 \\ \{1 - \Sigma P(R_j)\}\{P_a\theta_i + c_{ga}(1 - P_a\theta_i)\} \end{cases} \qquad (3.7)$$

if $h > 1$

## Reliability and Coefficient of Effective Length

One point of comparison between the binary and rank order scoring formulas is their respective reliabilities. The reliability of a test is defined as the correlation between observed score (x) and true score (t).

$$\rho(x,t) = \frac{\sigma_t^2}{\sigma_x^2} \tag{3.8}$$

where: $\sigma_t^2$ is the true score variance, and

$\sigma_x^2$ is the observed score variance.

Since improvement in reliability is a main point of interest it is necessary to provide a suitable metric for expressing this factor. The Coefficient of Effective Length for Reliability (CEL-R) serves this purpose. (Gulliksen, 1950, p 83)

$$\text{CEL-R} = \frac{(1 - r_{11})R_{kk}}{(1 - R_{kk})r_{11}} \tag{3.9}$$

where: $r_{11}$ is the reliability of a binary scored test item, and $R_{kk}$ is the reliability of a rank order scored test item. The CEL-R is interpreted as the factor by which the binary scored test would have to be lengthened or shortened to yield the reliability of the same test administered using the rank order scoring procedure.

## Validity and Coefficient of Effective Length

A second point of comparison between the two formula scoring procedures is their respective validities. The validity of a test item is defined as the correlation between observed test score (x) and underlying ability ($\theta$).

$$\rho_{(x,\theta)} = \frac{\mathrm{cov}(x,\theta)}{\sigma(\theta)\sigma(x)} \tag{3.10}$$

But since the distribution of $\theta$ is assumed to be N(0,1), equation (3.10) becomes

$$\rho_{(x,\theta)} = \frac{\mathrm{cov}(x,\theta)}{\sigma(x)} \tag{3.11}$$

Since improvement in validity is a main point of interest it is necessary to provide a suitable metric for expressing this factor. The Coefficient of Effective Length for Validity (CEL-V) serves this purpose. (Gulliksen, 1950, p. 93)

$$\text{CEL-V} = \frac{R_{ki}^2(1 - r_{11})}{r_{1i}^2 - r_{11}R_{ki}^2} \tag{3.12}$$

where; $r_{1i}$ is the validity of a binary scored item,

$r_{11}$ is the reliability of a binary scored item, and

$R_{ki}$ is the validity of a rank order scored item.

The CEL-V is interpreted as the factor by which a binary scored test would have to be lengthened or shortened to yield the validity of the same test administered under the rank order scoring procedure.

Conditional, Joint, and Marginal Distributions

The variances and covariances necessary for the computation of item reliability and validity are constructed from the conditional distribution of test score $x_g$ for a fixed $\theta_i$, and the joint distribution of $x_g$ and $\theta_i$. These distributions follow directly from the definition of $\theta_i$ and its probability distribution.

It can be seen from the relationship of equation (3.1) to (3.2) that the binary model is a special case of the rank order model. This relationship allows the definition of the conditional, joint, and marginal distributions to follow a general form. Since the probability of a point on a continuous function is equal to zero, ability is specified as a set of discrete points in units of standard deviation. The area contained within the interval $\Phi(\theta_i - \theta_{i-1})/2$ to $\Phi(\theta_i - \theta_{i+1})/2$ is used as an estimate of the probability of the point $\theta_i$. This area is calculated for each point $\theta_i$ from $-3\sigma$ to $+3\sigma$ in increments of $0.2\sigma$.

The conditional distribution of test score is a (k,n) matrix with k ranks and n $\theta$ points if $\theta$ is discrete.

$$
\begin{array}{llll}
P(X_g=1|\theta_1) & P(X_g=1|\theta_{i+1}) & \cdots\cdots\cdots & P(X_g=1|\theta_n) \\
P(X_g=2|\theta_1) & \cdots\cdots\cdots\cdots\cdots\cdots\cdots & & P(X_g=2|\theta_n) \\
\vdots & & & \vdots \\
P(X_g=k|\theta_1) & \cdots\cdots\cdots\cdots\cdots\cdots\cdots & & P(X_g=k|\theta_n)
\end{array}
$$

It can be seen that the conditional distribution of test score for the binary model is found in the first row of this matrix. The joint distribution $(P(X_k,\theta))$ of observed score (X) and $(\theta)$ is obtained by multiplying each entry in the matrix of conditional probabilities by its corresponding probability of $P(\theta_i)$.

The marginal distribution of observed score ($P(X_g=k)$) is obtained by summing the rows of the joint distribution. This yields a k element vector.

Having specified the conditional, joint, and marginal distributions of test score, true and observed score variances are calculated.

True Score and Observed Score Variance

The binary and rank order normal ogive models assume $\theta_i$ to be the only source of true variance among people. It follows then that when $\theta_i$ is fixed true score is also fixed. As a result, the expected value of observed score for a fixed $\theta_i$ is the true score for $\theta_i$.

Let $\tau_i$ equal the true score corresponding to the ability level $\theta_i$ and let $X_h$ equal observed test score. The item true score takes the general form

$$\begin{aligned}\tau_i &= E(X_h|\theta_i) \\ &= \Sigma P(X_g=k|\theta_i)\cdot w_g \\ &= P(X_g=1|\theta_i)\cdot w_1 + P(X_g=2|\theta_i)\cdot w_2 + \ldots. \\ &\qquad \ldots. P(X_g=k|\theta_i)\cdot w_k \end{aligned} \tag{3.13}$$

where $w_g$ are the item alternative weights. For the binary model, there are only two possible outcomes and the correct one receives a weight of one while all other alternatives receive weights of zero.

$$\begin{aligned}\tau_b &= P(X_g=1|\theta_i)\cdot 1 + P(X_g=2,3,4,5|\theta_i)\cdot 0 \\ &= P(X_g=1|\theta_i)\end{aligned} \tag{3.14}$$

True score variance ($\sigma_\tau^2$) follows from the expected values of the sum and sum of squares of true score.

$$\sigma_\tau^2 = E\tau^2 - E(\tau)^2 \qquad (3.15)$$

$$= \Sigma\, P(\tau_i)\cdot\tau_i^2 - \{\, \Sigma\, P(\tau_i)\cdot\tau_i \,\}^2 \qquad (3.16)$$

$$= \Sigma\, P(\theta_i)^2\cdot\theta_i - \{\, \Sigma\, P(\theta_i)\cdot\theta_i \,\}^2 \qquad (3.17)$$

Observed score variance follows from the expected value of the marginal distribution.

$$\sigma_x^2 = Ex^2 - E(x)^2 \qquad (3.18)$$

$$= \Sigma\, P(X_g{=}k)\cdot w_k - \Sigma\, P(X_g{=}k)\cdot w_k^2 \qquad (3.19)$$

where $P(X_g{=}k)$ is the marginal probability for the kth alternative and $w_k$ is the scoring weight for the kth alternative.

## Procedure

Test items were simulated using the normal ogive models for binary and rank order scoring of multiple-choice items discussed earlier. The marginal distributions of test score and true score for the simulated items were used to compute item reliability. The joint distribution of observed score and ability were used to compute validity. The resulting reliabilities and validities were contrasted by expressing them as coefficients of effective lengths for reliability (CEL-R) and validity (CEL-V). The thirty-six items simulated were made up of all combinations of item discrimination ( 0.5 to 2.5) and item difficulty ( -1.5 to 2.5) in increments of 0.5.

# CHAPTER IV

## RESULTS

A computer simulation of the conditional, joint, and marginal distribution of test score for items scored using the binary and rank order normal ogive formula scoring procedures was performed. By varying the item difficulty and item discrimination, thirty-six different test items were simulated. The item reliabilities and validities calculated for each item using the two different procedures and the coefficients of effective length for reliability (CEL-R) and validity (CEL-V) for the thirty-six items are presented in Table 1.

An inspection of Table 1 shows item reliability decreases as item difficulty ($b_g$) increases for a fixed level of item discrimination. Except for items 31 to 35, items scored using the rank order procedure have reliabilities equal to or higher than the same items scored using the binary procedure. The greatest gains in reliability (largest CEL-R) result when the item discrimination index ($a_g$) is less than or equal to 1.0. Alternatively, if the item difficulty is held constant the CEL-R decreases as the item discrimination index ($a_g$) increases. Rank order scoring produces the greatest gain in reliability over the binary scoring for very easy and very difficult test items.

If the item discrimination index is held constant, item validity increases as item difficulty increases to $b_g = 0.0$ and then decreases. This is the attenuation paradox (Loevinger, 1954). Although improvement in validity does not always favor the rank order scoring procedure,

when item discrimination is held constant the improvement in validity (CEL-V) increases as item difficulty increases. In general, as item discrimination increases the CEL-V decreases with the smallest CEL-V occurring with the highest item discrimination and lowest item difficulty. The largest CEL-V's occur with the more difficult test items. Test items 1 to 9 and 12 to 18 represent combinations of item difficulty and item discrimination commonly found in aptitude and achievement testing (Lord, 1968). Scoring these items using the rank order procedure results in gains in reliability and validity. It should be noted that the greatest gains in reliability and validity only occur for the more difficult test items.

In order to further illustrate the relationship between item discrimination, item difficulty, and underlying ability, nine items (1, 5, 7, 10, 14, 16, 19, 23, 25 from Table 1.) representing combinations of easy, moderate, and high difficulty with moderate, high, and very high discrimination were chosen and their conditional distributions of rank order score were plotted (Figs. 1 - 9). From top to bottom, the curves represent $P(R_1)$, $P(R_2)$,....., $P(R_k)$. For each of these nine items the conditional error variance (scaled for total test variance) at each point on the ability continuum calculated (Tables 2 - 10).

Table 1

Summary Statistics

for

Binary and Rank Order Models

| | ITEM PARAMETERS | | BINARY ITEM | | RANK ORDER ITEM | | IMPROVEMENT | |
|---|---|---|---|---|---|---|---|---|
| | $a_g$ | $b_g$ | Reliability | Validity | Reliability | Validity | CEL-R | CEL-V |
| 1) | 0.5 | -1.5 | 0.09 | 0.28 | 0.17 | 0.29 | 1.82 | 1.09 |
| 2) | 0.5 | -1.0 | 0.09 | 0.29 | 0.15 | 0.31 | 1.65 | 1.18 |
| 3) | 0.5 | -0.5 | 0.09 | 0.29 | 0.14 | 0.32 | 1.58 | 1.26 |
| 4) | 0.5 | 0.0 | 0.09 | 0.29 | 0.14 | 0.33 | 1.55 | 1.35 |
| 5) | 0.5 | 0.5 | 0.08 | 0.27 | 0.12 | 0.32 | 1.55 | 1.43 |
| 6) | 0.5 | 1.0 | 0.07 | 0.26 | 0.11 | 0.31 | 1.57 | 1.51 |
| 7) | 0.5 | 1.5 | 0.06 | 0.23 | 0.09 | 0.28 | 1.61 | 1.58 |
| 8) | 0.5 | 2.0 | 0.04 | 0.20 | 0.07 | 0.25 | 1.66 | 1.64 |
| 9) | 0.5 | 2.5 | 0.03 | 0.16 | 0.05 | 0.21 | 1.73 | 1.69 |
| 10) | 1.0 | -1.5 | 0.22 | 0.39 | 0.32 | 0.38 | 1.46 | 0.89 |
| 11) | 1.0 | -1.0 | 0.24 | 0.44 | 0.31 | 0.44 | 1.31 | 0.99 |
| 12) | 1.0 | -0.5 | 0.24 | 0.46 | 0.30 | 0.48 | 1.24 | 1.08 |
| 13) | 1.0 | 0.0 | 0.22 | 0.46 | 0.27 | 0.48 | 1.21 | 1.16 |
| 14) | 1.0 | 0.5 | 0.19 | 0.42 | 0.23 | 0.46 | 1.20 | 1.23 |
| 15) | 1.0 | 1.0 | 0.15 | 0.36 | 0.18 | 0.40 | 1.20 | 1.28 |
| 16) | 1.0 | 1.5 | 0.10 | 0.27 | 0.12 | 0.31 | 1.23 | 1.34 |
| 17) | 1.0 | 2.0 | 0.05 | 0.18 | 0.07 | 0.22 | 1.29 | 1.40 |
| 18) | 1.0 | 2.5 | 0.02 | 0.11 | 0.03 | 0.13 | 1.45 | 1.49 |

Table 1 (cont.)

| | $a_g$ | $b_g$ | Reliability | Validity | Reliability | Validity | CEL-R | CEL-V |
|---|---|---|---|---|---|---|---|---|
| 19) | 1.5 | -1.5 | 0.33 | 0.43 | 0.42 | 0.39 | 1.29 | 0.79 |
| 20) | 1.5 | -1.0 | 0.35 | 0.50 | 0.41 | 0.48 | 1.17 | 0.88 |
| 21) | 1.5 | -0.5 | 0.35 | 0.54 | 0.39 | 0.53 | 1.11 | 0.95 |
| 22) | 1.5 | 0.0 | 0.33 | 0.54 | 0.35 | 0.54 | 1.07 | 1.01 |
| 23) | 1.5 | 0.5 | 0.28 | 0.48 | 0.29 | 0.49 | 1.04 | 1.05 |
| 24) | 1.5 | 1.0 | 0.20 | 0.39 | 0.20 | 0.40 | 1.02 | 1.08 |
| 25) | 1.5 | 1.5 | 0.12 | 0.27 | 0.12 | 0.28 | 1.02 | 1.13 |
| 26) | 1.5 | 2.0 | 0.05 | 0.15 | 0.06 | 0.16 | 1.09 | 1.20 |
| 27) | 1.5 | 2.5 | 0.02 | 0.07 | 0.02 | 0.08 | 1.31 | 1.32 |
| 28) | 2.0 | -1.5 | 0.41 | 0.44 | 0.49 | 0.39 | 1.19 | 0.72 |
| 29) | 2.0 | -1.0 | 0.43 | 0.53 | 0.46 | 0.50 | 1.09 | 0.80 |
| 30) | 2.0 | -0.5 | 0.43 | 0.58 | 0.44 | 0.56 | 1.03 | 0.86 |
| 31) | 2.0 | 0.0 | 0.40 | 0.58 | 0.39 | 0.56 | 0.99 | 0.90 |
| 32) | 2.0 | 0.5 | 0.33 | 0.51 | 0.31 | 0.50 | 0.95 | 0.93 |
| 33) | 2.0 | 1.0 | 0.24 | 0.40 | 0.21 | 0.39 | 0.91 | 0.95 |
| 34) | 2.0 | 1.5 | 0.13 | 0.26 | 0.12 | 0.25 | 0.90 | 0.98 |
| 35) | 2.0 | 2.0 | 0.06 | 0.13 | 0.05 | 0.14 | 0.96 | 1.05 |
| 36) | 2.0 | 2.5 | 0.02 | 0.05 | 0.02 | 0.06 | 1.21 | 1.19 |

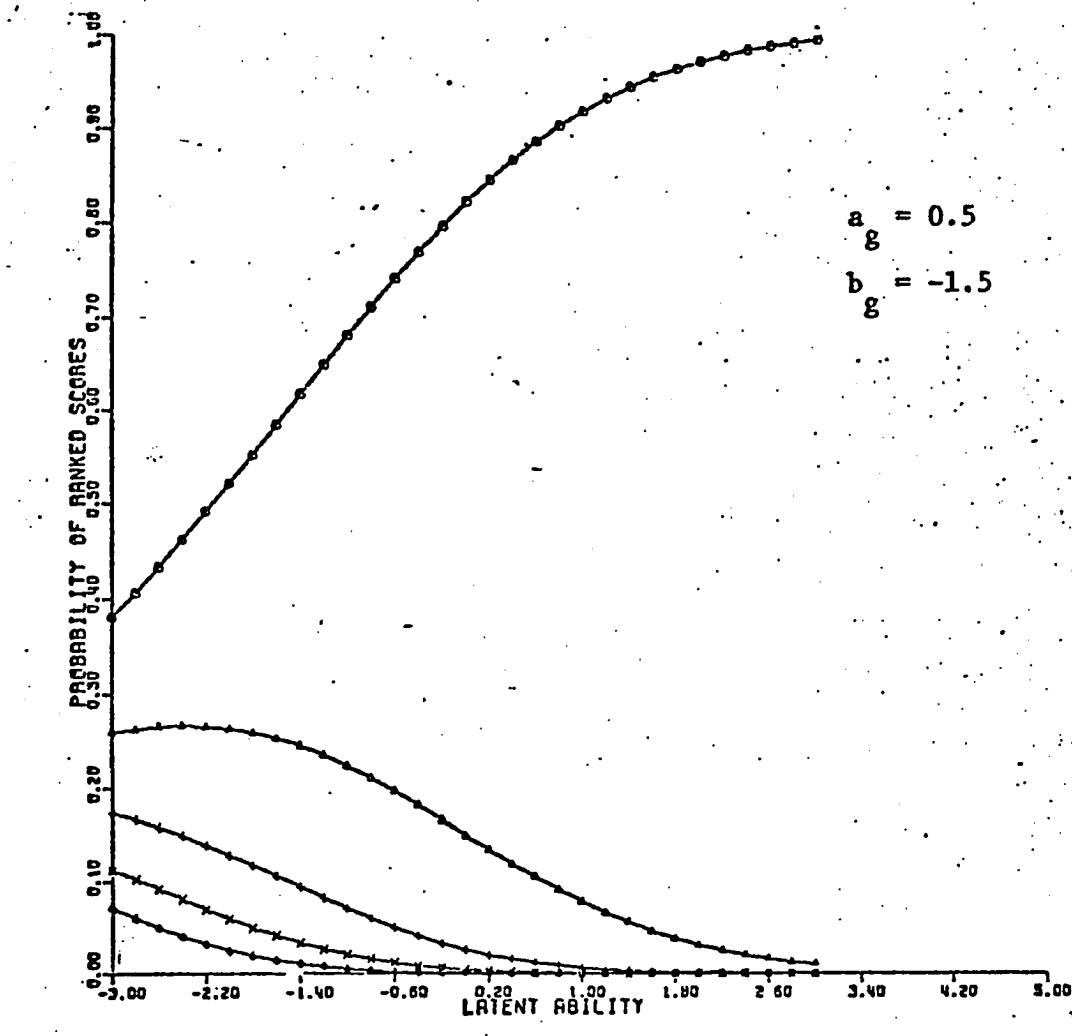



Figure 1. Plot of Conditional Distributions of Rank-Order Score for Item 1

TABLE 2

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 1.329 | 3.368 |
| 2) | -2.8 | 1.359 | 3.179 |
| 3) | -2.6 | 1.383 | 2.973 |
| 4) | -2.4 | 1.400 | 2.751 |
| 5) | -2.2 | 1.406 | 2.520 |
| 6) | -2.0 | 1.406 | 2.285 |
| 7) | -1.8 | 1.393 | 2.051 |
| 8) | -1.6 | 1.369 | 1.823 |
| 9) | -1.4 | 1.333 | 1.605 |
| 10) | -1.2 | 1.286 | 1.400 |
| 11) | -1.0 | 1.228 | 1.212 |
| 12) | -0.8 | 1.161 | 1.042 |
| 13) | -0.6 | 1.087 | 0.889 |
| 14) | -0.4 | 1.007 | 0.754 |
| 15) | -0.2 | 0.923 | 0.636 |
| 16) | 0.0 | 0.836 | 0.534 |
| 17) | 0.2 | 0.750 | 0.446 |
| 18) | 0.4 | 0.666 | 0.371 |
| 19) | 0.6 | 0.584 | 0.308 |
| 20) | 0.8 | 0.507 | 0.254 |
| 21) | 1.0 | 0.436 | 0.208 |
| 22) | 1.2 | 0.371 | 0.170 |
| 23) | 1.4 | 0.312 | 0.138 |
| 24) | 1.6 | 0.260 | 0.112 |
| 25) | 1.8 | 0.214 | 0.090 |
| 26) | 2.0 | 0.175 | 0.072 |
| 27) | 2.2 | 0.141 | 0.057 |
| 28) | 2.4 | 0.113 | 0.045 |
| 29) | 2.6 | 0.090 | 0.035 |
| 30) | 2.8 | 0.070 | 0.027 |
| 31) | 3.0 | 0.055 | 0.021 |

† $a_g = 0.5$, $b_g = -1.5$

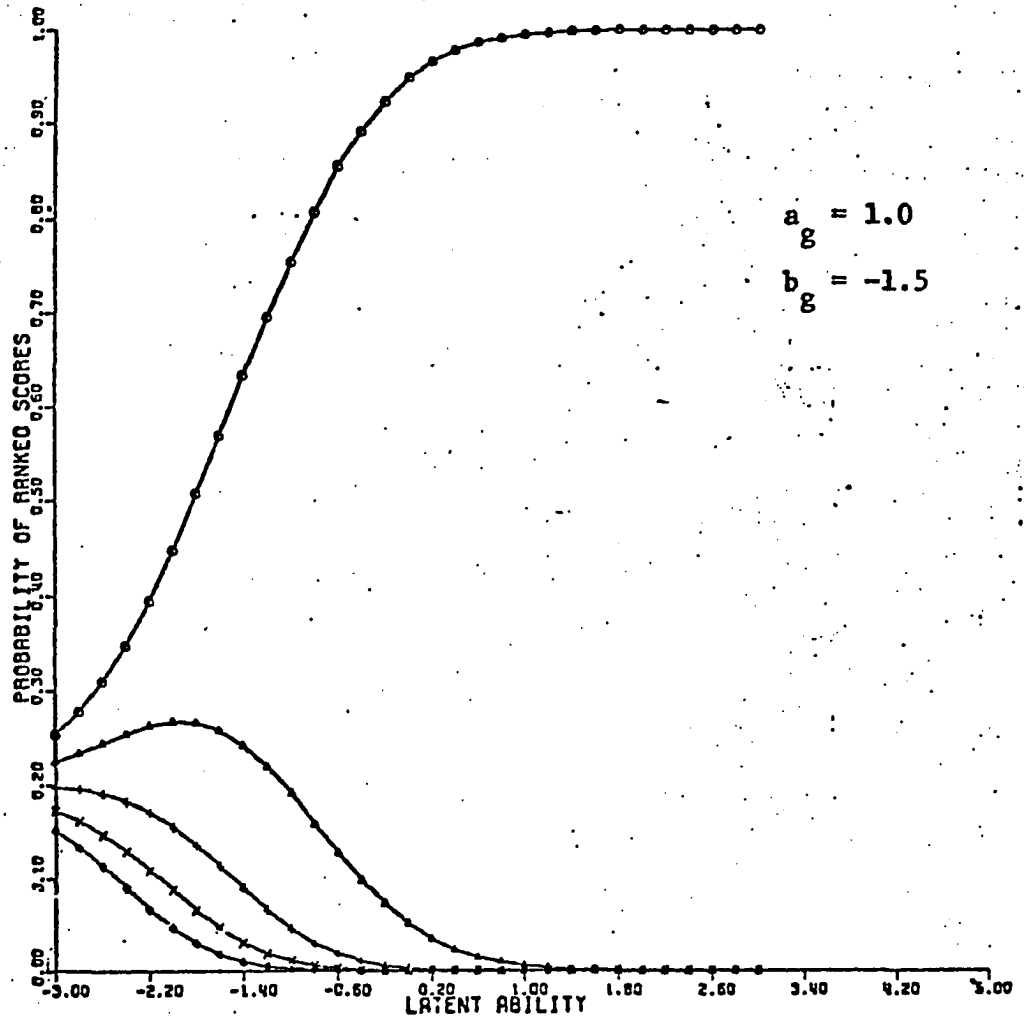



Figure 2. Plot of Conditional Distributions of Rank-Order Score for Item 10

TABLE 3.

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 0.658 | 1.406 |
| 2) | -2.8 | 0.673 | 1.400 |
| 3) | -2.6 | 0.689 | 1.391 |
| 4) | -2.4 | 0.708 | 1.379 |
| 5) | -2.2 | 0.729 | 1.363 |
| 6) | -2.0 | 0.752 | 1.341 |
| 7) | -1.8 | 0.775 | 1.314 |
| 8) | -1.6 | 0.800 | 1.280 |
| 9) | -1.4 | 0.825 | 1.239 |
| 10) | -1.2 | 0.849 | 1.190 |
| 11) | -1.0 | 0.871 | 1.133 |
| 12) | -0.8 | 0.890 | 1.070 |
| 13) | -0.6 | 0.906 | 1.000 |
| 14) | -0.4 | 0.917 | 0.926 |
| 15) | -0.2 | 0.923 | 0.848 |
| 16) | 0.0 | 0.921 | 0.769 |
| 17) | 0.2 | 0.913 | 0.690 |
| 18) | 0.4 | 0.897 | 0.613 |
| 19) | 0.6 | 0.873 | 0.540 |
| 20) | 0.8 | 0.842 | 0.471 |
| 21) | 1.0 | 0.805 | 0.408 |
| 22) | 1.2 | 0.761 | 0.350 |
| 23) | 1.4 | 0.712 | 0.299 |
| 24) | 1.6 | 0.660 | 0.254 |
| 25) | 1.8 | 0.604 | 0.214 |
| 26) | 2.0 | 0.548 | 0.180 |
| 27) | 2.2 | 0.491 | 0.150 |
| 28) | 2.4 | 0.436 | 0.125 |
| 29) | 2.6 | 0.383 | 0.103 |
| 30) | 2.8 | 0.332 | 0.085 |
| 31) | 3.0 | 0.286 | 0.070 |

† $a_g = 1.0$, $b_g = -1.5$

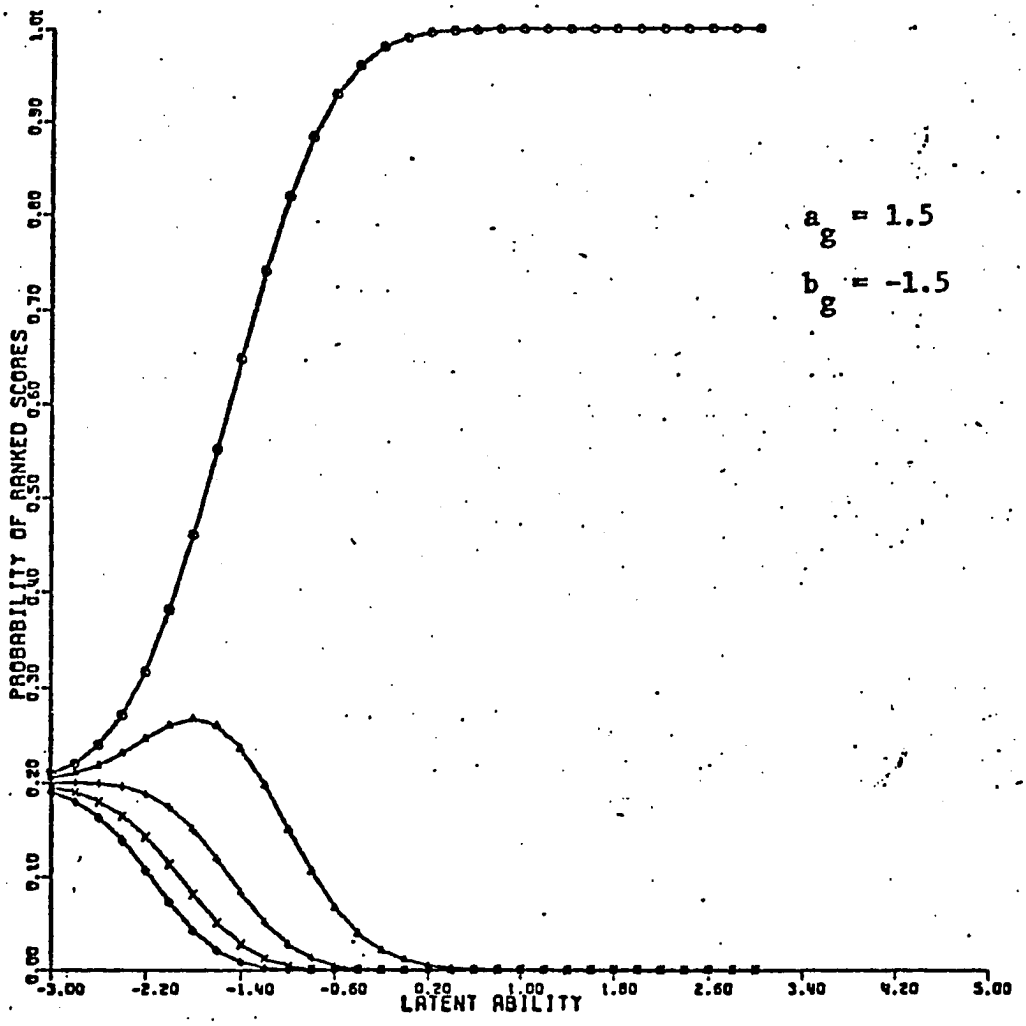



Figure 3. Plot of Conditional Distributions of Rank-Order Score for Item 19

TABLE 4

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 0.651 | 1.096 |
| 2) | -2.8 | 0.658 | 1.095 |
| 3) | -2.6 | 0.666 | 1.094 |
| 4) | -2.4 | 0.675 | 1.093 |
| 5) | -2.2 | 0.687 | 1.090 |
| 6) | -2.0 | 0.700 | 1.087 |
| 7) | -1.8 | 0.716 | 1.082 |
| 8) | -1.6 | 0.734 | 1.075 |
| 9) | -1.4 | 0.754 | 1.066 |
| 10) | -1.2 | 0.776 | 1.054 |
| 11) | -1.0 | 0.800 | 1.037 |
| 12) | -0.8 | 0.825 | 1.016 |
| 13) | -0.6 | 0.851 | 0.990 |
| 14) | -0.4 | 0.878 | 0.958 |
| 15) | -0.2 | 0.903 | 0.920 |
| 16) | 0.0 | 0.927 | 0.876 |
| 17) | 0.2 | 0.948 | 0.827 |
| 18) | 0.4 | 0.964 | 0.773 |
| 19) | 0.6 | 0.976 | 0.716 |
| 20) | 0.8 | 0.982 | 0.656 |
| 21) | 1.0 | 0.980 | 0.595 |
| 22) | 1.2 | 0.971 | 0.534 |
| 23) | 1.4 | 0.954 | 0.474 |
| 24) | 1.6 | 0.929 | 0.418 |
| 25) | 1.8 | 0.896 | 0.364 |
| 26) | 2.0 | 0.856 | 0.315 |
| 27) | 2.2 | 0.810 | 0.271 |
| 28) | 2.4 | 0.758 | 0.231 |
| 29) | 2.6 | 0.702 | 0.196 |
| 30) | 2.8 | 0.643 | 0.166 |
| 31) | 3.0 | 0.583 | 0.139 |

† $a_g = 1.5$, $b_g = -1.5$

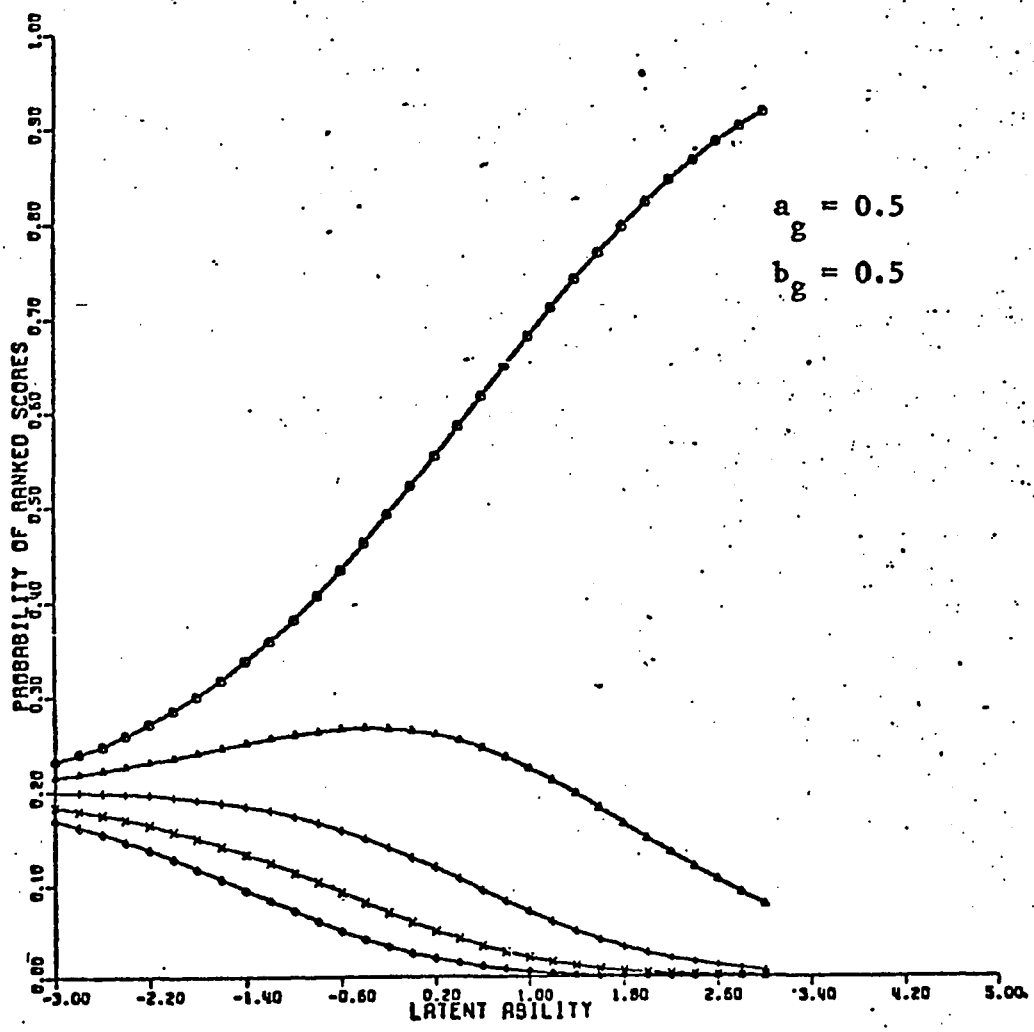



Figure 4. Plot of Conditional Distributions of Rank-Order Score for Item 5

TABLE 5

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 1.428 | 4.498 |
| 2) | -2.8 | 1.513 | 4.392 |
| 3) | -2.6 | 1.610 | 4.214 |
| 4) | -2.4 | 1.711 | 3.946 |
| 5) | -2.2 | 1.802 | 3.580 |
| 6) | -2.0 | 1.866 | 3.129 |
| 7) | -1.8 | 1.887 | 2.626 |
| 8) | -1.6 | 1.852 | 2.115 |
| 9) | -1.4 | 1.756 | 1.639 |
| 10) | -1.2 | 1.602 | 1.229 |
| 11) | -1.0 | 1.403 | 0.895 |
| 12) | -0.8 | 1.178 | 0.637 |
| 13) | -0.6 | 0.948 | 0.445 |
| 14) | -0.4 | 0.730 | 0.305 |
| 15) | -0.2 | 0.539 | 0.206 |
| 16) | 0.0 | 0.382 | 0.136 |
| 17) | 0.2 | 0.260 | 0.088 |
| 18) | 0.4 | 0.169 | 0.055 |
| 19) | 0.6 | 0.106 | 0.034 |
| 20) | 0.8 | 0.064 | 0.020 |
| 21) | 1.0 | 0.037 | 0.012 |
| 22) | 1.2 | 0.021 | 0.006 |
| 23) | 1.4 | 0.011 | 0.003 |
| 24) | 1.6 | 0.006 | 0.002 |
| 25) | 1.8 | 0.003 | 0.001 |
| 26) | 2.0 | 0.001 | 0.000 |
| 27) | 2.2 | 0.001 | 0.000 |
| 28) | 2.4 | 0.000 | 0.000 |
| 29) | 2.6 | 0.000 | 0.000 |
| 30) | 2.8 | 0.000 | 0.000 |
| 31) | 3.0 | 0.000 | 0.000 |

† $a_g = 0.5$, $b_g = 0.5$

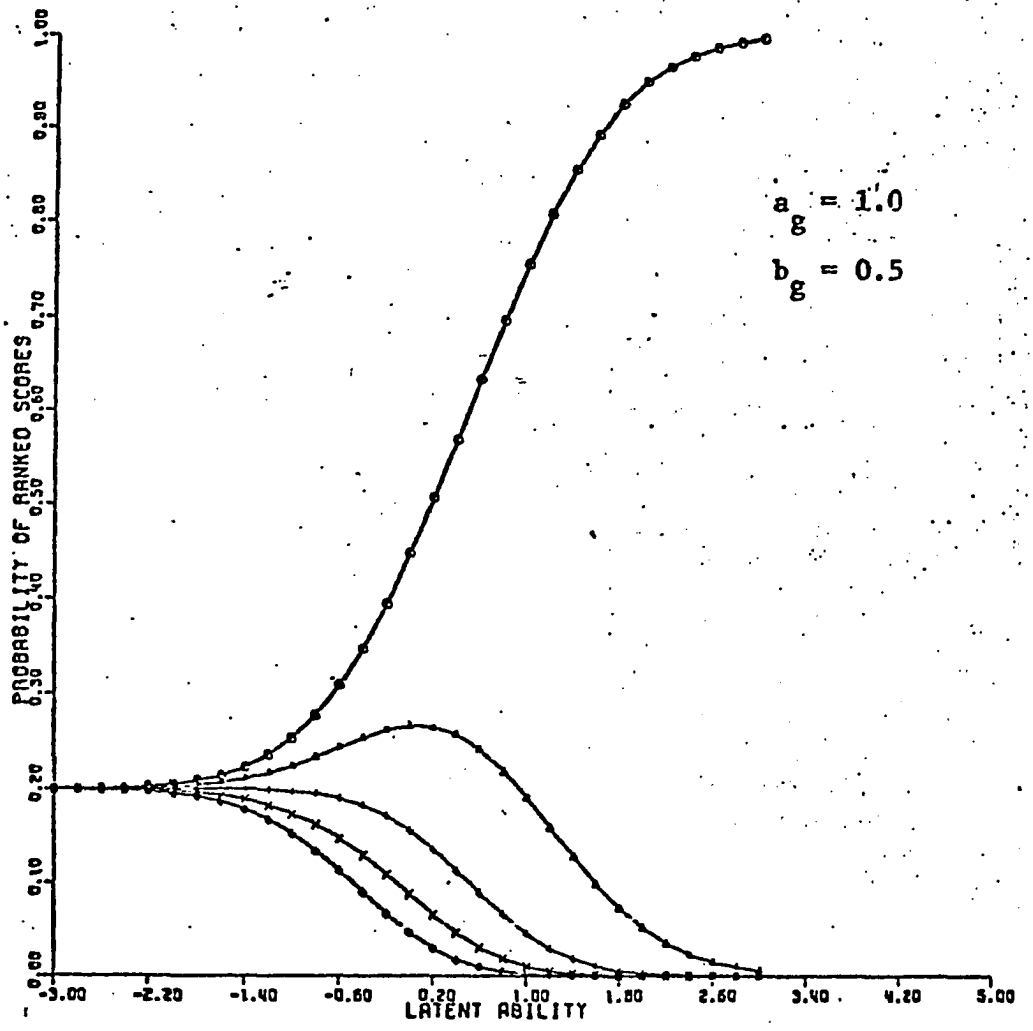



Figure 5. Plot of Conditional Distributions of Rank-Order Score for Item 14

TABLE 6

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 0.517 | 0.960 |
| 2) | -2.8 | 0.517 | 0.960 |
| 3) | -2.6 | 0.518 | 0.960 |
| 4) | -2.4 | 0.519 | 0.960 |
| 5) | -2.2 | 0.522 | 0.960 |
| 6) | -2.0 | 0.526 | 0.960 |
| 7) | -1.8 | 0.533 | 0.959 |
| 8) | -1.6 | 0.543 | 0.958 |
| 9) | -1.4 | 0.559 | 0.955 |
| 10) | -1.2 | 0.581 | 0.949 |
| 11) | -1.0 | 0.610 | 0.937 |
| 12) | -0.8 | 0.647 | 0.915 |
| 13) | -0.6 | 0.688 | 0.878 |
| 14) | -0.4 | 0.731 | 0.822 |
| 15) | -0.2 | 0.770 | 0.746 |
| 16) | 0.0 | 0.797 | 0.652 |
| 17) | 0.2 | 0.806 | 0.547 |
| 18) | 0.4 | 0.792 | 0.441 |
| 19) | 0.6 | 0.750 | 0.342 |
| 20) | 0.8 | 0.685 | 0.256 |
| 21) | 1.0 | 0.600 | 0.187 |
| 22) | 1.2 | 0.504 | 0.133 |
| 23) | 1.4 | 0.405 | 0.093 |
| 24) | 1.6 | 0.312 | 0.064 |
| 25) | 1.8 | 0.230 | 0.043 |
| 26) | 2.0 | 0.163 | 0.028 |
| 27) | 2.2 | 0.111 | 0.018 |
| 28) | 2.4 | 0.072 | 0.011 |
| 29) | 2.6 | 0.045 | 0.007 |
| 30) | 2.8 | 0.027 | 0.004 |
| 31) | 3.0 | 0.016 | 0.002 |

† $a_g = 1.0$, $b_g = 0.5$

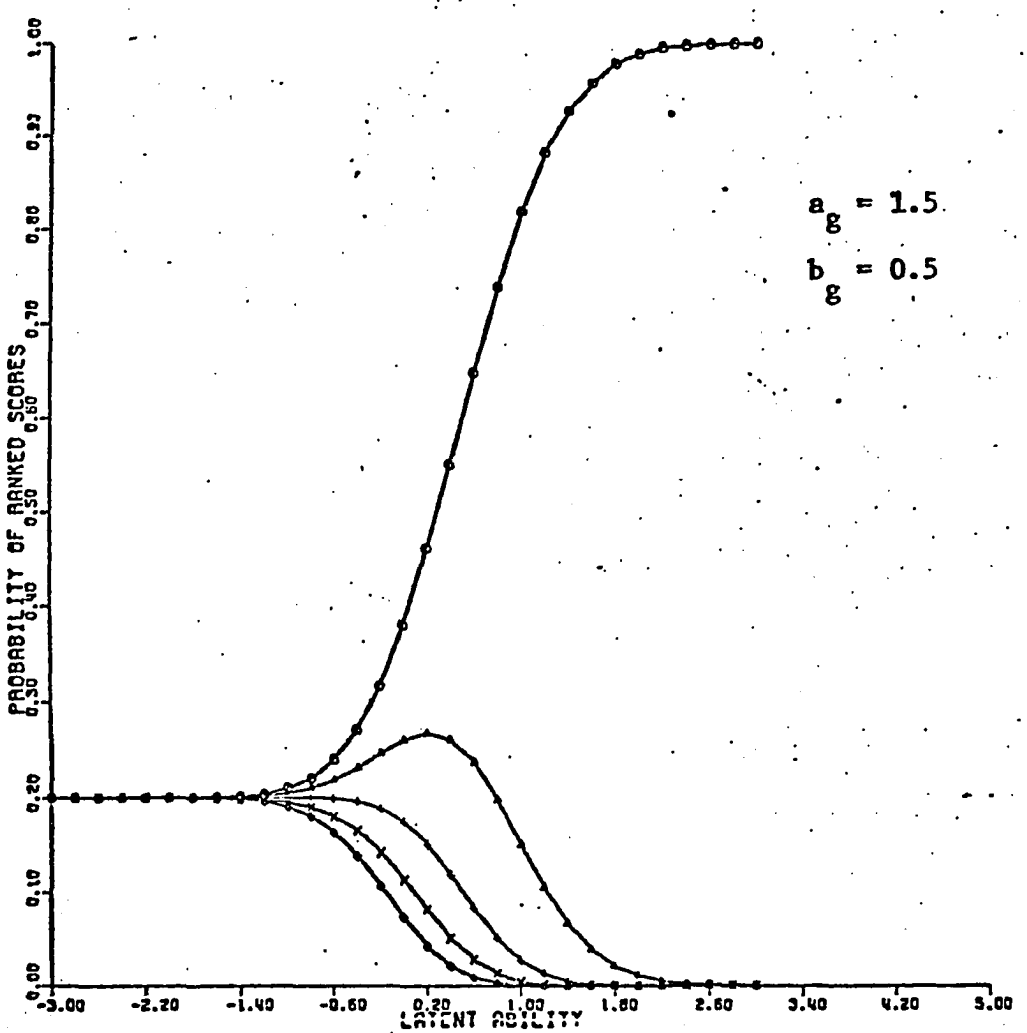



Figure 6. Plot of Conditional Distributions of Rank-Order Score for Item 23

TABLE 7

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 0.668 | 0.892 |
| 2) | -2.8 | 0.668 | 0.892 |
| 3) | -2.6 | 0.668 | 0.892 |
| 4) | -2.4 | 0.668 | 0.892 |
| 5) | -2.2 | 0.668 | 0.892 |
| 6) | -2.0 | 0.669 | 0.892 |
| 7) | -1.8 | 0.669 | 0.892 |
| 8) | -1.6 | 0.670 | 0.892 |
| 9) | -1.4 | 0.672 | 0.892 |
| 10) | -1.2 | 0.675 | 0.892 |
| 11) | -1.0 | 0.681 | 0.891 |
| 12) | -0.8 | 0.689 | 0.891 |
| 13) | -0.6 | 0.703 | 0.890 |
| 14) | -0.4 | 0.724 | 0.888 |
| 15) | -0.2 | 0.752 | 0.882 |
| 16) | 0.0 | 0.790 | 0.871 |
| 17) | 0.2 | 0.837 | 0.850 |
| 18) | 0.4 | 0.891 | 0.816 |
| 19) | 0.6 | 0.947 | 0.764 |
| 20) | 0.8 | 0.997 | 0.693 |
| 21) | 1.0 | 1.032 | 0.606 |
| 22) | 1.2 | 1.044 | 0.508 |
| 23) | 1.4 | 1.025 | 0.409 |
| 24) | 1.6 | 0.971 | 0.317 |
| 25) | 1.8 | 0.886 | 0.238 |
| 26) | 2.0 | 0.776 | 0.173 |
| 27) | 2.2 | 0.652 | 0.123 |
| 28) | 2.4 | 0.524 | 0.086 |
| 29) | 2.6 | 0.404 | 0.059 |
| 30) | 2.8 | 0.298 | 0.040 |
| 31) | 3.0 | 0.211 | 0.026 |

† $a_g = 1.5$, $b_g = 0.5$

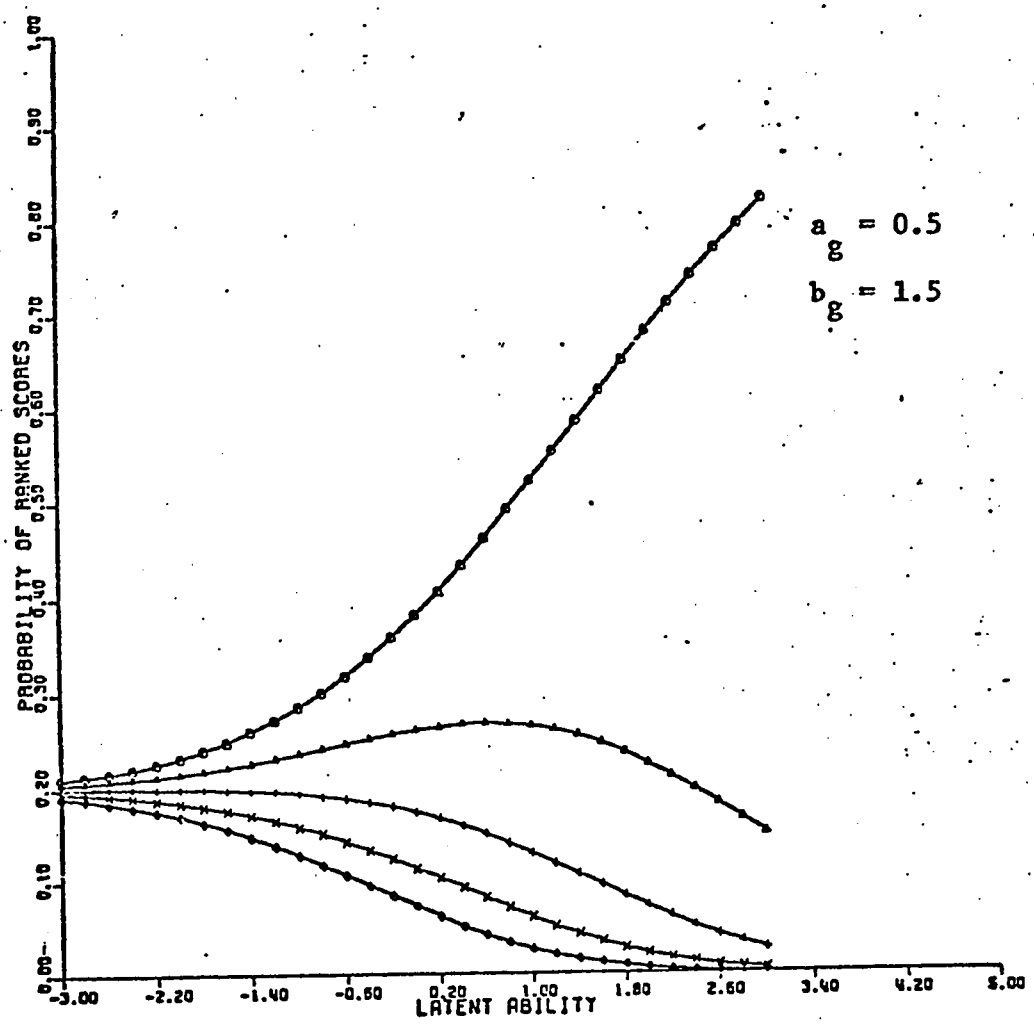



Figure 7. Plot of Conditional Distributions of Rank-Order Score for Item 7

TABLE 8

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 1.415 | 4.149 |
| 2) | -2.8 | 1.467 | 4.137 |
| 3) | -2.6 | 1.555 | 4.098 |
| 4) | -2.4 | 1.685 | 3.989 |
| 5) | -2.2 | 1.849 | 3.747 |
| 6) | -2.0 | 2.013 | 3.318 |
| 7) | -1.8 | 2.121 | 2.711 |
| 8) | -1.6 | 2.110 | 2.020 |
| 9) | -1.4 | 1.947 | 1.379 |
| 10) | -1.2 | 1.646 | 0.876 |
| 11) | -1.0 | 1.267 | 0.526 |
| 12) | -0.8 | 0.885 | 0.303 |
| 13) | -0.6 | 0.561 | 0.167 |
| 14) | -0.4 | 0.324 | 0.088 |
| 15) | -0.2 | 0.171 | 0.044 |
| 16) | 0.0 | 0.083 | 0.021 |
| 17) | 0.2 | 0.037 | 0.009 |
| 18) | 0.4 | 0.015 | 0.004 |
| 19) | 0.6 | 0.006 | 0.001 |
| 20) | 0.8 | 0.002 | 0.000 |
| 21) | 1.0 | 0.001 | 0.000 |
| 22) | 1.2 | 0.000 | 0.000 |
| 23) | 1.4 | 0.000 | 0.000 |
| 24) | 1.6 | 0.000 | 0.000 |
| 25) | 1.8 | 0.000 | 0.000 |
| 26) | 2.0 | 0.000 | 0.000 |
| 27) | 2.2 | 0.0 | 0.0 |
| 28) | 2.4 | 0.0 | 0.0 |
| 29) | 2.6 | 0.0 | 0.0 |
| 30) | 2.8 | 0.0 | 0.0 |
| 31) | 3.0 | 0.0 | 0.0 |

† $a_g = 0.5$, $b_g = 1.5$

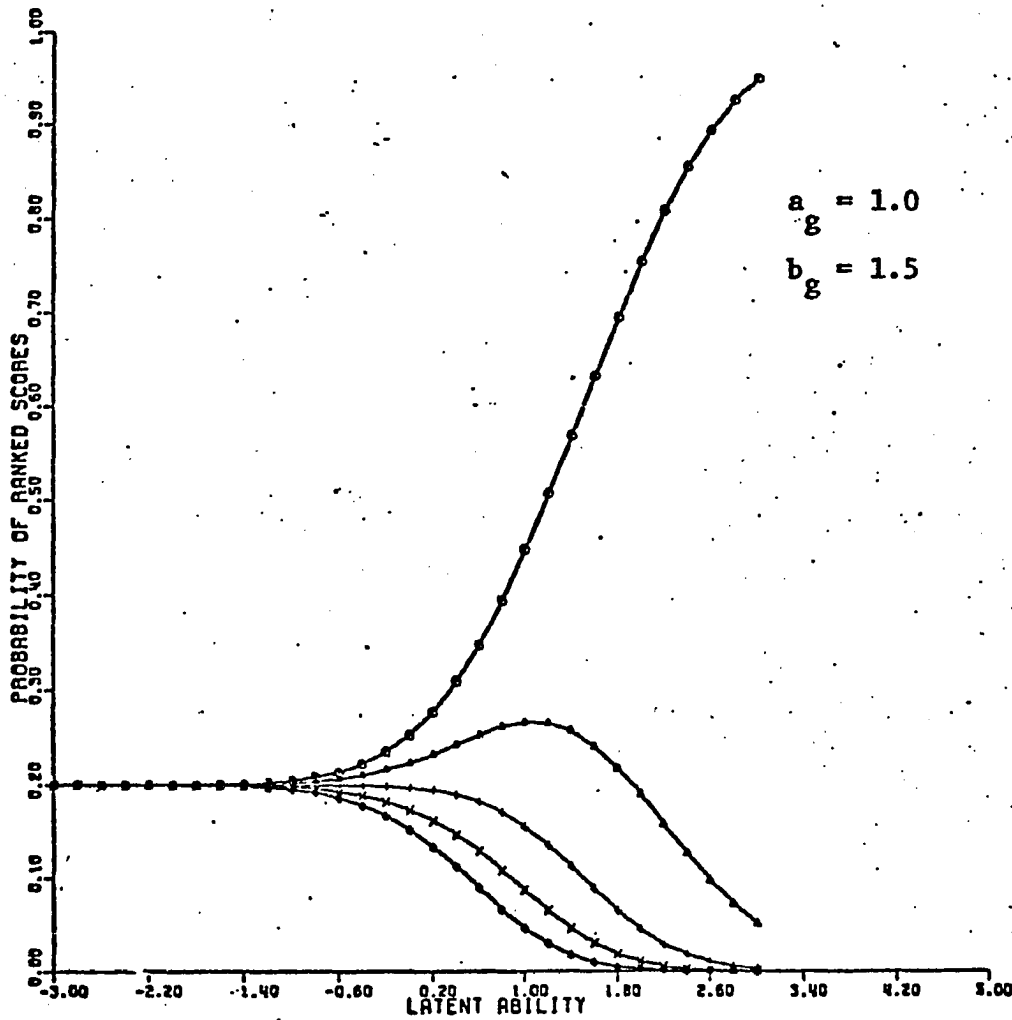



Figure 8. Plot of Conditional Distributions of Rank-Order Score for Item 16

TABLE 9

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 0.465 | 0.794 |
| 2) | -2.8 | 0.465 | 0.794 |
| 3) | -2.6 | 0.465 | 0.794 |
| 4) | -2.4 | 0.465 | 0.794 |
| 5) | -2.2 | 0.465 | 0.794 |
| 6) | -2.0 | 0.466 | 0.794 |
| 7) | -1.8 | 0.466 | 0.794 |
| 8) | -1.6 | 0.467 | 0.794 |
| 9) | -1.4 | 0.468 | 0.794 |
| 10) | -1.2 | 0.473 | 0.794 |
| 11) | -1.0 | 0.482 | 0.793 |
| 12) | -0.8 | 0.500 | 0.791 |
| 13) | -0.6 | 0.530 | 0.783 |
| 14) | -0.4 | 0.574 | 0.763 |
| 15) | -0.2 | 0.630 | 0.717 |
| 16) | 0.0 | 0.686 | 0.634 |
| 17) | 0.2 | 0.723 | 0.518 |
| 18) | 0.4 | 0.719 | 0.386 |
| 19) | 0.6 | 0.664 | 0.264 |
| 20) | 0.8 | 0.561 | 0.167 |
| 21) | 1.0 | 0.432 | 0.101 |
| 22) | 1.2 | 0.302 | 0.058 |
| 23) | 1.4 | 0.191 | 0.032 |
| 24) | 1.6 | 0.111 | 0.017 |
| 25) | 1.8 | 0.058 | 0.008 |
| 26) | 2.0 | 0.028 | 0.004 |
| 27) | 2.2 | 0.012 | 0.002 |
| 28) | 2.4 | 0.005 | 0.001 |
| 29) | 2.6 | 0.002 | 0.000 |
| 30) | 2.8 | 0.001 | 0.000 |
| 31) | 3.0 | 0.000 | 0.000 |

† $a_g = 1.0$, $b_g = 1.5$

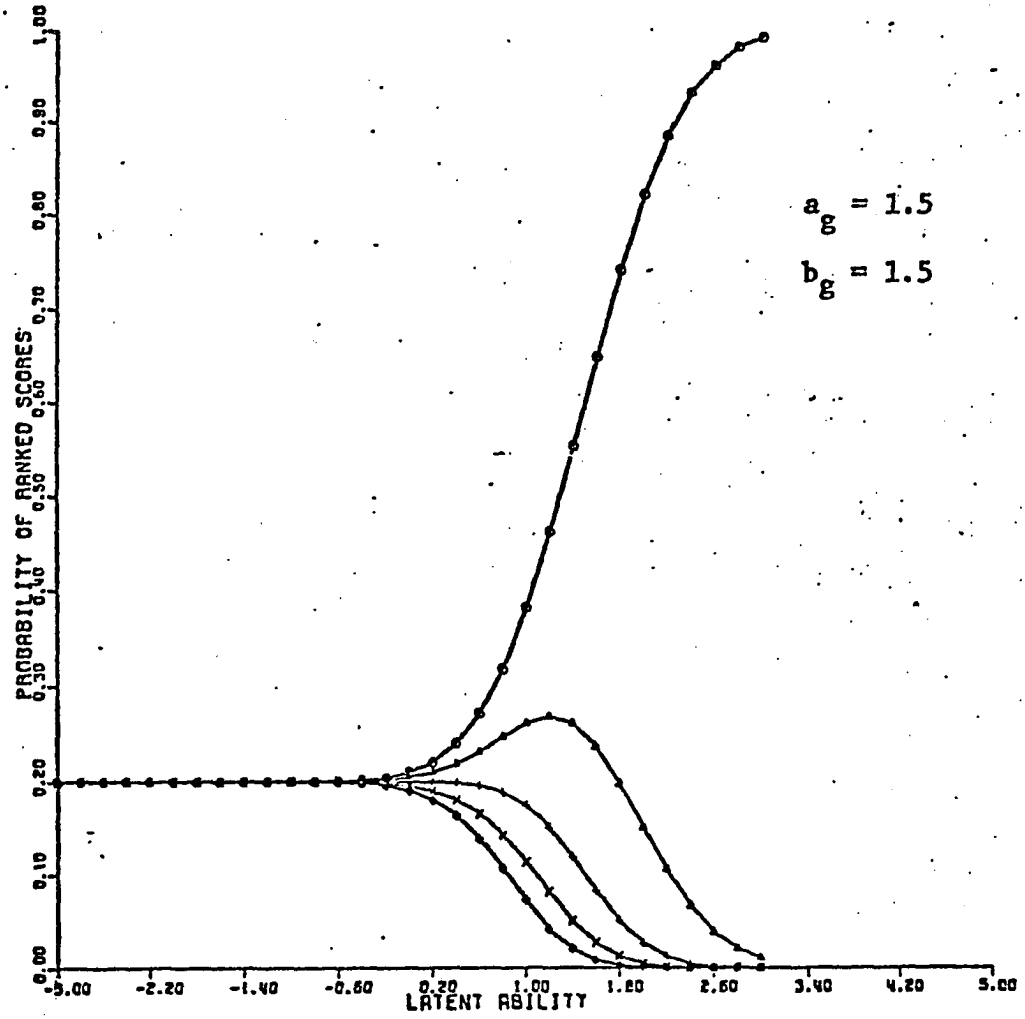



Figure 9. Plot of Conditional Distributions of Rank-Order Score for Item 25

TABLE 10

CONDITIONAL ERROR VARIANCES †

| | ABILITY | BINARY | RANKED |
|---|---|---|---|
| 1) | -3.0 | 0.693 | 0.859 |
| 2) | -2.8 | 0.693 | 0.859 |
| 3) | -2.6 | 0.693 | 0.859 |
| 4) | -2.4 | 0.693 | 0.859 |
| 5) | -2.2 | 0.693 | 0.859 |
| 6) | -2.0 | 0.693 | 0.859 |
| 7) | -1.8 | 0.693 | 0.859 |
| 8) | -1.6 | 0.693 | 0.859 |
| 9) | -1.4 | 0.693 | 0.859 |
| 10) | -1.2 | 0.693 | 0.859 |
| 11) | -1.0 | 0.693 | 0.859 |
| 12) | -0.8 | 0.693 | 0.859 |
| 13) | -0.6 | 0.694 | 0.859 |
| 14) | -0.4 | 0.697 | 0.859 |
| 15) | -0.2 | 0.704 | 0.859 |
| 16) | 0.0 | 0.718 | 0.858 |
| 17) | 0.2 | 0.744 | 0.856 |
| 18) | 0.4 | 0.789 | 0.847 |
| 19) | 0.6 | 0.855 | 0.825 |
| 20) | 0.8 | 0.938 | 0.775 |
| 21) | 1.0 | 1.021 | 0.686 |
| 22) | 1.2 | 1.076 | 0.561 |
| 23) | 1.4 | 1.071 | 0.418 |
| 24) | 1.6 | 0.988 | 0.285 |
| 25) | 1.8 | 0.835 | 0.181 |
| 26) | 2.0 | 0.643 | 0.109 |
| 27) | 2.2 | 0.449 | 0.063 |
| 28) | 2.4 | 0.285 | 0.035 |
| 29) | 2.6 | 0.165 | 0.018 |
| 30) | 2.8 | 0.087 | 0.009 |
| 31) | 3.0 | 0.042 | 0.004 |

† $a_g = 1.5$, $b_g = 1.5$

# CHAPTER V

# DISCUSSION

Despite the fact that formula scoring in partial knowledge studies has been characterized by a long history of disappointing results, it is obvious that response methods presently used in paper and pencil testing probably extract only a very small fraction of the information potentially available from each question. The amount of residual information which can in fact be recovered by introducing a more refined response method has been the subject of this study.

## Reasonableness of Assumptions Underlying the Models

It is clear that if an examinee's marks on an answer sheet are viewed without any assumptions at all, the amount of knowledge he may possess can not be estimated. The assumptions of these two models have been chosen so that the scoring formulas will depend upon a set of parameters for which consistent estimates may be found. The basic assumptions of the binary model were listed in Chapter III. Those numbered 1 and 2 have been reviewed in great detail by Lord (1952, 1953, 1968) and lead to equation (3.1). Assumption 3 was introduced to eliminate the possibility of omission which is not the subject of concern in this study. The assumption that the rank placed on the correct alternative can be used as an index of the partial knowledge possessed by an examinee becomes a device constructed to make it profitable for examinees to respond to test items in a specific way. The rank order procedure scores examinees according to a rule which relates the examinee's ranking decision to the examinee's beliefs about the relative correctness of each of the item alternatives.

In light of work reported by Coombs, Milholland, and Womer (1956); de Finetti (1965); Nedelsky (1954); Powell (1968) and others, the rank order responding and scoring procedure seems very reasonable, less arbitrary, and much less demanding to teach examinees.

Rank-Order Responding vs Binary Responding

The main purpose for studying a rank order scoring and responding procedure was to determine if an examinee's ability can be measured with greater precision than is possible using binary scoring. A review of the basic trends in item reliability, item validity, CEL-R, and CEL-V for the thirty-six simulated test items were described in Chapter III. It is obvious from Table 1 that rank order scoring is superior to binary scoring in specific situations only. Insight as to why this is so can be gained from a careful inspection of the plots of conditional distributions of ranked score for each of the nime sample items (Figs. 1 - 9) and their respective conditional error variances (Tables 2 - 10). For example, Figures 1, 2, and 3 are plots of conditional distributions of rank order scores for items of equal difficulty (very easy) and increasing discrimination. Each curve in the plot represents the regression of rank order score on ability for each of the ranks 1, 2,...., k. From equation (3.2) it can be seen that the top curve represents the item characteristic curve for the binary model and the probability that a rank of one is placed on the correct alternative in the rank order model. It can be seen that this curve provides differential information about the probability that an examinee with ability $\Theta_i$ will rank the correct

alternative one. Further, the curve functions over the entire range of examinee ability ( $-3.0\sigma$ to $3.0\sigma$ ). The second curve represents the probability that an examinee with ability $\theta_i$ will place a rank of two on the correct alternative. This curve functions from $-3.0\sigma$ below the mean ability to $2.7\sigma$ above the mean ability providing differential information about the probability of rank order scores. Curves representing the probability of ranking the correct alternative 3, 4, and 5 (the 3rd, 4th, and 5th curve respectively) indicate additional information about the probability of rank order score although the range of examinee ability over which these curves function becomes smaller as the rank increases. Only examinees of very low ability are likely to rank the correct alternative 3, 4, or 5.

Conditional error variances for item 1 (Figure 1) are presented in Table 2. For low examinee ability ( $-3.0\sigma$ to $-1.2\sigma$ ) rank order error variances are much higher than the corresponding binary variance. This indicates that the rank order scoring system is not discriminating very well among examinees of low ability. Rank order variance is larger because of the noise introduced by guessing at these low abilities. It is not surprising to find a CEL-R of 1.82 and a CEL-V of only 1.09.

In Figure 2 (item 10) we find a plot of an item of equal difficulty to item 1 but a higher item discrimination. The effect of increasing the item discrimination is to increase the slope of all the curves. The lower asymptote of each curve is nearer 0.2 indicating more random response for examinees of very low ability. The binary item (top curve) is becoming more discriminating over a narrower range of

examinee ability. Curves 2, 3, 4, and 5 are functioning over smaller ranges of examinee ability than they did in item 1 (Figure 1). The conditional error variances for this item are presented in Table 3. Again it is found that the rank order procedure is not effective at low abilities while almost equal precision results for examinees of high ability as the item becomes more discriminating. Table 1 indicates that item reliability increases as item discrimination increases. The increase in item reliability, however, is much greater for the binary scored item than it is for the rank order scored item. Thus as item discrimination increases, at a fixed difficulty, the CEL-R decreases. This is true for CEL-V also. This pattern becomes even more pronounced in Figure 3. Here the item difficulty remains the same but the item discrimination is increased still further. The slope of P(R1) becomes almost vertical and the range of examinee ability over which each of the curves functions become smaller. Table 4 records the conditional error variances for this item (item 19). For examinees below -1.8 the rank order procedure is not effective while for examinees above 0.6 either scoring system will do. Table 1 shows CEL-R and CEL-V to decrease.

Item sets (5, 14, 23 & 7, 16, 25) have different difficulties (0.5 & 1.5) but have equally increasing discriminations. The effect of increasing the item difficulty is to shift the curves to the right side of the plot although the pattern within each set of items is the same as that described above. Thus if item difficulty is held constant and item discrimination is increased, the range of examinee

ability over which the test item functions becomes narrower and concentrated around the point $\theta = b_g$. For a binary item, increasing item discrimination to infinity would yield a vertical slope for $P(R_1)$ resulting in an item with perfect reliability and no validity. Examinees below the point $\theta = b_g$ would miss the item (would place ranks at random) and examinees above $\theta = b_g$ would would be getting a perfect score (placing the rank of one on the correct alternative). Rank order scoring would not be expected to result in improvement because it would have low precision below $\theta = b_g$ and equal precision above $\theta = b_g$. CEL-R $\leq 1.0$ and CEL-V $\leq 1.0$ would be expected with items of this type. It is easily seen why gains in reliability and validity would not result from the rank order scoring of test items with high discrimination indices.

High item discrimination at a fixed item difficulty is one of the few situations in which rank order scoring is not superior to binary scoring. This occurs when $a_g$ exceeds unity ( see Table 1). However, for items found in practice, values of $a_g$ exceeding unity are rare (Lord, 1968). Thus, items found in practice have moderate to high difficulty and moderate discrimination ($a_g \leq 1.0$). An inspection of Table 1 reveals substantial gains in reliability and validity are had when items with these characteristics are scored using the rank order procedure. It must be realized that the greatest improvement in rank order scoring over binary scoring will be found for examinees of moderate to high ability.

## Other Problems

Further research attention might be directed toward estimating reliability and validity within a truncated range of examinee ability. This would provide clearer pictures of the effectiveness of rank order scoring for examinees of specified abilities and more precise information about how and where testing could be benefited. In addition, the item information structure proposed by Lord (1968) should be used as an alternative in evaluating an item's effectiveness. Such research would provide estimates of the information content of item alternatives. This type of knowledge would be helpful in item construction and diagnostic feedback to the instructor and examinee.

It should be noted that what has been proposed and simulated in this study is a procedure for scoring individual test items which utilizes ranking. No rationale has been provided for the combination of test items into a total test. There has been no suggestion that scoring items so combined using the rank order procedure would result in gains in reliability and validity over binary scoring. This would certainly be an important question to be answered by future research. Other questions regarding cost in time, in effort, and money necessary to obtain partial knowledge must be evaluated within the empirical framework.

## Conclusion

This study provides evidence that the main arguments for and against the use of rank order scoring are not to be found in group statistics but rather in the undesirable effects of one kind of

of scoring procedure or another for certain examinees. It has been demonstrated that for examinees of moderate to high ability, substantial gains in reliability and validity may result from the rank order scoring of items of moderate discrimination and varying difficulty. Items commonly found in practice in aptitude and achievement testing possess these characteristics. Despite the problems noted above, the rank order scoring model does present a promising line of investigation for studying and extracting partial examinee knowledge in multiple-choice testing.

# BIBLIOGRAPHY

## BIBLIOGRAPHY


Bayuk, R. J. The effects of choice weights and item weights on reliability and predictive validity of aptitude tests. Philadelphia, Pennsylvania University, 1973. (ERIC: ED 078061)

Calandra, A. Scoring formulas and probability considerations. Psychometrika, 1941, 6, 1 - 9.

Chernoff, H. The scoring of multiple-choice questionnaires. Annals of Mathematical Statistics, 1962, 33, 375 - 393.

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

Davis, F. B. Use of correction for chance success in test scoring. Journal of Educational Research, 1958, 52, 279 - 280.

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

Davis, F. B. Estimation and use of scoring weights for each choice in multiple-choice test items. Educational and Psychological Measurement, 1959, 19, 291 - 298. (b)

De Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 18, 87 - 123.

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

Ebel, R. L. Measuring Educational Achievement. N.J.: Prentice-Hall, Inc., 1965. (a)

Ebel, R. L. Confidence weighting and test reliability. Journal of Educational Measurement, 1965, 2, 150 - 153. (b)

Ebel, R. L. Valid confidence testing-demonstration kit. Journal of Educational Measurement, 1968, 5, 353 - 354.

Echternacht, G. The use of confidence testing in objective tests. Review of Educational Research, 1972, 2, 217 - 236. (a)

Echternacht, G. Personality influences on confidence test scores. Journal of Educational Measurement, 1972, 3, 235 - 241. (b)

Finney, D. J. The application of probit analysis to the results of mental tests. Psychometrika, 1944, 19, 31 - 39.

Frary, R. B. The reliability of a multiple-choice test is not the proportion of variance which is true variance. Educational and Psychological Measurement, 1969, 29, 359 - 365. (a)

Frary, R. B. Elimination of the guessing component of multiple-choice test scores: Effects on reliability and validity. Educational and Psychological Measurement, 1969, 29, 665 - 680. (b)

Garvin, A. D. Confidence weighting. A paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972). (ERIC: ED 062401)

Grier, J. B. & Ditrichs, R. The estimation of knowledge by multiple-choice tests. The American Statistian, 1968, 22, 35 - 36.

Guilford, J. P., Lovell, C., & Williams, R. M. Completely weighted versus unweighted scoring in achievement exams. Educational and Psychological Measurement, 1942, 2, 15 - 18.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

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

Hansen, R. The influence of variables other than knowledge on probabilistic test. Journal of Educational Measurement, 1971, 8, 9 - 14.

Hendrickson, G. E. The effect of differential option wieghting on multiple-choice objective tests. Baltimore, Maryland, Johns Hopkins University, 1971 (ERIC: ED 050168)

Hevner, K. A. A method of correcting for guessing in true-false tests and empirical evidence in support of it. Journal of Social Psychology, 1932, 3, 359 - 362.

Jacobs, P. I. & Vandeventer, M. Information in wrong responses. Research Bulletin 68 - 25, Princeton, N.J.: Educational Testing Service, 1968.

Kelly, T. L. Scoring of alternative responses with reference to some criterion. Journal of Educational Psychology, 1934, 25, 504 - 510.

Kogan, N. & Wallach, M. A. Risk-taking: A Study In Cognition and Personality. New York: Holt, Rinehart and Winston, 1964.

Kuder, G. F. A comparative study of some methods of developing occupational keys. Educational and Psychological Measurement, 1957, 17, 105 - 114.

Lawley, D. N. On problems with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61, 273 - 278.

Lazarsfeld, P. F. Chapters 10 and 11 in S. A. Stouffer et al. (Eds.) Measurement and Prediction. Princeton, N.J.: Princeton University Press, 1950. (a)

Lazarsfeld, P. F. Latent structure analysis. In S. Koch (Ed.). Psychology: A Study of a Science. Vol. 3, New York: McGraw-Hill, 1959, 476 - 542. (b)

Loevinger, J. The attenuation paradox in test theory. Psychological Bulletin, 1954, 51, 493 - 504.

Lord, F. M. A theory of test scores. Psychometrika Monograph, 1952, No. 7.

Lord, F. M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517 - 548.

Lord, F. M. Formula scoring and test validity. Educational and Psychological Measurement, 1963, 23, 663 - 672.

Lord, F. M. The effect of random guessing on test validity. Educational and Psychological Measurement, 1964, 24, 745 - 747.

Lord, F. M. & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistics model. Educational and Psychological Measurement, 1968, 28, 989 - 1020.

Merwin, J. C. Rational and mathematical relationship of six scoring procedures applicable to three choice items. Journal of Educational Psychology, 1959, 50, 153 - 161.

Michael, J. J. The reliability of a multiple-choice examination under various test-taking instructions. Journal of Educational Measurement, 1968, 5, 332 - 337.

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

Powell, J. C. The interpretation of wrong answers from a multiple-choice test. Educational and Psychological Measurement, 1968, 28, 403 - 412.

Richardson, M. W. The combination of measures. Pages 379 - 401 in P. Horse (Ed.). The Prediction of Personal Adjustment. New York: Social Science Research Council, 1941.

Rippey. R. M. Scoring and analyzing confidence tests. Chicago, Illinois University, 1971. (ERIC: ED 054236)

Roby, T. B. Belief States: A Preliminary Empirical Study. ESD-TDR-64-238. Bedford, Mass.: Decision Sciences Laboratory, 1965.

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

Shuford, E. H., Albert, A. & Massengill, H. E. Admissable probability measurement procedures. Psychometrika, 1966, 31, 125 - 145.

Slakter, M. J. Risk taking behavior on objective examinations. Journal of the American Educational Research Association, 1967, 4, 31 - 43.

Soderquist, H. O. A new method of weighting scores in a true-false test. Journal of Educational Research, 1936, 30, 290 - 292.

Stanley, J. C. & Wang, M. D. Weighting test items and test-item options, an overview of the analytical and empirical literature. Educational and Psychological Measurement, 1970, 30, 21 - 35.

Staffelback, E. H. Weighting responses in true-false examinations. Journal of Educational Psychology, 1930, 21, 136 - 139.

Strong, E. K. Vocational Interests of Men and Women. Stanford: Stanford University Press, 1943.

Swineford, F. The measurement of a personality trait. Journal of Educational Psychology, 1938, 29, 289 - 292.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 29, 438 - 444.

Toda, M. Measurement of Subjective Probability Distributions. ESD-TRD-63-407. Bedford, Mass. Decision Sciences Laboratory, 1963.

Wiley, L. N. & Trimble, O. C. The ordinary objective test as a possible criterion of certain personality traits. School and Society, 1936, 43, 446 - 448.

Ziller, R. C. A measure of the gambling response set in objective tests. Psychometrika, 1957, 22, 289 - 292.

# APPENDIX A

```
      REAL*4            T(1001),TS(1001),ANT(1001),W(5),X(5,1001),THETA(1
     *001),ZTSC(1001),ZQ(1001),Q(3,31)
      REAL*4            SUM,SSQ,RANK,RZERO,RTV,RTEV,ZTS,ZTSSQ,RTS,RTSSQ,
     *RTES,RTESSQ,ZTT,Y,Z,UP,DL,UP1,DL1,D
      DIMENSION AA(10),BB(10)
C... THE VECTOR T HOLDS THE Z-SCORE VALUES OF THETA
C... THE VECTOR TS HOLDS THE TRUE SCORES FOR THE RANK-ORDER MODEL
C... THE VECTOR INT HOLDS THE INTERVAL WIDTHS WHICH REPRESENT THE
C...  PROBABILITY OF OCCURANCE OF THETA
C... THE VECTOR W CONTAINS THE ITEM WEIGHTS
C... THE MATRIX CONTAINS A WORK AREA FOR THE COMPUTATION OF CONDITIONAL
C...  DISTRIBUTIONSAND AN AREA FOR THE COMPUTATION OF JOINT AND MARGINAL
C...  DISTRIBUTIONS.
C... THE VECTOR THETA CONTAINS THE POINTS (STANDARD DEVIATION UNITS)
C...  OF THETA OR LEVEL OF ABILITY.
C
      REWIND 2.
      KP=5
      READ(5,7777) KI,ADD
 7777 FORMAT(I5,F8.3)
      KU=KI+6
      N=(KI-1)/2
      K=N+1
      READ(5,4)(W(I),I=1,KP)
    4.FORMAT(5F10.5)
      ZQ(K)=0.0
      START= ZQ(K)
      DO 99999 I=1,N
      START=START+ADD
      ZQ(K-I)=-1.0*START
      ZQ(K+I) = START
99999 CONTINUE
      READ(5,76) (AA(I),I=1,4)
      READ(5,76) (BB(I),I=1,9)
   76 FORMAT(10F3.1)
      DO 2000 IJ=1,4
      DO 2000 II=1,9
      A=AA(IJ)
      B=BB(II)
C
C... KP = NUMBER OF ITEM RESPONSES
C... KI = THE NUMBER OF ABILITY POINTS
C... KJ = KP + KI + 1  AND IS USED AS A DIMENTION OF THE MATRIX X
C
C... ZERO OUT
C
      SUM=0.
      SSQ=0.
      RANK=0.
      RZERO=0.
      RTV=0.
      RTEV=0.


      ZTS=0.
      ZTSSQ=0.
      RTS=0.
      RTSSQ=0.
      RTES=0.
      RTESSQ=0.
C
C
C..ZERO TRUE (RANK) SCORE VECTOR
C
      DO 100 J=1,KI
      ZTSC(J)=0.
  100 TS(J)=0.
C
```

```
C... PREFORM ITEM PARAMETER MANIPULATIONS AND TRANSFORM TO Z-SCORE
C
      DO 888 I=1,KI
  888 THETA(I)=A*(ZQ(I)-B)
      DO 900 I=1,KI
      Y=THETA(I)
      CALL NTD (Y,Z,D)
      T(I)=Z
  900 CONTINUE
C
C...CALCULATE THE INTERVAL WIDTH
C
      SUM=0.
      C=(ZQ(1)-ZQ(2))/2.
      DO 901 J=1,KI
      UP=ZQ(J)+C
      DL=ZQ(J)-C
      CALL NTD(UP,UP1,D)
      CALL NTD(DL,DL1,D)
      ANT(J)= ABS(UP1-DL1)
  901 CONTINUE
C
C...CALCULATE CONDITIONAL DISTRIBUTIONS
C
      DO 101 J=1,KI
      CP=0.
      DO 102 I=1,KP
      X(I,1)=T(J)
      THET =(1.-CP)*T(J)
      XK=1.-CP
      X(I,2)=T(J)*XK
      CX=CP+X(I,2)
      XG=1.-CX
      CGUESS=1./(KP-II-1)
      X(I,3)=XG*CGUESS
      X(I,4)=X(I,2)+X(I,3)
      X(I,J+5)=X(I,4)
      X(I,5)=CP+X(I,4)

      CP=X(I,5)
  102 CONTINUE
  101 CONTINUE
C
C
C...COMPUTE SUM & SSQ FOR ONE-ZERO TRUE SCORE VARIANCE
C...COMPUTE SUM & SSQ FOR RANK-ORDER TRUE SCORE VARIANCE
C
      KZ=KU-1
      I=0
      DO 104 J=6,KZ
      I=I+1
      ZTS=ZTS+X(1,J)*ANT(I)
      ZTSSQ=ZTSSQ+X(1,J)*X(1,J)*ANT(I)
      DO 104 K=1,KP
      TS(I)=TS(I)+X(K,J)*W(K)
  104 CONTINUE
C... COMPUTE STD OF MEASUREMENT
      DO 501 J=1,31
  501 Q(1,J  )= X(1,J+5) - X(1,J+5)*X(1,J+5)
C
C
      DO 503 I=1,31
      Q(2,I  )=0.
      Q(3,I  )=0.
      DO 504 J=1,5
      Q(2,I  )=Q(2,I  )+X(J,I+5)*W(J)
      Q(3,I  )=Q(3,I  )+X(J,I+5) * W(J)*W(J)
  504 CONTINUE
      Q(3,I)=Q(3,I)-Q(2,I)*Q(2,I)
  503 CONTINUE
C
```

```
C...COMPUTE JOINT DISTRIBUTION
C
      DO 203 J=1,KI
      DO 203 I=1,KP
  203 X(I,J+5)=X(I,J+5)*ANT(J)
C...SUM ROWS OF JOINT DISTRIBUTION TO OBTAIN MARGINAL
      DO 204 I=1,KP
      X(I,KU)=0.
      DO 204 J=6,KZ
  204 X(I,KU)=X(I,KU)+X(I,J)
C
C
C...COMPUTE SUM & SSQ FOR RANK-ORDER TOTAL TEST VARIANCE
C...COMPUTE SUM & SSQ FOR RANK-ORDER TTRUE SCORE VARIANCE
C
      DO 233 I=1,KP
      RTES=RTES+W(I)*X(I,KU)
  233 RTESSQ=RTESSQ+W(I)*W(I)*X(I,KU)
      DO 207 I=1,KI
      RTS=RTS+ANT(I)*TS(I)
```

---

```
  207 RTSSQ=RTSSQ+TS(I)*TS(I)*ANT(I)
C
C...COMPUTE VARIANCES
      ZTT=X(1,KU)*(1.-X(1,KU))
      ZTRUE=ZTSSQ-ZTS*ZTS
      RTEV=RTESSQ-(RTES*RTES)
      RTV=RTSSQ-(RTS*RTS)
C...COMPUTE RELIABILITIES
      RANK=RTV/RTEV
      RZERO=ZTRUE/ZTT
C
C...  CONVERT STD OF MEASUREMENT TO PROPORTIONS
C
      DO 5511 I=1,31
      Q(1,I)=Q(1,I)/ZTT*(1.0-RZERO)
      Q(3,I)=Q(3,I)/RTEV *(1.0-RANK)
 5511 CONTINUE
      WRITE(6,502) A,B,(Q(1,J),J=1,31)
      WRITE(6,502) A,B,(Q(3,I),I=1,31)
  502 FORMAT(' DISC = ',F4.1,' DIFF = ',F4.1/10F8.4/10F8.4/10F8.4/F8.4)
C COMPUTE COEFFICIENT OF EFFECTIVE LENGTH
      RC=RANK*(1.-RZERO)
      RCC=RZERO*(1.-RZERO)
      RC=RC/RCC
C
C... COMPUTE VALIDITY
C
      SUM=0.
      DO 206 J=1,KI
  206 SUM=SUM+X(1,J+5)*ZQ(J)
      ZVAL=SUM/ SQRT(ZTT)
      SUM=0.
      DO 205 J=1,KI
      DO 205 I=1,5
  205 SUM=SUM+W(I)*X(I,J+5)*ZQ(J)
      VAL=SUM/ SQRT(RTEV)
      ZV=VAL*VAL*(1.-RZERO)
      RV=ZVAL*ZVAL-VAL*VAL*RZERO
      ZV=ZV/RV
      WRITE       (2) A,B,ZTRUE,ZTT,RZERO,ZVAL,RTV,RTEV,RANK,VAL,RC,ZV
      WRITE(6,681) A,B
  681 FORMAT(///' ITEM DISCRIMINATION INDEX = ',F10.5/
     *' ITEM DIFFICULTY INDEX      = ',F10.5//)
      WRITE(6,15) ZTRUE,ZTT,RZERO,ZVAL,RTV,RTEV,RANK,VAL,RC,ZV
```

```
   15 FORMAT(///' ZERO-ONE SCORING SYSTEM'/
     *'      TRUE SCORE VARIANCE = ',F10.5/
     *'      TOTAL TEST VARIANCE = ',F10.5/
     *'      RELIABILITY FOR ONE ITEM = ',F10.8/
     *'      VALIDITY FOR ONE ITEM = ',F10.5///
     *' RANK-ORDER SCORING SYSTEM'/
     *'      TRUE SCORE VARIANCE = ',F10.5/
     *'      TOTAL TEST VARIANCE = ',F10.5/


     *'      RELIABILITY FOR ONE ITEM = ',F10.8/
     *'      VALIDITY FOR ONE ITEM = ',F10.5/
     *' CEL FOR RELIABILITY = ',F10.5/
     *' CEL FOR VALIDITY = ',F10.5///)
 2000 CONTINUE
C
      REWIND 2
      WRITE(6,666)
  666 FORMAT(15X,'   ITEM        BINARY      RANKED'/
     115X,'PARAMETERS      ITEM          ITEM    IMPROVEMENT'/
     215X,'A(G)  B(G)  REL   VAL   REL   VAL  CEL-R CEL-V')
      DO 661 KK=1,36
      READ       (2) A,B,ZTRUE,ZTT,RZERO,ZVAL,RTV,RTEV,RANK,VAL,RC,ZV
      WRITE(6,662) KK,A,B,RZERO,ZVAL,RANK,VAL,RC,ZV
  662 FORMAT(10X,I2,')',8(3X,F4.2))
  661 CONTINUE
      WRITE(6,664)
  664 FORMAT('1')
      STOP
      END
```

```
      SUBROUTINE ZPLOT(X,ZQ,KI,KP)
      DIMENSION X(5,1001)
      DIMENSION XC(1001)
      DIMENSION ZQ(1001),XX(4)
C
C... SET PEN 3 IN. FROM RIGHT
C
      CALL PLOT(0.0,-29.5,-3)
      CALL PLOT(0.0,3.0,-3)
C
C... SET MAX. AND MIN. VALUES FOR X
C
      XX(1)=0.0
      XX(2)=1.0
C
C... SCALE X
C
      CALL SCALE(XX,10.0,2,1)
C
C... SET SCALED MIN. = START & MAX. = DEL
C
      START=XX(3)
      DEL=XX(4)
      XC(KI+1)=START
      XC(KI+2)=DEL
C.
C... CALCULATE X-AXIS
C
      CALL AXIS(0.0,0.0,'TRUE SCORE',10,10.0,90.0,START,DEL)
C
C... SCALE AND SET Y-XIS
C
      XX(1)=-3.0
      XX(2)=3.0
      CALL SCALE(XX,10.0,2,1)
      ZQ(KI+1)=XX(3)
      ZQ(KI+2)=XX(4)
      CALL AXIS(0.0,0.0,'LATENT ABILITY',-14,10.0,0.0,ZQ(KI+1),ZQ(KI+2))
C
C... PLOT LINES
C
      DO 100 I=1,KP
      DO 10 J=1,KI
   10 XC(J)=X(I,J+5)
      IS=I
      CALL PLOT(ZQ(1),XC(1),3)
      CALL LINE(ZQ,XC,KI,1, 1,IS)
  100 CONTINUE
      RETURN
      END
```

---

```
      SUBROUTINE NTD(X,P,D)
      REAL*4           AX,T,D,P,X
      AX=  ABS(X)
      T = 1.0/(1.0 + 0.2316419 * AX)
      D = 0.3989423 * EXP(-X * X/2.0)
      P = 1.0-D*T*(((((1.330274*T -1.821256)*T + 1.781478)*T - 0.3565638)
     **T + 0.3193815)
      IF (X) 1,2,2
    1 P=1.0-P
    2 RETURN
      END
```