A COMPARISON OF THREE TYPES OF ITEM ANALYSIS IN TEST DEVELOPMENT USING CLASSICAL AND LATENT TRAIT METHODS

By

IRIS G. BENSON

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

UNIVERSITY OF FLORIDA

1977

ACKNOWLEDGMENTS

I am deeply indebted to two special people who have greatly influenced by graduate education, Dr. William Ware, chairman of my guidance committee, and Dr. Linda Crocker, unofficial cochairman of my committee. Their continued encouragement and support has resulted in my reaching this point in my graduate studies. I shall always be extremely grateful to Dr. Ware and Dr. Crocker for whatever skills I have developed as a researcher and as a teacher are in large part due to their advice and guidance. To them I owe the high value I place on objective, quantitative research methods. Further, I would like to acknowledge the tremendous amount of time they spent in molding the final copy of this manuscript.

I would also like to express my appreciation to the members of my committee, Dean John Newell and Dr. William Powell, for their suggestions and editorial comments on this dissertation. Special thanks are extended to Dr. Wilson Guertin for his assistance with portions of the study, and as an unofficial member of my committee.

I would like to thank Dr. Jeaninne Webb, Director of the Office of Instructional Resources, and Mr. Robert Feinberg and Ms. Arlene Barry of the Testing Division, for providing the data used in this study.

Finally, I would like to express my sincere appreciation to my friends and family who stood by me during very trying times in my graduate education.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF TABLES - Continued

LIST OF FIGURES

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

A COMPARISON OF THREE TYPES OF ITEM ANALYSIS IN TEST DEVELOPMENT USING CLASSICAL AND LATENT TRAIT METHODS

By

Iris G. Benson

December 1977

Chairman: William B. Ware
Major Department: Foundations of Education

Test reliability and validity are determined by the quality of the items in the tests. Through the application of item analysis procedures, test constructors are able to obtain quantitative, objective information useful in developing and judging the quality of a test and its items.

Classical test theory forms the basis for one method of test development. An integral part of the development of tests based on the classical model is selection of a final set of items from an item pool based on classical item analysis or factor analysis. Classical item analysis requires identification of single items which provide maximum discrimination between individuals on the latent trait being measured. The biserial correlation between item score and total score is commonly used as an index of item discrimination.

An alternative method of test development, but based on the classical model, is factor analysis. Factor analysis is a more complex test development procedure than classical item analysis. It is a

statistical technique that takes into account the item correlation with all other individual items in the test simultaneously. Thus, classical item analysis can be viewed as a unidimensional basis for item analysis, less sophisticated than the multidimensional procedure of factor analysis.

Recently, the field of latent trait theory has provided a new approach to test construction. Several latent trait models have been developed; however, this study was concerned only with the one-parameter logistic Rasch model. The Rasch model was chosen because it is the most parsimonious of the latent trait models and has recently been used in the development and equating of tests.

A review of the literature revealed numerous studies conducted in each of the three areas of item analysis, but no comparative studies were reported among all three item analytic techniques. Therefore, the present study was designed to compare the methods of classical item analysis, factor analysis, and the Rasch model in terms of test precision and relative efficiency.

An empirical study was designed to compare the effects of the three methods of item analysis on test development across different sample sizes of 250, 500, and 995 subjects. Item response data were obtained from a sample of 5,235 high school seniors on a 50 item cognitive test of verbal aptitude. The subjects were divided into nine independent samples, one for each item analytic technique and sample size. The study was conducted in three phases: item selection, computation of item and test statistics for selected items on double cross-validation samples, and statistical analyses of item characteristics. For each item analytic procedure two tests were developed:

a 15 item test, and a 30 item test. Four dependent variables were obtained for each test to assess precision: internal consistency estimates, standard error of measurement, item difficulties, and item discriminations. In addition, the relative efficiencies of the 30 item tests developed by each item analytic technique were compared for the sample of 995 subjects.

The results of the analysis revealed that there were no differences between the tests developed by the three methods of item analysis, in terms of the precision of measurement. In terms of efficiency, substantive differences between the tests produced by the three item analytic methods were observed. Specifically, the tests based on classical test theory were more effective for measuring very low and very high ability students. The Rasch developed test was more efficient for assessing average and high ability students.

CHAPTER 1

INTRODUCTION

The systematic approach to test development was initiated by Binet and Simon in 1916. Since that time psychometricians have been concerned with the extent to which accurate measurement of a person's "ability" is possible. Most measurement experts agree that upon repeated testing an individual's observed score will vary even though his true ability remains constant. This variability is the essence of classical test theory.

Classical test theory is based upon the assumption that a person's observed score (X) is made up of a true score (T) and error score (E) denoted:

$$X = T + E. \qquad (1)$$

Limited by few assumptions, this theory has wide applications. The few assumptions pertain to the error score (Magnusson, 1966, p. 64):

1. The mean of an examinee's error scores on an infinite number of parallel tests is zero.

2. The correlation between examinee's error scores on parallel tests is zero.

3. The correlation between examinees' error scores and true scores is zero.

Relying upon these assumptions, psychometricians have used the observed score (X) to represent the best estimate of a person's true score (T).

The accuracy of the observed score (X) in representing an examinee's true score (T) is described by the reliability coefficient. One definition of reliability is given by the coefficient of precision. This coefficient is the correlation between truly parallel tests, assuming the examinee's true score does not change between two measurements. Lord and Novick (1968) have defined truly parallel tests to be those for which, "the expected values [true scores] of parallel measurements are equal; and the observed score variances of parallel measurements are equal (p. 48)."

The reliability coefficient for the population is defined as (Lord and Novick, 1968, p. 134):

$$r_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}, \qquad (2)$$

where σ_T^2 is the true score variance, σ_X^2 is the observed score variance, and σ_E^2 is the error score variance. When this expression is used to represent the coefficient of precision, it can be interpreted as the extent to which unreliability is due solely to inadequacies of the test form and testing procedure rather than due to changes in examinees over time.

The coefficient of precision is a theoretical value because the components σ_T^2 and σ_E^2 cannot be observed. The coefficient of precision is usually estimated by internal consistency methods. Internal consistency is a measure of the relationship between random parallel tests. Random parallel tests are composed of items drawn from the same population of items (Magnusson, 1966, p. 102-103). Scores on these tests may differ somewhat from true scores in means, standard deviations, and correlations because of random errors in the sampling of items. However, random parallel tests are more often encountered in practice than are

truly parallel tests. Cronbach's coefficient alpha (1951) is the internal consistency coefficient commonly used to represent the average correlation among all possible tests created by dividing the domain into random halves. Thus, the internal consistency coefficient indicates the extent to which all the items are measuring the same ability or trait. Psychological traits are often described as latent because they cannot be directly observed. Therefore, psychological tests are developed in an attempt to measure these latent traits.

Classical test theory forms the basis for one method of test development. An integral part of the development of tests based on the classical model is the utilization of classical item analysis or factor analysis. Classical item analysis is a procedure to obtain a description of the statistical characteristics of each item in the test. This approach requires identification of single items which provide maximum discrimination between individuals on the latent trait being measured. Theoretically, selecting items which have high correlations with total test score will result in a discriminating test which is homogeneous with respect to the latent trait. Therefore, classical item analysis is an aid to developing internally consistent tests.

An alternative method of test development, but based on the classical model, is factor analysis. Factor analysis is a more complex test development procedure than classical item analysis. It is a statistical technique that takes into account the item correlation with all other individual items in the test simultaneously. Groups of similar items tend to cluster together and comprise the latent traits (factors) underlying the test. Under the classical model then, classical item analysis can be viewed as a unidimensional basis for item analysis, less sophisiticated than the multidimensional procedure of factor analysis.

The purpose of factor analysis is to represent a variable in terms of one or several underlying factors (Harman, 1967). Depending upon the objective of the analysis, two general approaches are used in factor analysis: (a) common factor analysis, and (b) principal components analysis. A common factor solution would be warranted if the researcher were interested in determining the number of common and unique factors underlying a given test. A principal component solution would be warranted if it were of interest to extract the maximum amount of variance from a given test.

Regardless of the approach used, factor analysis is an item analytic technique in which all test items are considered simultaneously to produce a matrix of item correlations with factors. It is these correlations or item loadings that indicate the strength of the factor and also the number of factors underlying the test. However, factor analysis shares the weakness of classical item analysis, that of being sample dependent.

Critics of classical test theory contend that a major weakness of tests developed from this model is that the item statistics vary when the examinee group changes; item statistics may also vary if a different set of items from the same domain is used with the same examinee group (Hambleton and Cook, 1977; Wright, 1968). Thus, the selection of a final set of test items will be sample dependent.

Until recently, classical item analysis and factor analysis were the only techniques described in measurement texts for use in item analysis and test development (Baker, 1977). However, with the publication of Lord and Novick's Statistical Theories of Mental Test Scores (1968) and the availability of computer programs, considerable attention is being directed now toward the field of latent trait theory as a new area in

test development. Latent trait theory dates back to Lazarsfeld (1950) who introduced the concept; however, Fredrick Lord is generally given credit as the father of latent trait theory (Hambleton, Swaminathan, Cook, Eignor, and Gifford, 1977). Proponents of this approach claim that the advantages of latent trait theory over classical test theory are twofold: (a) theoretically it provides item parameters which are invariant across examinee samples which will differ with respect to the latent trait, and (b) it provides item characteristic curves that give insight into how specific items discriminate between students of varying abilities. These properties of latent trait theory will be presented in more detail in Chapter II.

Four latent trait models have been developed for use with dichotomously scored data: the normal ogive, and the one-, two-, and three-parameter logistic model (Hambleton and Cook, 1977; Lord and Novick, 1968). This study is concerned with the one-parameter logistic Rasch model because it is the simplest of the four models.

Tests developed using the Rasch model are intended to provide objective measurement of the examinee's true ability on the latent trait in question, as well as providing for invariant item parameters (Rasch, 1966; Wright, 1968). That is, any subset of items from a population of items that have been calibrated by the Rasch model should accurately measure the examinee's true ability regardless of whether the items are very easy or very difficult; also, the item parameters should remain constant over different examinees. In measurements obtained from classical test theory this objective feature is rarely attained. The item parameters associated with classical test theory are group and item specific. That is, the item parameters are determined by the

ability of the people taking the test and the subset of items chosen. Wright (1968) has stated, "The growth of science depends on the development of objective methods for transforming an observation into measurement (p. 86)." Latent trait theory is an attempt to develop mental measurement into a technique similar to measurement in the physical sciences.

Latent trait theory is based on strong assumptions that are restrictive and hence limit its application (Hambleton and Cook, 1977). The assumptions required for the Rasch model are the following (Rasch, 1966):

1. The test is unidimensional, e.g., there is only one factor or trait underlying test performance.

2. The item responses of each examinee are locally independent, e.g., success or failure on one item does not hinder other item responses.

3. The item discriminations are equal, e.g., all items load equally on the factor underlying the test.

Lord and Novick (1968) noted that the assumptions of unidimensionality and local independence are synonymous. To say that only one underlying ability is being tested means the items are statistically independent for persons at the same ability level. The third assumption relates to item characteristic curves. The item characteristic curve is a mathematical function that relates the probability of success on an item to the ability measured by the test. Curves vary in slope and intercept to reflect how items vary in discrimination and difficulty. The one-parameter logistic Rasch model (the one parameter is item difficulty) assumes all item discriminations are equal. Thus all item characteristic curves should be similar with respect to their slopes.

The Problem

Several studies have been conducted to varify the invariant properties of tests constructed using the Rasch model (Tinsley and Dawis, 1975; Whitely and Dawis, 1974; Wright, 1968). If we assume that tests developed using latent trait theory possess the quality of invariant item statistics, why then hasn't latent trait theory been more visible in the psychometric community? There appear to be three main reasons for this slow acceptance. First, the Rasch procedure is based on a mathematical model involving restrictive assumptions, e.g., the unidimensionality of the items, the local independence of the items, and equal item discriminations. A further restriction of the Rasch model is the assumption of minimal guessing. However, several researchers have demonstrated the robustness of the model with regard to departures from the basic assumptions (Anderson, Kearney and Everett, 1968; Dinero and Haertel, 1976; Rentz, 1976). Second, latent trait theory has not been used in practical testing situations because until recently there was a lack of available computer programs to handle the complex mathematical calculations. Hambleton et al. (1977) described four computer programs now available to the consumer. Third, measurement experts who are knowledgeable about latent trait models have been skeptical as to the real gains that may be available through this line of research. Are tests developed using latent trait models superior to tests developed using classical item analysis or factor analysis?

The purpose of this study was to compare the precision and efficiency of cognitive tests constructed by the three methods (classical item analysis, factor analysis and the Rasch model) from a common item and examinee population. Precision, as measured by internal consistency,

is an overall estimate of a test's homogeneity, but provides no information on how the test as a whole discriminates for the various ability groups taking the test. For that reason measures test efficiency (Lord, 1974a, 1974b) were incorporated into the study. Test efficiency provides information on the effectiveness of one test over another as a function of ability level. A cognitive college admissions subtest was used in this study for several reasons. First, tests of this type are widely used by educational institutions for a large number of examinees each year, in the areas of selection, placement, and academic counseling. Most college admission examinations traditionally have been developed using classical item analysis. Second, because of the importance of the decisions made using such test scores, it would be worth investing considerable time and expense in the development of these instruments. Thus, the use of factor analysis or the Rasch model would be justified if superiority of either of these methods over classical item analysis could be determined. Third, the items on college admission tests have been written by experts, and each subtest is intended to be unidimensional, e.g., items measuring a single ability. Thus, assumptions from all models should be met. Fourth, because of the time required to take such examinations, it is important to maximize the precision and the effectiveness of the tests. The possibility of using fewer items while maintaining precision would be desirable. Therefore, the question of which test development procedure can best accomplish this is not a trival one.

Purpose of the Study

The purpose of this study was to compare empirically the Rasch model with classical item analysis and factor analysis in test development. Five research questions guided this study.

1. Will the three methods of test development produce tests with superior internal consistency estimates when compared to the projected

internal consistency of the population as the number of items decreases?

2. Will the three methods of test development produce tests with stable estimates of internal consistency when the number of examinees decreases?

3. Will the three methods of test development produce tests with similar standard errors of measurement?[1]

4. Will the three methods of test development select items that are similar in terms of difficulty and discrimination?

5. Will the three methods of test development produce equally efficient tests for all ability levels?

Hypotheses

This study investigated the capacities of three methods of test development to increase precision and efficiency of measurement in test construction. The five questions posited in the previous section were phrased as testable hypotheses:

1. There are no significant differences in the internal consistency estimates of the tests produced by the three methods, as the number of items decreases, when compared to the projected internal consistency estimates for the population for tests of similar length.

---

[1] The standard error of measurement (SEM) is defined in the classical sense as (Magnusson, 1966, p. 79):

$$SEM = S_X \sqrt{1 - r_{xx'}}$$

where S_X is the standard deviation of the test, and r_{xx} is the reliability coefficient.

2. There are no differences in the internal consistency estimates of the tests produced by the three methods when the number of examinees is decreased.

3. There are no meaningful[2] differences in the magnitude of the standard error of measurement of the tests produced by the three methods.

4. There are no significant differences in the difficulties or discriminations of the items selected by the three methods.

5. There are no differences across ability levels in the efficiency of the tests produced by the three methods.

Significance of the Study

Objective measurement has always been assumed in the physical sciences. It has only been recently that objective measurement in the behavioral sciences has been deemed possible with the advent of latent trait theory. Since the introduction of latent trait theory by Lazarsfeld (1950) and Lord (1952a, 1953a, 1953b) much of the research on latent trait models has been confined to theoretical research journals. Wright (1968), speaking at a conference on testing problems, discussed at an applied level the need to seriously consider latent trait theory and the Rasch model in particular as a major test development technique far superior to classical item analysis and factor analysis. However, even in 1968 computer programs were not yet available to run the analyses

[2]Because test scores are usually reported and interpreted in whole numbers, a "meaningful" difference in the standard error of measurement is defined as a difference of ≥ 1.00.

should anyone beyond academicians be interested. Today this obstacle has been overcome, but many test developers remain unconvinced of the value of latent trait theory because its superiority to classical test theory has not been conclusively demonstrated. This study is an attempt to provide an empirical comparison of classical test theory and latent trait theory methods of test construction.

Of the various logistic models that represent latent trait theory the Rasch model was chosen for comparison with traditional item analysis procedures in the present study because it is the most parsimonious latent trait model and has been used recently in the development of the equating of tests (Rentz and Bashaw, 1977; Woodcock, 1974). The Rasch model provides a mathematical explanation for the outcome of an event when an examinee attempts an item on a test. Rasch (1966) stated that the outcome of an encounter is governed by the product of the ability of the examinee and the easiness of the item and nothing more. The implication of this simple concept (objectivity of measurement) would seem to revolutionize mental measurement. If invariant properties of items and ability scores can be identified and used to improve the psychometric quality of tests to an extent greater than now possible with classical and factor analytic procedures then we truly are in the age of modern test theory.

Organization of the Study

The theoretical and empirical studies related to the three methods of item analysis are described in Chapter II. An empirical investigation to compare the three methods of item analysis under varying conditions is described in Chapter III. The results of the study are reported in Chapter IV. A discussion of the results, conclusions of the study, and

implications for future research in this area have been presented in the fifth chapter. A summarization of the study has been provided in Chapter VI.

CHAPTER II

REVIEW OF THE LITERATURE

The quality of the items in a test determine its validity and reliability. Through the application of item analysis procedures, test constructors are able to obtain quantitative objective information useful in judging the quality of test items. Item analysis thus provides an empirical basis for revising the test, indicating which items can be used again and which items have to be deleted or rewritten (Lange, Lehmann, and Mehrens, 1967). Item analysis data also help settle arguments and objections to specific items that might be raised by administrators, test experts, examinees, or the public.

This study is focused on three approaches to item analysis (classical item analysis, factor analysis, and the Rasch model) as test construction techniques. It is assumed throughout this study that the test under construction is unidimensional, e.g., all items are measuring only one ability. These three approaches to item analysis and the relevant research related to each method are discussed in this chapter.

Item Analysis Procedures for the Classical Model

Item analysis as a test development technique emerged at the beginning of this century. Binet and Simon (1916) were among the first to systematically validate test items. They noted the proportion of students at particular age levels passing an item. This statistic was

measuring the relative difficulty of the items for different age groups. The item difficulty index, defined as the percentage of persons passing an item and denoted by p, is one of the statistics used in classical item analysis.

Item difficulty is related to item variance and hence to the internal consistency of the test. Test constructors are usually concerned with achieving high test reliability, e.g., precision of measurement. Therefore, an item difficulty of .50 is considered to be the ideal value necessary to maximize test reliability. This is because half the examinees are getting the item correct and half the examinees are missing the item. The proportion missing an item is defined as 1-p or q. Thus, when p is equal to .50, q is equal to .50. Because the variance of a dichotomized item is p x q the maximum variation an item can contribute to total test variance and ultimately to true-score variance is .25. As an item's difficulty index deviates from .50, its contribution to total test variance is always some value less than .25. Hence test constructors have been advised (Gulliksen, 1945) to select items with difficulty indices at or near .50. However, when items are presented in multiple choice or alternate choice format, the ideal level of difficulty is adjusted to accommodate for guessing.[3]

A second important item statistic in classical item analysis is the item discrimination index. An item discrimination index is a measure

---

[3] The ideal value of p = .50 assumes there has been no guessing on the item. The effects of guessing on item difficulty tends to increase the ideal value of p. For example, on a four option multiple choice item the chance of guessing the correct answer is (¼)(.50)=.12. The value of .12 is added to .50 to correct for the effect of guessing and the ideal p would now be .62 (Lord, 1952b; Mehrens and Lehmann, 1973).

of how well the item discriminates between persons who have high test scores and persons who have low test scores. The discrimination index is often expressed as a correlation between the item and total test score. When the criterion is total test score, the correlation coefficient indicates the contribution that item makes to the test as a whole. Thus, on tests of academic achievement it is a measure of item validity as well as a contributor to internal consistency. Noting an increasing use of item analytic procedures for the improvement of objective examinations, Richardson (1936) pointed out that the development of the procedures of item analysis had centered primarily around the invention of various indices of association between the test item and the total test score, e.g., item discrimination indices.

The two most popular item-test correlation indices are the biserial and point biserial correlations. The point biserial was developed by Pearson (1900) and is a special case of the more general Pearson Product Moment (PPM) correlation coefficient (Magnusson, 1966). This index is recommended when one of the variables being correlated (the item score) represents a true dichotomy and the other variable (total test score) is continuously distributed. Pearson (1909) also derived the biserial correlation which is an estimate of the PPM. The biserial correlation is recommended when one of the variables (the item score) has an underlying continuous and normal distribution which has been artifically dichotomized and the other variable (total test score) is continuously distributed. The assumption for the point biserial correlation is often hard to justify when it is suspected that knowledge required to answer an item is continuously distributed.

In considering the dichotomized item (pass/fail), McNemar (1962) has commented, "It is obvious that failing a test item represents anything from a dismal failure up to a near pass, whereas passing the item involves barely passing up to passing with the greatest of ease" (p. 191). Thus, the biserial correlation is usually favored over the point biserial correlation as a measure of item discrimination. Also, the biserial is often chosen over the point biserial because the magnitude of the point biserial correlation for an item is not independent of the item difficulty (Davis, 1951; Henrysson, 1971; Swineford, 1936). Specifically, values of the point biserial are systematically depressed as p approaches the extremes of .00 or 1.00. Lord and Novick (1968) have pointed out that because of this bias, the point biserial correlation tends to favor medium difficulty items over easy or very difficulty items.

The formulae for the biserial and point biserial correlation respectively are (Magnusson, 1966, p. 200 & 203):

$$r_{bis} = \frac{\bar{X}_p - \bar{X}_q}{s_y} \cdot \frac{pq}{Y} \quad , \qquad (3)$$

$$r_{pbis} = \frac{\bar{X}_p - \bar{X}_q}{s_y} \cdot \sqrt{pq} \quad , \qquad (4)$$

where $\bar{X}_p$ is the mean of y scores for persons who correctly solved the item, $\bar{X}_q$ is the mean of y scores for persons who incorrectly solved the item, s_y is the standard deviation of the y test scores, p and q have been previously defined, and Y is the ordinate of the dividing line between the proportions p and q in a unit normal distribution (Magnusson, 1966).

One of the main objectives of classical test theory is to improve the internal consistency of the test under construction where internal consistency was defined as the extent to which all items are measuring the same ability. To ensure high internal consistency the random error in the test must be minimized. As stated previously in Equation 2, reliability, in the classical model, was defined as:

$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \quad .$$

Thus, the relationship among the test items can be noted in the coefficient alpha formulae for estimating internal consistency for a sample (Magnusson, 1966, pp. 116-117):

$$r_{xx} = \frac{n}{n-1} \cdot 1 - \frac{\Sigma Si^2}{S_X^2} \quad , \qquad (5)$$

or

$$r_{xx} = \frac{n^2 \bar{C}_{ik}}{S_X^2} \quad , \qquad (6)$$

where n is the number of test items, ΣS_i^2 is the sum of the item variances, S_X^2 is the variance of the test, and $\bar{C}_{ik}$ is the mean of the item covariances. By comparing equation 2 with 5, it is seen that the sum of the unique item variances is used as an estimate of σ_E^2, and that when the unique item variation is minimized internal consistency will be high. Furthermore, the mean of the item covariances (equation 6) serves as an estimate of σ_T^2. The size of the covariance term is in turn determined by the intercorrelations and standard deviations of the items (Magnusson, 1966). Therefore, internal consistency is directly dependent upon the correlation among the items in the test.

The item discrimination index provides a measure of how well an item contributes to what the test as a whole measures. When items with the highest item-test correlations are selected, the homogeneity of the test is increased; that is, σ_t^2 is increased. So it is the item discrimination that directly affects test reliability. When items with low item-test correlations are eliminated, the remaining item inter-correlations are raised. When item-test correlations are high, the test is able to discriminate between high and low scorers and hence internal consistency is increased. If too few items are discarded in an item analysis the internal consistency of the test tends to decrease because items with little power of measuring what the entire test is intended to measure will dilute the measuring power of the efficient items (Beddell, 1950).

Research Related to Classical Item Analysis in Test Development

Several articles have been published concerning standards for item selection to maximize test validity and increase internal consistency. Flanagan (1939) stated two considerations in selecting test items: (a) the item must be valid, that is, it should discriminate between high and low scorers, and (b) the level of item difficulty should be suitable for the examinee group. Gulliksen (1945) agreed with Flanagan on these two points and added a third; items selected with $p = .50$ would produce the most valid tests; however, Gulliksen noted that current practice was opposed to selecting items with difficulty near .50. Test developers were selecting items based upon spreading difficulty indices over a broad range.

Several studies have been conducted to examine the effects of varying item difficulty on test development. Brogden (1946), in a

study of test homogeneity, has shown empirically that a test of 45 items with varying levels of item difficulty produced a reliability of .96 (measured by the Kuder-Richardson$_{20}$ formula). However, a similar but longer test of 153 items, that had item difficulties at .50 for all items, produced reliability of .99. Thus, Brogden concluded that effective item selection was based more on selecting a test with fewer items that possessed varying difficulty, than a longer test with equal item difficulty.

Davis (1951), in commenting on item difficulty, stated that if all test items had a difficulty of .50 and were uncorrelated then maximum discrimination was achieved. But when test items were correlated, maximum discrimination would only be achieved when the difficulty index for all test items was spread out, e.g., several difficult items, several easy items, and several items with difficulty near .50. Davis recommended the latter procedure for test development because test items are usually correlated to some degree. Davis also recognized the need for the approval of subject matter specialists in addition to statistical criteria in item selection.

In a study of test validity, Webster (1956) found results similar to Brogden (1946), but different from Gulliksen (1945). By selecting fewer items with high discrimination indices and varying item difficulty levels, a more valid test was produced. Webster's results indicated that a test of 178 items with difficulty indices near .50 had a validity coefficient of .66. However, a test of 124 similar items with varying item difficulties had a validity coefficient of .76, statistically significant at $p < .03$ (based on r to z transformations).

Myers (1962), concerned by the current practice of selecting items based on varying item difficulties instead of the theoretical idea of p = .50, compared the effect of the current practice to the theoretical idea on reliability and validity of a scholastic aptitude test. The ideal item difficulty ranged from .40 to .74 in what he called the peaked test. Items selected by the current practice were outside the above range, and Myers called this the U-shaped test. Two sets of items were selected for the peaked test and the U-shaped test, four tests in all. Myers reported no statistically significant differences in test validity when the different tests were correlated with freshman grades. Test reliability was statistically significant at $P < .02$ (using the Wilcoxon matched pairs sign test) in favor of the peaked test. The reliability of the peaked test was .69. The reliability of the U-shaped test was .63. The author noted that the results above were based on a 24 item test, and that when test length was projected to 48 items (via Spearman-Brown Prophecy Formula) there were no significant differences in test reliability. The studies of Brogden (1946) and Webster (1956) indicate that selecting items of varying item difficulty tends to increase internal consistency and test validity. The results from Myer's (1962) study indicated just the opposite, that item difficulty near .50 produced the more internally consistent test. But this was only true for a relatively short test of 24 items, and that when the test length was projected to 48 items, there were no differences in the reliability of either test based upon the two methods of selecting items.

Simplified Methods of Obtaining Item Discriminations

A second major group of articles on classical item analysis has dealt with simplified methods of obtaining indices of item discrimination. Because of the lack of computers in the early years of test development many psychometricians concerned themselves with devising tables to provide quick estimates of item discrimination. Kelley (1939) found that in the computation of item discrimination only 54 percent of the examinee group (based on total test score) needed to be used. Considering the top 27 percent and the bottom 27 percent of the test scorers resulted in a considerable savings in computational time. Flanagan (1939) developed a table of item discriminations to estimate the PPM correlation between item and test score based on Kelley's extreme score groups of top and bottom 27 percent.

Fan (1952) developed a table for the estimation of the tetrachoric correlation coefficient using the upper and lower 27 percent of the scorers. The tetrachoric correlation is similar to the biserial correlation, where the correlation is between two variables, which are assumed to have a normal and continuous underlying distribution, but have been artifically dichotomized.

Guilford (1954) presented several short cut tabular and graphic solutions for estimating various types of correlation coefficients to measure test item validity. These methods result in saving a considerabove amount of time when one is forced to use hand calculations. Today these short cut methods can be used by classroom teachers who often do not have the aid of calculators or computers. However, many test constructors still use these classical methods of item analysis even

though computers are available with which more sophisticated item analytic techniques such as factor analysis or latent trait models can be used.

Item Analysis Procedures for the Factor Analytic Model

Charles Spearman (1904) proposed a theory of measurement based on the idea that every test was composed of one general factor and a number of specific factors. In order to test his idea Spearman developed the statistical procedure known as factor analysis.

> "Factor analysis is a method of analyzing a set of observations from their intercorrelations to determine whether the variations represented can be accounted for adequately by a number of basic categories smaller than that with which the investigation started" (Fruchter, 1954, p. 1).

Factor analysis is a mathematical procedure which produces a linear representation of a variable in terms of other variables (Harman, 1967). In the case of test items being factor analyzed, a matrix of item intercorrelations is obtained first. Subsequently, the matrix of item correlations is submitted to the factoring process. There are two basic alternatives within the framework of factor analysis for analyzing a set of data: common factor analysis, based on the work of Spearman and later Thurstone (1947); and principal components, developed by Hotelling (1933). The major distinction between the two methods relates to the amount of variance analyzed, e.g., the values placed in the diagonal of the intercorrelation matrix. Factoring of the correlation matrix with unities in the diagonal leads to principal components, while factoring the correlation matrix with communalities[4] in the diagonal

[4] The communality (h^2) of a variable is defined as the sum of the squared factor loadings $h^2 = a_{j1}^2 + a_{j2}^2 + \ldots a_{jn}^2$ (Harman, 1967, p. 17), see formula 8.

leads to common factor analysis (Harman, 1967). If it is of interest to know what the test items share in common, a common factor solution is warranted. But if it is of interest to make comparisons to other tests or other test development procedures, a principal components solution is warranted. Since the present study was initiated to compare three different test development techniques, a principal components solution was used in this study to analyze the date under the factor analytic model.

The linear model for the principal components procedure is defined as (Harman, 1967, p. 15):

$$Z_{ji} = a_{j1}F_1 + a_{j2}F_2 + \ldots a_{jn}F_n. \qquad (7)$$

Z_{ji} is the variable (or item) of interest, and a_{j1} is the coefficient, or more frequently referred to as the loading of variable Z_{ji} on component, (F_1). An important feature of principal components is that the extracted components account for the maximum amount of variance from the original variables. Each principal component extracted is a linear combination of the original variables and is uncorrelated with subsequent components extracted. Thus, the sum of the variances of all n principal components is equal to the sum of the variances of the original variables (Harman, 1967). According to Guertin and Bailey (1970), the principal components solution was designed basically for prediction, hence the need to use the maximum amount of variance in a set of variables.

Since factor analysis is based upon a matrix of intercorrelations, it is important that care be taken in selecting the appropriate coefficient. Several item coefficients are available: phi, phi/phi max, and the tetrachoric correlation coefficient. Carroll (1961)

pointed out several problems concerning the choice of a correlation coefficient to be used in factor analysis. The phi coefficient (used where both variables are true dichotomies) was found to be affected by disparate marginal distributions and often underestimated the PPM. The phi/phi max coefficient was developed to correct for the underestimation of phi, but the correction is not enough to counter the effect of extreme dichotomizations. Carroll recommended the tetrachoric coefficient as being the least biased by extreme marginal splits providing the variable under consideration was normally distributed in the population. Wherry and Winer (1953) had made conclusions similar to Carroll, but went on to say that when the normality assumption was met and the regression of test score on the item was linear the PPM and tetrachoric are identical. The tetrachoric correlation was used in the present study to obtain item intercorrelations.

Research Related to Factor Analysis in Test Development

The early use of factor analysis to construct and refine tests was suggested by the work of McNemar (1942) in revising the Stanford-Binet scales, and Burt and John (1943) in analyzing the Terman-Binet scales.

Several contemporary psychometricians have advocated the use of factor analysis in developing unidimensional tests (Cattell, 1957; Hambleton and Traub, 1973; Henrysson, 1962; Lord and Novick, 1968). A unidimensional test was defined briefly in the introduction to this chapter, but a more precise definition is warranted. Lumsden (1961) noted that a unidimensional test can be determined by the examinee response patterns. If the test items are arranged from easiest to hardest, $person_1$ who misses $item_1$ will miss all the other items, and

person$_2$ who gets item$_1$ correct but misses item$_2$ will miss all the subsequent items and so on. The above statement assumes infallible items. However, most tests constructed today contain fallible items, thus the response pattern will be disturbed by random error. Lumsden suggested in developing unidimensional tests factorially that the items be carefully selected on empirical grounds, thus reducing the problem of too many heterogeneous items and the possibility of obtaining multiple factors. By preselecting items one increases the chances of the items converging on one factor.

The importance of developing unidimensional tests is demonstrated most clearly in considering the concepts of test reliability and validity. For a test to be valid it must actually measure the trait it was intended to measure. For a test to be reliable it must provide similar results upon repeated measurement. It should be easier to estimate these two important aspects of a test when the test is unidimensional than when the test is multidimensional, hence the use of a unidimensional test in the present study.

Cattell (1957) has suggested that in the development of a factor homogeneous scale, one should preselect items, carry out a preliminary factor analysis, then select for further analysis those items which load on the first factor. Cattell defined an index of unidimensionality as the ratio of the variance of the first factor to the total test variance. This index has no set criterion and the sampling distribution is unknown.

Comparison of Factor Analysis to Classical Item Analysis

One measure of item validity, the biserial correlation was described for classical item analysis procedures. This same index is also obtained

by factor analysis. When the test items are factor analyzed, the factor loading a_{ij}, is the item-factor association that is considered a measure of item validity, e.g., the higher the factor loading, the greater the relationship between the item and the factor it measures. The factor loadings can be viewed as similar to the biserial correlations discussed under classical test theory. This relationship between factor loadings and biserial correlations has been discussed by several authorities (Guertin and Bailey, 1970; Henrysson, 1962; Richardson, 1936).

Factor analysis as an item analytic technique was not realistically possible for most psychometricians until the advent of high speed computers. Guertin and Bailey (1970) have predicted that with the increasing use of computers factor analysis will replace classical item analysis as a test development technique. Because it is possible for a test to reach the highest degree of homogeneity and yet be factorially a very odd mixture of factors (Cattell and Tsujioka, 1964), classical item analysis alone is not sufficient to determine if a test is unidimensional. However, factor analysis not only provides a measure of item-test correlation (the factor loading), it also provides an indication of how many items form a unifactor test. Thus, factor analysis has been advocated as a superior technique to classical item analysis (Guertin and Bailey, 1970). Using factor analysis in test development, psychometricians have advanced beyond an independent analysis of item intercorrelations to a simultaneous analysis of item intercorrelations with other individual items to obtain a measure of test unidimensionality and item-factor association.

However, there is an inherent flaw in factor analysis as there was in classical item analysis in test development. The flaw is that both procedures are sample dependent. When an item analysis procedure, or any procedure in general is sample dependent, it means that the results will vary from group to group. When the groups are very dissimilar, there is much variability. Gulliksen (1950) noted that a significant advance in item analysis theory would be made when a method of obtaining invariant item parameters could be discovered. To that end latent trait theory is an attempt to identify invariant item parameters.

Item Analysis Procedures for the Latent Trait Model

Latent trait theory specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on a test (Hambleton et al., 1977). The relationship is described by a mathematical function; hence latent trait models are mathematical models. As noted earlier, there are four major latent trait models for use with dichotomously scored data: the normal ogive, and the one-, two-, and three-parameter logistic models (Hambleton and Cook, 1977; Lord and Novick, 1968). All four models are based on the assumption that the items in the test are measuring one common ability and that the assumption of local independence exists between the items and examinees. These two assumptions imply that a test which measures only one trait or ability will have less measurement error in the test score than a test that is multidimensional, and that the response of an examinee to one item is not related to his response on any other item. Where the latent trait models begin to differ is with respect to the shape of their item characteristic curves.

The normal ogive, developed by Lord (1952a, 1953a), produces an item characteristic curve based on the following formula:

$$P_g(\theta) = \int_{-\infty}^{a_g(\theta - b_g)} \theta(t)dt, \qquad (8)$$

where $P_g(\theta)$ is the probability that an examinee with ability θ correctly answers item g, $\theta(t)$ is the normal density function, b_g represents item difficulty and a_g represents item discrimination.

The item characteristic curve of the two-parameter logistic model developed by Birnbaum (1968) has the same shape as the normal ogive, and Baker (1961) has shown them to be equivalent mathematical procedures. The shape of the item characteristic curve of the two-parameter logistic function is developed from the following formula:

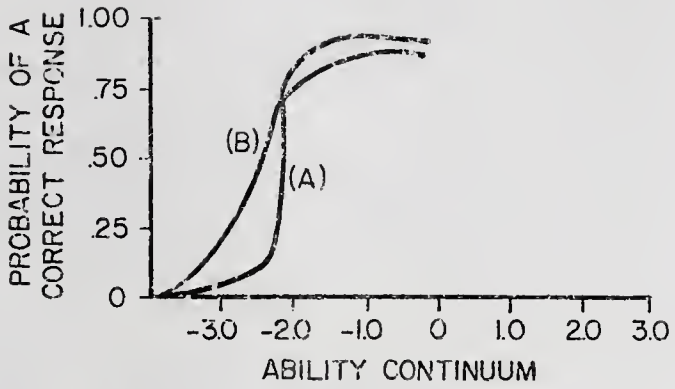
$$P_g(\theta) = \frac{e^{Dag(\theta - b_g)}}{1 + e^{Dag(\theta - b_g)}} . \qquad (9)$$

$P_g(\theta)$, a_g and b_g have the same interpretation as in the normal ogive. D is a scaling factor equal to 1.7 (the adjustment between the logistic function and normal density function), and e is the natural log function. In Figure 1a the shape of the normal ogive and the two-parameter logistic curve has been illustrated. In the Figure, item A is more discriminating than item B as noted by the steepness of the slopes.

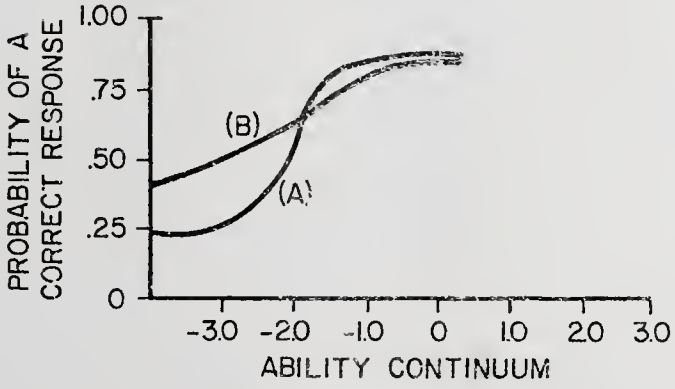
The three-parameter logistic model also developed by Birnbaum (1968) includes as an additional parameter, an index for guessing. The mathematical form of the three-parameter logistic curve is denoted,

$$P_g(\theta) = c_g + (1 - c_g)\frac{e^{Dag(\theta - b_g)}}{1 + e^{Dag(\theta - b_g)}} . \qquad (10)$$

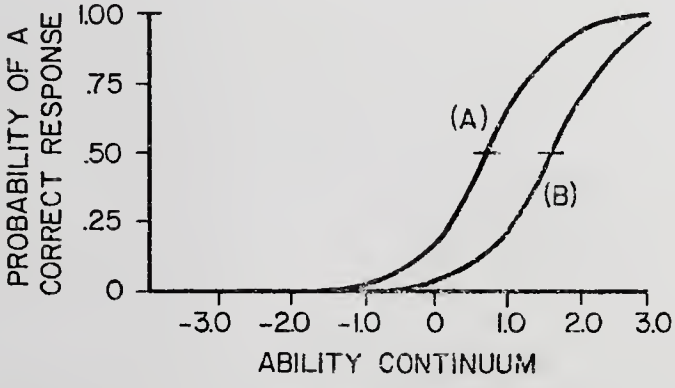
The parameter c_g, the lower asympotote of the item characteristic curve, represents the probability of low ability examinees correctly answering



NORMAL OGIVE & TWO-PARAMETER LOGISTIC CURVE (a)



THREE-PARAMETER LOGISTIC CURVE (b)



RASCH ONE-PARAMETER LOGISTIC CURVE (c)

FIGURE I. HYPOTHETICAL ITEM CHARACTERISTIC CURVES FOR THE FOUR LATENT TRAIT MODELS.

an item (Hambleton et al., 1977). In Figure 1b the shape of the three-parameter logistic curve has been illustrated. In the Figure, item A is more discriminating and has less guessing involved than item B.

The one-parameter logistic model, developed by Rasch (1960) is commonly referred to as the Rasch model. The Rasch model, though similar to the other latent trait models, was developed independently from the other models. The Rasch model is based upon two propositions: (a) the smarter an examinee, the more likely he is to answer the item correctly, and (b) an examinee is more likely to answer an easy item correctly than a difficult item. Mathematically the above propositions can be stated in terms of odds or probability of success on an item. The odds of an examinee with ability θ correctly answering an item with difficulty ζ is given by the ratio of θ to ζ (Rasch, 1960):

$$\text{odds} = \frac{\theta}{\zeta} \quad . \tag{11}$$

The derivation of equation 11 was presented in Appendix A. Equation 11 more formally written in the following equation is the Rasch model.

$$P\ (X_{ki} = 1 | \beta_k, \delta_i) = \frac{e^{(\beta_k - \delta_i)}}{1 + e^{(\beta_k - \delta_i)}} \quad . \tag{12}$$

In equation 12, the probability of examinee k making a correct response to item i, noted X = 1, given an examinee of ability β_k (where β_k is the log transformation of θ) taking an item of difficulty δ_i (where δ_i is the log transformation of ζ) is a function of the difference between the examinee's ability and the item's difficulty. The derivation of equation 11 to equation 12 is presented in Appendix A.

The assumptions for the Rasch model were discussed in Chapter I. Essentially the three assumptions are as follows:

1. There is only one trait underlying test performance.
2. Item responses of each examinee are statistically independent.
3. Item discriminations are equal.

The first two assumptions can be checked by conducting a factor analysis of the test items as suggested by Lord and Novick (1968), and Hambleton and Traub (1973). The assumptions are met if one dominant factor emerges from the analysis. The third assumption can be checked by plotting item characteristic curves for each item. In Figure 1c the item characteristic curves for two hypothetical items based on the Rasch model have been illustrated. The difficulty for items A and B is .5 and 1.5 respectively (point where p = .50), and the discriminations of the two items are equal. The assumption that all items have equal discriminations is quite restrictive; however, Rentz (1976) demonstrated, in a simulation study, that the item slopes can deviate from 1 (where all slopes are equal) $\pm$ .25 and still fit the model. In a similar simulation study, Dinero and Haertel (1976) concluded that the lack of an item discrimination parameter in the Rasch model does not result in poor item calibrations when discriminations are varied as much as .25.

The estimates for the Rasch parameters β_k and δ_i, examinee ability estimate and item difficulty estimate respectively, are sufficient, consistent, efficient, and unbiased (Anderson, 1973; Bock and Wood, 1971). That is, the examinee's test score will contain all the information necessary to measure the person ability parameter β_k, and the sum of the right answers to a given item will contain all the information used to calibrate the item parameter δ_i (Wright, 1977). Of the latent trait models, the Rasch model is unique in this respect.

The mathematical rationale of the Rasch model is based upon the separation of the ability and item difficulty parameters. As shown in Appendix A, the estimation of the item parameters is independent of the distribution of ability and ability independent of the distribution of item difficulty (Rasch, 1966). Several studies have demonstrated this (Anderson et al., 1968; Tinsley and Dawis, 1975; Whitely and Dawis, 1974; Whitely and Dawis, 1976; Wright, 1968; Wright and Panchapakeson, 1969). The separation of the ability and item parameters leads to what Rasch has termed specific objectivity. Specific objectivity relates to the fact that the measurement of a person's ability is not dependent upon the sample of items used, nor the examinee group in which a person is tested. Once a set of items has been calibrated to the Rasch model, any subset of the calibrated items will produce the same estimate of the examinee's ability. This type of objectivity is possessed by the physical sciences and the goal toward which mental measurement should be aimed in the future. Toward the goal of objective measurement several researchers have conducted empirical studies comparing classical factor analytic test development procedures to the latent trait models, and also comparisons have been made between the various latent trait models.

Research Related to Latent Trait Models in Test Development

Baker (1961) conducted one of the earlier comparative studies between two latent trait models. He compared the effect of fitting the normal ogive and the two-parameter logistic model to the same set of data, a scholastic aptitude test. The two-parameter model as well as the normal ogive provide item difficulty and item discrimination estimates.

The empirical results suggest there is little difference between the two procedures as measured by a chi-square test of fit. However, Baker noted the computer running time of the logistic model was one-third that of the ogive model, thus he concluded the logistic model was more efficient in terms of cost than the ogive.

Hambleton and Traub (1971) compared the efficiency of ability estimates provided by the Rasch model and the two-parameter model to the three-parameter logistic model using Birnbaum's concept of information (1968). The three-parameter model provides item difficulty and discrimination estimates as well as accounting for guessing on each item. Eleven simulated tests of fifteen items each were generated varying item discrimination and degree of guessing. The authors sought to determine how efficient the one- and two-parameter logistic models were under these conditions taking the three-parameter model to be the true model. The results indicated that when guessing was a factor the three-parameter model was most efficient in providing ability estimates, but when guessing was not a factor all models were equally efficient. Since the Rasch model has fewer parameters to estimate, hence it takes less computer time to run than the other two models, it would be preferred in the absence of guessing. In considering item discrimination, when the guessing parameter was set to zero, the Rasch model was as efficient as the two-parameter model when item discrimination varied from .39 to .79. As item discrimination deviated from this range the two-parameter model was more efficient.

Hambleton and Traub (1973) compared the one- and two-parameter models with three sets of real data (the verbal and mathematics subtests of a scholastic aptitude test used in Ontario (items = 45 and 20

respectively), and the verbal section of the Scholastic Aptitude Test (SAT, items = 80). Their results indicated that generally the two-parameter model fit the data better than the one-parameter model. The loss in predicting performance was greatest on the shorter mathematics test and smallest on the longer SAT. These findings confirm Birnbaum's conjecture (1968, p. 492) that if the number of items in a test is very large the inferences that can be made about an examinee's ability will be much the same whether the Rasch model or the two-parameter logistic model is used. The authors questioned whether the gain obtained with the two-parameter model is worth the increased computer cost of estimating the item discrimination parameter. Based on the results of these studies, it is concluded that the Rasch model is the most efficient of the latent trait models and hence will be used in comparison to the more traditional methods of test development included in the present study.

Comparison of the Rasch Model to Factor Analysis

Two recent studies have been completed comparing the Rasch model to factor analysis. Anderson (1976) posed two questions concerning the Rasch model and factor analysis: (a) what types of items would be excluded in terms of difficulty and discrimination using Rasch and factor analysis as item analytic techniques, and (b) what effect would the two procedures have on validity? Anderson chose to use 235 middle school students' responses to a 15 item Likert-type scale that was dichotomized for use with the Rasch model and the factor analytic procedures. A principal component factor analysis based upon tetrachoric correlation coefficients was compared to the Rasch model using the CALFIT computer program (Wright and Mead, 1975). Only items fitting the model were used. His results indicated that the Rasch procedure

eliminated the more difficult items and the factor analytic procedure eliminated the easier items; a statistically significant difference as determined by chi-square test at $p < .01$. For item discriminations the Rasch procedure eliminated very low and very high item discriminations, while the factor analytic procedure tended to reject only very low discriminations. The difference here was not statistically significant. The second question of test validity showed very similar results for the two procedures when test score was correlated with course grade point average.

In a similar study Mandeville and Smarr (1976) developed a two stage design. First they compared the Rasch procedure to factor analysis, then they combined the two analytic procedures. The authors felt the combined approach would be a more effective item analytic approach than any single method in determining which items fit the Rasch model. Two cognitive data sets (one standardized and one classroom) and one simulated set were used in the study. A rotated principal axis factor analysis based upon phi correlation coefficients were compared to the Rasch model using the CALFIT program.

The results indicated that for the standardized and simulated data sets the double procedure of factor analyzing the items, then submitting only the items loading on the first factor to the Rasch procedure was not really useful. The Rasch procedure alone was just as effective as the double procedure in selecting items that fit the model.

For the classroom data set the investigators found that 92 percent of the items fit the Rasch model, but upon factor analyzing these items only seven percent of the total test variance was associated with the first factor. Their results tend to indicate that factor analysis

and the Rasch procedure do not always identify the same unidimensional trait underlying test performance. However, the results of the Mandeville and Smarr study may be suspect for three reasons. First, the phi coefficient, which can be seriously affected when p and q take on extreme values, was used as a basis to form the intercorrelation matrix that was factor analyzed. The greater the difference in p and q the smaller will be the maximum correlation, hence very easy and very difficult items will have systematically lower coefficients and will tend to bias the results of the analysis in favor of moderately difficult items. Second, the factor analysis was based on a principal axis solution, using some value less than 1.00 in the diagonal hence less variance is being used in the total solution for comparison with the Rasch procedure that is utilizing all the test variance available. Third, the principal axis solution was rotated so that the total variance associated with the first factor has been distributed out among the other factors and was no longer as strong as it once had been.

Summary

In the development of tests based upon classical item analysis two main statistics are used in reviewing and revising test items, e.g., item difficulty and item discrimination. The item discrimination index provides information as to the validity of the item in relation to total test score, while item difficulty indicates how appropriate the item was for the group tested. A serious limitation of classical item analysis is that the statistics obtained for examinees and items are sample dependent (Hambleton and Cook, 1977; Wright, 1968).

The same problem of sample dependency also exists for factor analysis. However, factor analysis is viewed as a superior technique

to classical item analysis for two reasons: (a) factor analysis compares item intercorrelations with other items simultaneously, and (b) factor analysis provides an indication of how many factors or abilities the test is measuring. Also in factor analysis, the factor loading is comparable to the item discrimination index of classical item analysis, thus providing a measure of item validity for each item on each factor in the test.

Not until the development of latent trait models was a solution suggested to the problem of sample dependency of the statistics for items and examinees. The Rasch model in particular has been shown to provide item statistics that are independent of the group on which they were obtained, as well as examinee statistics that are independent of the group of items on which they were tested. This feature of the Rasch model provides for more objective mental measurement.

The Rasch model has been compared to other latent trait models and has been shown to be as efficient in many cases as the more complex models. The Rasch model has also been compared with factor analytic procedures in determining test unidimensionality, validity, and types of items retained and excluded by the two procedures. Missing from this review is a comparative study of the three item analytic techniques using the same data base and a comparison of the efficiency of tests developed from the three techniques across ability levels. Also missing from the literature is the effect of varying sample size and number of items as well as the kinds of items each of the three procedures would either retain or exclude in test development.

It is apparent that an empirical investigation into these areas seems warranted to determine which procedure under the various conditions would produce the superior test in terms of internal consistency and efficiency. It was for this reason that the present study was undertaken comparing the three methods of classical item analysis, factor analysis, and the Rasch model used in test development. The design of the study is described in Chapter III.

CHAPTER III

METHOD

An empirical study was designed to compare the effects of three methods of item analysis on test development for different sample sizes. The three methods of item analysis studied were classical item analysis, factor analysis, and Rasch analysis. The sample sizes used to compare the three item analytic methods were 250, 500, and 995 subjects. The study was designed in three phases: (a) item selection, (b) a double cross-validation of the selected items, and (c) statistical analyses of the selected items. For each item analytic procedure two tests were developed, a 15 item test, and a 30 item test. Four dependent variables were obtained for each test: (a) an estimate of internal consistency, (b) the standard error of measurement, (c) item difficulty, and (d) item discrimination. A description of the subjects, instrument used, research design, and statistical analyses is presented in this chapter.

The Sample

In the fall of 1975, all high school seniors in the State of Florida (N = 78,751) were tested as part of the State assessment program. The population was from 435 high schools throughout the state. From this population a 1 in 15 systematic sample of 5,250 subjects was chosen (Mendenhall, Ott, and Scheaffer, 1971). A systematic sample was selected to ensure samples from every high school in the state. The

types of data obtained on each subject were sex, race, item responses, and total score.

The data file was edited to remove those subjects who either answered all the items correctly or incorrectly. The rationale for this procedure was that the Rasch model cannot calibrate items when a person has a perfect score or the alternative, when a person has no items correct (Wright, 1977). Through the editing procedure 15 subjects were removed, thus the available sample size was 5,235. Because such a small number of subjects were removed, it seems unlikely that the elimination of these subjects would bias the results in favor of any of the three item analytic techniques.

The Instrument

The instrument selected for use in this study was the Verbal Aptitude subtest of the Florida Twelfth Grade Test, developed by the Educational Testing Service. It is a statewide assessment battery which has been administered every year since 1935 (Benson, 1975). The Verbal Aptitude subtest is comprised of 50 verbal analogies, in a multiple choice format, from which a single score based on the number of items correct is reported. Descriptive information on the Verbal Aptitude subtest for the population tested in 1975 is presented in Table 1.

This particular instrument was selected for three reasons. First, it is a cognitive measure of verbal ability and much of classical test theory has been build upon tests in the cognitive domain. Second, it is similar to and hence representative of other national aptitude tests used for college admissions. Third, it has a large data pool from which to sample.

TABLE 1

DESCRIPTIVE DATA ON THE VERBAL APTITUDE SUBTEST
OF THE FLORIDA TWELFTH GRADE TEST
1975 ADMINISTRATION

Number of Schools = 435 Number of Students = 78,751

Number of items	50
Mean	25.95
Standard Deviation	8.23
Reliability[a]	.88
Standard Error of Measurement	2.85

Note: Data obtained from the Florida Twelfth Grade Testing Program, Report No. 1-75, Fall 1975.

[a]Reliability based on the split-half method, and corrected by the Spearman-Brown formula.

Classical test theory has been built mainly around the development of cognitive tests. Therefore, it seemed desirable to compare the new procedures of latent trait theory, via the Rasch model to the procedures of classical test theory, e.g., factor analysis and classical item analysis by using a cognitive test. Thus, the results may be more generalizable to the major type of tests developed by practitioners in the field.

The Procedure

Design

The sample of 5,235 was divided into nine systematic samples in the following manner:

$Group_1$ = three independent samples of 250 students each;

$Group_2$ = three independent samples of 500 students each;

$Group_3$ = three independent samples of 995 students each.

From the initial editing of the data file, previously described, 15 subjects were removed from the total sample of 5,250. Therefore, it was decided that this loss of subjects would only affect $Group_3$ since it was the largest. Thus, the number of subjects in each of the three independent samples was reduced by five, resulting in three independent samples of 995 subjects each.

The purpose of obtaining the three separate samples for the three groups was to insure that each item analytic and double cross-validation procedure used an independent sample, so that tests of statistical significance could be performed. The scheme shown in Table 2 was used to obtain the nine samples. In the present study the independent variables were sample size and item analytic procedure.

TABLE 2

SYSTEMATIC SAMPLING DESIGN OF THE STUDY[a]
N = 5,235

Group	Sampling Procedure	Sample Number	Number Selected	Item Analytic Procedure	Total Sample Remaining
$Group_1$	1 in 20	1	250	Classical	4,985
	1 in 19	2	250	Factor Analysis	4,735
	1 in 18	3	250	Rasch	4,485
$Group_2$	1 in 8	4	500	Classical	3,985
	1 in 7	5	500	Factor Analysis	3,485
	1 in 6	6	500	Rasch	2,985
$Group_3$[b]	1 in 3	7	995	Classical	1,995
	1 in 2	8	995	Factor Analysis	995
	remaining	9	995	Rasch	0

[a]The sampling procedure was randomly assigned to item analytic technique in $group_1$ and the same pattern carried out for $group_2$ and $group_3$.

[b] Those subjects edited from the data file were removed equally from $group_3$, hence the reduced sample size.

The item data were analyzed in three phases: (a) selection of the items, (b) computation of item and test statistics for selected items on double cross-validation samples, and (c) statistical analyses of item characteristics to test the hypotheses.

Item Selection

The three independent samples, within each of the groups of subjects (N = 250, N = 500, N = 995), were submitted to one of the three item analytic procedures (in accordance with Table 2) in order to select a specified number of items, e.g., the "best" 15 and 30 items. Each of these two sets of items comprised two separate tests; however, all of the items on the 15 item tests were always included on each of the 30 item tests. A different process for selecting the items was used with each item analytic technique, and has been described in the following three sections.

Classical item analysis. The definition of the "best" items was based on the numerical magnitude of the items' biserial correlations. The biserial correlation was defined as the correlation between the artifically dichotomized item score (1 or 0) and total test score. In using the biserial correlation the assumption was made that the artifically dichotomized variable (the item) had a continuous and normal distribution (Magnusson, 1966).

In order to obtain biserial correlations for the items under the classical item analysis procedure, the 50 verbal items were submitted to the item analysis program, GITAP[5] for each of the three sample sizes.

---

[5]The Generalized Item Analysis Program (GITAP) is a part of the test analysis package developed by F. B. Baker and T. J. Martin, Occasional Paper No. 10, Michigan State University, 1970.

The 15 and 30 items with the highest biserial correlations were selected as the best items from the total subtest. Item difficulties were also obtained for the "best" 15 and 30 items selected. Item difficulty has been defined as the proportion of persons getting a particular item correct out of the total number of persons attempting that item (Mehrens and Lehman, 1973).

Factor analysis. Item selection based on factor analysis was accomplished using the computer programs developed for the Education Evaluation Laboratory at the University of Florida. These programs have been described by Guertin and Bailey (1970). The present study was concerned only with the items that load on the first principal component, in order to adhere to the unidimensionality assumption of the test. The principal components analysis was based on a matrix of tetrachoric item intercorrelations with unities in the diagonal.

The tetrachoric correlation was chosen to produce the intercorrelation matrix for the same reason the biserial correlation was chosen: Knowledge of an item was assumed to be normal and continuously distributed. In the case of the tetrachoric correlation each item (scored 1 or 0) was correlated with every other item.

The 15 and 30 items with the highest loadings on the first unrotated principal component were selected from the total subtest. These component loadings are analogous to biserial correlations previously described, where the loading refers to the relationship of the item to the principal component or factor (Guertin and Bailey, 1970, Henrysson, 1962).

Rasch analysis. The selection of items based on the Rasch model was accomplished in two stages. First, in order to check the assumption

of a unidimensional test, a factor analysis using a principal components solution was used. Items were selected with loadings between .39 and .79 on the first unrotated factor, to hold the discrimination index of the items constant. Hambleton and Traub (1971) have shown that the efficiency of a test developed using the Rasch model will remain very high (over 95 percent) when the range on the discrimination index was held between .39 and .79. Second, the items selected from the principal components solution using the above criteria were submitted to a Rasch analysis using the BICAL program (Wright and Mead, 1976). Items were selected based upon the mean square fit of the items to the Rasch model. The best 15 and 30 items fitting the model were chosen from the total subtest, and their corresponding item difficulties reported.

Double Cross-Validation

A double cross-validation design (Mosier, 1951) was used to obtain item parameter estimates for the best 15 and 30 items selected by the three item analytic techniques for the three sample sizes. In this study a 3 X 3 latin square was used to reassign samples. This procedure ensured that the estimates of the item parameters would be based upon a different sample of subjects than the original sample used to identify the best items. Each item analytic technique was randomly reassigned, using a latin square procedure (Cochran and Cox, 1957, p. 121), to a different sample within each of the three groups (N = 250, N = 500, N = 995). The double cross-validation design is shown in Table 3.

The best 15 and 30 items selected by each item analytic procedure in the first phase of the study, were submitted to a standard item

TABLE 3

DOUBLE CROSS-VALIDATION DESIGN OF THE STUDY

Group	Sample Number[a]	Number Selected	Item Analytic Procedure	Double Cross-Validation Procedure[b]
$Group_1$	1	250	Classical	Factor Analysis
	2	250	Factor Analysis	Rasch
	3	250	Rasch	Classical
$Group_2$	4	500	Classical	Rasch
	5	500	Factor Analysis	Classical
	6	500	Rasch	Factor Analysis
$Group_3$	7	995	Classical	Rasch
	8	995	Factor Analysis	Classical
	9	995	Rasch	Factor Analysis

[a]The sample number is the same as referred to in Table 2.

[b]Assignment to sample was based on a randomized 3 X 3 latin square procedure.

analysis program (GITAP) from which were obtained the dependent variables in the study:

- indices of internal consistency as measured by the analysis of variance procedure (Hoyt, 1941)
- the standard error of measurement
- item difficulty
- biserial correlations

By submitting the best 15 and 30 items selected by each item analytic procedure in the study to a common item analysis program comparable measures of the dependent variables were obtained.

Statistical Analyses

The third phase of the study focused on obtaining measures of statistical significance for three of the dependent variables: internal consistency, item difficulties, and biserial correlations. Only visual comparisons were made for the remaining dependent variable, the standard error of measurement. The internal consistency estimates from each test were compared to the projected population value for tests of similar length via confidence intervals as suggested by Feldt (1965). (Projected population values were obtained using the Spearman-Brown Prophecy Formula.)

Item difficulties for the 15 and 30 best items were submitted to a two-way analysis of variance,[6] the two factors being sample size and item analytic technique. This procedure was used to test for differences in the types of items selected, in terms of item difficulty, by each technique. If statistical significance was observed, with $\alpha = .05$, Tukey's HSD (honestly significant difference) post hoc procedure (Kirk, 1968) was

---

[6]The analysis of variance procedure is appropriate only if the distribution of the item difficulties and (transformed biserial correlations

employed to determine which item analytic technique(s) resulted in a test with the highest item difficulties.

The biserial correlations were transformed to an interval scale of measurement using a linear function of $\underline{z}$ suggested by Davis (1946). The linear transformation was based upon converting the biserial correlation to $\underline{z}$ values, and then eliminating the decimals and negative values of $\underline{z}$ by multiplying the constant 60.241 to each $\underline{z}$ value (Davis, 1946, pp. 12-15). Thus, the range of the transformed biserials ranged between 0 and 100. A two-way analysis of variance [7] (sample size by item analytic technique) was performed on the transformed biserial correlations for the best 15 items. This type of analysis was used to test for differences in the types of items selected, in terms of biserial correlations by each technique. If statistical significance was observed, $\alpha = .05$, Tukey's HSD post hoc procedure was employed to determine which item analytic technique(s) resulted in higher transformed biserial correlations.

The two-way analysis of variance and post hoc analysis, where indicated, for the transformed biserial correlations was performed on the 30 best items.

In addition to tests of statistical significance, a measure of the efficiency of the 30 best items selected by each procedure was compared for the sample of 995 subjects. Birnbaum (1968) defined the relative efficiency of two testing procedures as the ratio of their

---

approximates normality and the variances are homogeneous (Ware and Benson, 1975).

[7] The analysis of variance procedure is appropriate only if the distribution of the item difficulties and (transformed) biserial correlations approximates normality and the variances are homogeneous (Ware and Benson, 1975).

information curves. Lord (1974a) has described a procedure to compare the relative efficiency of one test with another at different ability levels. If two tests to be compared vary in difficulty, then the relative efficiency of each will usually be different at different ability levels (Lord, 1974b; 1977). In classical test theory it is common to compare two tests that measure the same ability in terms of their reliability coefficients, but this only gives a single overall comparison. The formula developed by Lord for relative efficiency provides a more precise way of comparing two tests that measure the same ability. The formula for approximating relative efficiency is (Lord, 1974b, p. 248):

$$\text{R.E. } (y,x) = \frac{n_y}{n_x} \cdot \frac{x\,(n_x - x) f_x^2}{y\,(n_y - y) f_y^2}, \qquad (13)$$

where R.E. denotesthe relative efficiency of y compared to x, n_y and n_x denote the number of items in the two tests, x and y are the number-right scores having the same percentile rank, and f_x^2 and f_y^2 are the squared observed frequencies of x and y. Lord has suggested that formula 13 only be used with a large sample of examinees and tests that are not extremely short, hence this comparison was restricted to the case where N = 995 and the 30 item test.

Three relative efficiency comparisons using the 30 item tests were made: (a) the test based on factor analysis was compared to the test based on classical item analysis, (b) the test based on the Rasch analysis was compared to the test based on classical item analysis, and (c) the test based on the Rasch analysis was compared to the test based on factor analysis.

Summary

An empirical study was designed to compare the effects of classical item analysis, factor analysis, and the Rasch model on test development. Item response data were obtained from a sample of 5,235 high school seniors on a cognitive test of verbal aptitude.

The subjects were divided into 9 samples: three independent groups of 250 subjects each, three independent groups of 500 subjects each, and three independent groups of 995 subjects each. The independent groups were obtained so that tests of statistical significance could be performed.

The item response data were then analyzed in three phases. First, the "best" 15 and 30 items were selected using each item analytic technique. Under classical item analysis, the best 15 and 30 items were selected based on the highest biserial correlations. For factor analysis, the best 15 and 30 items were selected based on the highest item loadings on the first (unrotated) principal component. The selections of the best 15 and 30 items using the Rasch model were based upon the mean square fit of the items to the model. These procedures were used for each group of subjects. Second, a double cross-validation design was employed to obtain estimates on the item parameters for the best 15 and 30 items. The three item analytic techniques were reassigned randomly to different samples of subjects within each level of sample size. Then, the best 15 and 30 items chosen by each method were submitted to a common item analytic procedure in order to obtain estimates for comparing the three item analytic methods. Third, a two-way analysis of variance and a Tukey post hoc comparison test, when indicated, were used to test for differences in

the properties of items selected by each item analytic procedure. Also confidence intervals were calculated to compare the internal consistency estimates to a population value. In addition, the relative efficiencies of the 30 item tests developed by each item analytic technique were compared for the sample of 995 subjects.

CHAPTER IV

RESULTS

The study was designed to compare empirically the precision and efficiency of tests developed using three item analytic techniques: classical item analysis, factor analysis, and the Rasch model. The following five hypotheses were generated to compare the three techniques:

1. There are no significant differences in the internal consistency estimates of the tests produced by the three methods as the number of items decreases when compared to the projected internal consistency estimates for the population for tests of similar length.

2. There are no differences in the internal consistency estimates of the tests produced by the three methods when the nunber of examinees is decreased.

3. There are no meaningful[8] differences in the magnitude of the standard error of measurement of the tests produced by the three methods.

4. There are no significant differences in the difficulties or discriminations of the items selected by the three methods.

---

[8] A meaningful difference was previously defined to be ≥ 1.00.

5. There are no differences across ability levels in the efficiency of the tests produced by the three methods.

The Verbal Aptitude subtest of the Florida Twelfth Grade Test, was used to test the hypotheses. A sample of 5,235 examinees was systematically selected from a population of 78,751. A demographic breakdown of the sample by ethnic origin and sex is presented in Table 4.

The data were analyzed and reported in the following manner: item selection, double cross-validation, comparison of the 15 item tests on precision and comparison of the 30 item tests on precision and efficiency. These results were then summarized with respect to the five hypotheses.

Item Selection

The 50 items on the Verbal Aptitude subtest were submitted to each of the three item analytic techniques. The means, medians, and standard deviations of the biserial correlations and item difficulties, based on classical item analysis, are presented in Table 5. These descriptive statistics appear equivalent across the varying sample sizes.

From the factor analysis, the percentage of total test variance accounted for by the 50 verbal items on the first unrotated principal component has been reported in Table 6. The percentage of variance accounted for by the first principal component was obtained by summing the squared item loadings and dividing by the total number of items. The percentages of variance accounted for by the first principal component in each sample were very similar. A check on the unidimensionality of the test was made by rotating the principal components solution for the sample of 995 subjects. Upon rotation, the results indicated one dominate factor remained.

TABLE 4

DEMOGRAPHIC BREAKDOWN BY ETHNIC ORIGIN AND SEX FOR TOTAL SAMPLE[a]

Sex	White	Black	Cuban	Spanish American	American Indian	Mexican American	Puerto Rican	Oriental	Other	Total
Male	1775	353	87	36	52	15	7	8	54	2387
Female	1854	452	90	38	28	10	9	12	61	2554
Total	3629	805	177	74	80	25	16	20	115	4941[b]

[a]Cell entries are actual number of students.

[b]There were 294 students who did not indicate their race or sex.

TABLE 5

SUMMARY STATISTICS ON THE 50 TEST ITEMS BASED
ON CLASSICAL ITEM ANALYSIS FOR EACH SAMPLE SIZE

Item	N = 250		N = 500		N = 995	
	Biserial	Difficulty	Biserial	Difficulty	Biserial	Difficulty
1	.55	.70	.35	.75	.49	.73
2	.74	.92	.54	.89	.59	.90
3	.47	.74	.53	.81	.62	.77
4	.43	.79	.38	.81	.42	.81
5	.63	.85	.58	.87	.63	.85
6	.59	.79	.59	.81	.63	.81
7	.74	.64	.77	.57	.76	.60
8	.53	.71	.48	.72	.41	.77
9	.36	.72	.53	.69	.63	.67
10	.62	.75	.64	.74	.72	.75
11	.62	.69	.60	.66	.57	.70
12	.53	.63	.60	.63	.64	.66
13	.72	.50	.68	.54	.65	.53
14	.71	.79	.58	.76	.68	.78
15	.51	.66	.52	.63	.53	.65
16	.41	.70	.34	.68	.42	.68
17	.58	.50	.57	.52	.51	.55
18	.58	.68	.68	.65	.56	.70
19	.65	.76	.77	.71	.73	.73
20	.59	.55	.66	.53	.62	.51
21	.16	.52	.30	.49	.26	.52
22	.57	.50	.52	.57	.54	.56
23	.50	.49	.38	.47	.40	.50
24	.57	.46	.59	.49	.60	.48
25	.42	.52	.44	.55	.47	.52
26	.63	.48	.50	.56	.52	.55
27	.66	.29	.53	.36	.59	.32

TABLE 5 - Continued

Item	N = 250 Biserial	N = 250 Difficulty	N = 500 Biserial	N = 500 Difficulty	N = 995 Biserial	N = 995 Difficulty
28	.56	.57	.48	.60	.52	.58
29	.43	.40	.57	.44	.46	.45
30	.27	.41	.39	.48	.32	.46
31	.56	.55	.54	.56	.53	.55
32	.16	.43	.32	.42	.25	.49
33	.28	.54	.36	.51	.31	.55
34	.45	.26	.44	.41	.42	.41
35	.42	.41	.51	.44	.40	.45
36	.41	.36	.50	.42	.38	.41
37	.34	.37	.43	.36	.48	.36
38	.36	.35	.31	.33	.38	.40
39	.38	.36	.55	.37	.54	.37
40	.46	.33	.56	.35	.46	.34
41	.25	.30	.30	.31	.32	.28
42	.44	.29	.34	.24	.33	.25
43	.45	.26	.33	.26	.30	.28
44	.33	.33	.38	.33	.45	.30
45	.22	.33	.23	.32	.22	.30
46	.30	.25	.33	.29	.25	.27
47	.34	.28	.39	.24	.41	.27
48	.49	.19	.28	.21	.46	.23
49	.23	.14	.22	.15	.24	.15
50	.36	.18	.38	.17	.36	.19
Median Biserial	.48		.50		.48	
Mean Biserial	.48		.48		.48	
Standard Deviation	.15		.13		.14	
Mean Difficulty		.51		.51		.52
Standard Deviation		.20		.19		.19

TABLE 6

ITEM LOADINGS ON THE FIRST UNROTATED FACTOR FOR THE 50 TEST ITEMS BASED ON FACTOR ANALYSIS FOR EACH SAMPLE SIZE

Item	N = 250 Item Loading	N = 500 Item Loading	N = 995 Item Loading
1	.40	.35	.40
2	.63	.68	.68
3	.61	.53	.56
4	.42	.38	.36
5	.67	.63	.60
6	.60	.57	.52
7	.68	.70	.73
8	.47	.50	.42
9	.70	.67	.64
10	.64	.75	.75
11	.51	.63	.67
12	.59	.64	.63
13	.61	.63	.72
14	.64	.58	.64
15	.55	.55	.49
16	.49	.37	.46
17	.52	.54	.57
18	.60	.67	.58
19	.63	.77	.71
20	.60	.64	.68
21	.22	.17	.28
22	.42	.55	.52
23	.39	.35	.34
24	.45	.55	.58
25	.51	.41	.41
26	.48	.56	.53
27	.49	.58	.59

TABLE 6 - Continued

Item	N = 250 Item Loading	N = 500 Item Loading	N = 995 Item Loading
28	.42	.43	.47
29	.54	.55	.50
30	.23	.18	.21
31	.40	.49	.52
32	.20	.25	.23
33	.35	.31	.28
34	.40	.46	.34
35	.42	.31	.43
36	.48	.34	.33
37	.40	.35	.38
38	.34	.33	.28
39	.48	.40	.47
40	.48	.43	.46
41	.21	.26	.26
42	.22	.33	.34
43	.31	.20	.31
44	.47	.41	.38
45	.34	.31	.25
46	.29	.26	.19
47	.36	.42	.40
48	.27	.28	.36
49	.23	.20	.16
50	.25	.24	.30
Percentage of Total Test Variance Accounted for on 1st Unrotated Factor	22.4%	23.2%	23.5%

Three statistics are reported in Table 7 for the Rasch item analysis procedure. For each sample size, the percentage of total variance accounted for by the first unrotated principal conponent and the means and standard deviations of the mean square fit statistic and Rasch difficulties are presented for the items selected.

In order to select the best 15 and 30 items from the Rasch analysis, all 50 items were submitted to a principal components solution. This procedure was used to ensure that the items selected measured one trait, as required by the assumption of test unidimensionality. As noted in Table 7, the percentage of total test variance accounted for by the first principal component, based on 50 items, was nearly equal for each sample size.

From the principal components solution only items with loadings between .39 and .79 were selected for the Rasch analysis as suggested by Hambleton and Traub (1971), to adhere to the assumption of equal item discriminations. Using this procedure the number of items (out of 50) retained for the Rasch analysis varied slightly with sample size; when N = 250, 33 items were retained, when N = 500, 35 items were retained, and when N = 995, 33 items were retained. These items, loading between .39 and .79, were then submitted to the Rasch analysis to obtain mean square fit statistics and Rasch item difficulties. These statistics have been reported in Table 7.

Wright and Panchapakesan (1969) developed a measure to assess the fit of the item to the Rasch model. The measure, defined as the mean square fit statistic, is:

TABLE 7

SUMMARY STATISTICS ON THE 50 TEST ITEMS BASED ON THE RASCH MODEL FOR EACH SAMPLE SIZE

Item	N = 250			N = 500			N = 995		
	Loading	MS_a Fit	Difficulty[b]	Loading	MS_a Fit	Difficulty[b]	Loading	MS_a Fit	Difficulty[b]
1	.34	----	-----	.39	2.37	- .85	.43	2.09	- .82
2	.50	.80	-2.10	.59	.69	-2.36	.69	6.31	-2.14
3	.54	.85	-1.29	.51	.51	-1.08	.52	1.78	- .95
4	.34	----	-----	.50	.70	-1.11	.33	----	-----
5	.48	.16	-1.84	.60	2.30	-1.40	.61	3.32	-1.51
6	.70	2.40	-1.29	.60	1.62	-1.13	.63	1.80	-1.14
7	.74	2.39	.04	.72	5.27	- .09	.69	5.66	- .08
8	.56	.55	- .72	.43	1.34	- .70	.44	1.96	- .80
9	.54	.84	- .82	.60	1.48	- .36	.58	2.12	- .40
10	.77	4.40	-1.18	.66	2.59	-1.01	.72	7.91	- .80
11	.63	1.42	- .55	.56	.92	- .74	.64	1.96	- .50
12	.59	1.59	- .21	.60	.89	- .32	.59	2.33	- .23
13	.67	1.84	.33	.65	2.02	.42	.63	2.65	.40
14	.50	.83	-1.32	.68	2.68	-1.20	.59	1.32	- .92
15	.58	.45	- .42	.53	1.11	- .24	.56	.72	- .27
16	.39	1.74	- .57	.53	.22	- .40	.40	6.02	- .42
17	.47	.26	,16	.51	.54	.37	.54	.84	.36
18	.47	.54	- .42	.59	.51	- .51	.70	4.43	- .49
19	.75	2.97	- .72	.71	3.40	- .58	.71	5.83	- .65
20	.63	1.08	.23	.59	.68	.31	.60	.69	.40
21	.27	----	-----	.29	----	-----	.26	----	-----
22	.42	1.16	.50	.48	.58	.15	.48	2.58	.25
23	.43	1.85	.40	.37	----	-----	.44	2.44	.58
24	.40	1.35	.40	.51	.35	.61	.54	1.38	.57
25	.36	----	-----	.48	1.53	.56	.38	----	-----
26	.50	1.04	.06	.58	.90	.20	.56	1.98	.30

TABLE 7 - Continued

Item	N = 250 Loading	MS Fit[a]	Difficulty[b]	N = 500 Loading	MS Fit[a]	Difficulty[b]	N = 995 Loading	MS Fit[a]	Difficulty[b]
27	.49	.14	1.30	.57	.68	1.24	.53	1.73	1.14
28	.56	.38	- .17	.52	.19	.04	.51	1.09	- .10
29	.47	1.32	.77	.46	2.24	.74	.51	.87	.73
30	.23	----	-----	.31	----	-----	.36	----	-----
31	.42	1.10	.25	.47	1.74	.30	.47	1.71	.24
32	.27	----	-----	.19	----	-----	.26	----	-----
33	.19	----	-----	.28	----	-----	.27	----	-----
34	.52	.78	1.05	.43	1.88	1.26	.44	4.08	1.02
35	.34	----	-----	.45	3.82	.89	.44	3.82	.77
36	.45	.88	1.05	.46	.87	.97	.36	----	-----
37	.49	1.19	1.30	.38	----	-----	.38	----	-----
38	.28	----	-----	.29	----	-----	.25	----	-----
39	.34	----	-----	.53	.21	1.36	.45	3.22	1.15
40	.32	----	-----	.43	3.91	1.19	.51	2.44	1.26
41	.28	----	-----	.20	----	-----	.16	----	-----
42	.44	2.06	1.56	.31	----	-----	.28	----	-----
43	.22	----	-----	.31	----	-----	.24	----	-----
44	.40	1.44	1.05	.31	----	-----	.39	11.25	1.30
45	.22	----	-----	.26	----	-----	.25	----	-----
46	.33	----	-----	.32	----	-----	.34	----	-----
47	.48	2.41	1.37	.48	1.17	1.46	.48	1.14	1.75
48	.39	2.21	1.79	.39	4.33	2.01	.36	----	-----
49	.32	----	-----	.22	----	-----	.20	----	-----
50	.38	----	-----	.29	----	-----	.36	----	-----
Percentage of Total Test Variance Accounted for on First Unrotated Factor	22.1%			23.2%			23.4%		

TABLE 7 - Continued

Item	N = 250 MS Loading Fit[a]	N = 250 Difficulty[b]	N = 500 MS Loading Fit[a]	N = 500 Difficulty[b]	N = 995 MS Loading Fit[a]	N = 995 Difficulty[b]
Mean MS fit	1.35		1.61		3.02	
Standard Deviation	.91		1.28		2.35	
Mean Difficulty		- .00		- .00		- .00
Standard Deviation		1.01		.98		.90

[a]Mean square fit refers to the fit of the items to the Rasch model. The number of items used was based upon a principal components factor analysis where only the items with loadings between .39-.79 were selected for entry into the Rasch calibration.

[b]Item difficulties from the Rasch model are centered at a mean of zero with a standard deviation of one. Negative values indicate easy items and positive values indicate difficult items.

$$\chi^2 = \sum_{i=1}^{k-1} \sum_{j=1}^{n} y_{ij}^2 \quad . \qquad (14)$$

The quanity, χ^2, defined above has approximately the chi square distribution with degrees of freedom equal to (k-1) (n-1). The value, y_{ij} is the deviation of the item from the model, or item misfit, and is determined by taking the difference between the observed and expected frequency of the examinees at a given ability level who answered a given item correctly. This difference was then divided by the standard deviation of the observed frequency, squared and summed over items and score groups. The BICAL program standardizes these deviations (y_{ij}) in computing the mean square fit statistic; therefore, y_{ij} has a normal distribution with a mean of zero and standard deviation of one (Hambleton et al., 1977). Items with large mean square fit values are items which do not fit the model. As shown in Table 7, the mean and standard deviation of the mean square fit statistic increased with sample size.

The item difficulty estimates based on the Rasch model also have an expected mean of zero and standard deviation of one (Wright and Mead, 1975). These estimates remained very similar across sample size and exceptionally close to the expected values (Table 7).

The Rasch model does not provide a parameter for item discriminating power as all item discriminations are considered equal and centered at one (Wright and Mean, 1975). The BICAL program provided, as part of the normal output, estimates of the item's discriminating power to check the fit of the data to the model. The item discriminations were obtained by regressing the difficulty of the item for each ability group on the ability estimate of the group (Wright and Mead, 1975, p. 11).

The means and standard deviations for the item discrimination estimates were shown in Table 8 for each sample size.

TABLE 8

DESCRIPTIVE DATA ON ITEM DISCRIMINATION ESTIMATES BASED ON THE RASCH MODEL ACCORDING TO SAMPLE SIZE

	N = 250[a] K = 33[b]	N = 500 K = 35	N = 995 K = 33
Mean	1.03	1.02	1.03
Standard Deviation	.28	.19	.22

[a] N = sample size

[b] K = number of items

From the data in Table 8, the mean item discrimination estimates appear nearly equal for each sample size, and quite close to the mean expected value of one.

The best 15 and 30 items were then selected by each item analytic procedure based on the information in Tables 5-7, and have been listed in Tables 9 and 10 respectively.

The items selected under classical item analysis were determined by the magnitude of the biserial correlation, e.g., the 15 and 30 items having the highest biserial correlations with total test score were selected. Indices of item difficulty have been reported for inspection, but in no way influenced the selection of items for classical item analysis.

TABLE 9

THE 15 BEST ITEMS SELECTED UNDER EACH ITEM ANALYTIC PROCEDURE ACCORDING TO SAMPLE SIZE[a]

N = 250				N = 500				N = 995			
Item	Classical	Factor Analysis	Rasch	Item	Classical	Factor Analysis	Rasch	Item	Classical	Factor Analysis	Rasch
2	X	X	X	2		X	X	2	X	X	
3		X	X	3			X	3	X		X
5	X	X	X	4			X	5	X	X	
6	X	X		5	X	X		6	X		X
7	X	X		6	X	X		7	X	X	
8			X	7	X	X		8			X
9		X	X	9		X		9	X	X	
10	X	X		10	X	X		10	X	X	
11	X			11	X	X		11	X	X	X
12		X		12	X	X	X	12	X	X	
13	X	X		13	X	X		13	X	X	
14	X	X	X	14	X	X		14	X	X	X
15		X	X	16			X	15			X
17	X		X	17	X		X	17		X	X
18	X	X	X	18	X	X	X	18		X	
19	X	X		19	X	X		19	X	X	
20	X	X	X	20	X	X	X	20	X	X	X
26	X		X	22			X	24	X	X	X
27	X		X	24	X		X	26			X
28	X		X	26		X	X	27	X	X	X
29		X		27		X	X	28			X
34			X	28			X	29			X
36			X	29	X			31			X
				36			X	47			X
				39			X				
				40	X						

[a]"X" indicates that the item was selected.

TABLE 10

THE BEST 30 ITEMS SELECTED UNDER EACH ITEM ANALYTIC PROCEDURE ACCORDING TO SAMPLE SIZE[a]

N = 250				N = 500				N = 995			
Item	Classical	Factor Analysis	Rasch	Item	Classical	Factor Analysis	Rasch	Item	Classical	Factor Analysis	Rasch
1	X			1			X	1	X	X	X
2	X	X	X	2	X	X	X	2	X	X	
3	X	X	X	3	X	X	X	3	X	X	X
4		X		4			X	5	X	X	X
5	X	X	X	5	X	X	X	6	X	X	X
6	X	X	X	6	X	X	X	7	X	X	X
7	X	X	X	7	X	X		8		X	X
8	X	X	X	8	X	X	X	9	X	X	X
9		X	X	9	X	X	X	10	X	X	
10	X	X		10	X	X	X	11	X	X	X
11	X	X	X	11	X	X	X	12	X	X	X
12	X	X	X	12	X	X	X	13	X	X	X
13	X	X	X	13	X	X	X	14	X	X	X
14	X	X	X	14	X	X	X	15	X	X	X
15	X	X	X	15	X	X	X	16		X	X
16		X	X	16			X	17	X	X	X
17	X	X	X	17	X	X	X	18	X	X	X
18	X	X	X	18	X	X	X	19	X	X	X
19	X	X		19	X	X		20	X	X	X
20	X	X	X	20	X	X	X	22	X	X	X
22	X	X	X	22	X	X	X	23			X
23	X		X	24	X	X	X	24	X	X	X
24	X	X	X	25	X	X	X	25	X	X	
25		X		26	X	X	X	26	X	X	X

TABLE 10 - Continued

N = 250				N = 500				N = 995			
Item	Classical	Factor Analysis	Rasch	Item	Classical	Factor Analysis	Rasch	Item	Classical	Factor Analysis	Rasch
26	X	X	X	27	X	X		27	X	X	X
27	X	X	X	28	X	X	X	28	X	X	X
28	X	X	X	29	X	X	X	29	X	X	X
29	X	X	X	31	X	X	X	31	X	X	X
31	X		X	34	X	X	X	34			X
34	X		X	35	X			35		X	X
36		X	X	36	X		X	37	X		
37			X	37			X	39	X	X	X
39		X		39	X	X	X	40	X	X	X
40	X	X		40	X	X		44	X		
42	X		X	44		X		47			X
43	X			47		X	X	48	X		
44		X	X								
48	X		X								

[a]"X" indicates that the item was selected.

The selection of items under factor analysis was determined by the item loadings on the first unrotated principal component. The 15 and 30 items having the highest item-component biserial correlation were selected.

The selection of the 15 and 30 items from the Rasch analysis was determined by the mean square fit of the item to the Rasch model. The closer the mean square fit was to zero the better the item fit the model, thus items with the lowest mean square fit statistic were selected.

Double Cross-Validation

After the tests of the best 15 and 30 items were developed by each procedure, they were scored on independent samples, in a double cross-validation procedure as noted in Table 3, Chapter III. Item and test statistics, needed to test the five hypotheses were obtained for the 15 and 30 item tests based on the cross-validation samples using the GITAP program (Baker and Martin, 1970).

The GITAP program provided the following output:

- each subject's total test score
- test mean and standard deviation
- internal consistency estimates as measured by Hoyt's analysis of variance procedure
- estimates of the standard error of measurement
- indices of item difficulty and biserial correlations

Comparison of the 15 Item Tests on Precision

The descriptive statistics based on the double cross-validation samples for the 15 item tests have been presented in Table 11.

TABLE 11

DESCRIPTIVE STATISTICS FOR THE TEST COMPOSED OF THE 15 BEST ITEMS SELECTED BY EACH ITEM ANALYTIC PROCEDURE ACCORDING TO SAMPLE SIZE

	N = 250			N = 500			N = 995		
	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch
Test:									
Mean	9.80	10.40	9.41	9.27	10.21	8.88	10.00	9.72	8.76
Standard Deviation	3.41	3.41	3.24	3.69	3.50	3.18	3.63	3.62	3.22
Internal Consistency[a]	.790	.800	.754	.816	.811	.724	.826	.814	.735
Standard Error of Measurement	1.51	1.47	1.55	1.53	1.47	1.61	1.46	1.51	1.60
Difficulty:									
Mean	.65	.69	.63	.62	.68	.60	.67	.65	.58
Standard Deviation	.16	.14	.17	.15	.15	.16	.15	.14	.17
Range	.31-.89	.40-.92	.32-.90	.32-.83	.35-.92	.36-.89	.33-.91	.37-.90	.27-.81
Biserial[b]:									
Mean	50.93	56.13	46.53	53.53	53.53	41.20	56.33	53.80	42.80
Standard Deviation	14.52	17.05	8.94	12.45	8.20	7.95	9.19	9.44	7.26
Range	39-82	34-91	37-64	29-78	39-66	28-56	41-73	38-70	30-56

[a] Internal consistencies reported to three digits to observe slight differences.

[b] Based on linear transformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946).

The values of the internal consistency estimates for the tests developed using the Rasch model were consistently lower than the internal consistency estimates of the tests developed by classical item analysis and factor analysis across all sample sizes.

The observed internal consistency estimates were tested for significance using confidence intervals described by Feldt (1965), to see if they were statistically different from the internal consistency estimate for the projected population using the Spearman-Brown Prophecy Formula.

The internal consistency estimate for the population based on the original 50 item subtest was .88 (Table 1). By applying the Spearman-Brown Prophecy Formula (Mehrens and Lehman, 1973) the projected population internal consistency estimate for a 15 item test was found to be .687. The value .687 was the expected internal consistency if 35 of the 50 items were randomly deleted. Thus, confidence intervals were generated around the observed internal consistency estimates, presented in Table 11, for each procedure across all sample sizes to see if any of the three item analytic techniques would produce a more reliable test than would be expected from mere random item deletion.

The confidence intervals for the observed consistency estimates for each procedure have been reported in Table 12.

When the sample sizes were 250 and 995 each item analytic technique produced an internal consistency estimate that was significantly different from the projected population estimate (.687) at a confidence level of 95 percent. Each of the three techniques systematically retained the 15 most homogeneous items. These tests were more precise in terms of internal consistency than would have been found if the

items were randomly deleted as noted by comparisons to the projected population reliability coefficient.

Table 12

CONFIDENCE INTERVALS[a] FOR THE OBSERVED INTERNAL CONSISTENCY ESTIMATES BASED ON THE 15 ITEM TESTS ACCORDING TO SAMPLE SIZE

	95% Confidence Interval		
Procedure	N = 250	N = 500	N = 995
Classical	.748 - .828*	.792 - .838*	.810 - .841*
Factor Analysis	.760 - .837*	.786 - .854*	.797 - .831*
Rasch	.704 - .799*	.688 - .757	.711 - .759*

[a]The F values used in calculating the confidence intervals were obtained from Marisculo (1971).

*Statistical significance is indicated when the population internal consistency estimate is not concluded in the confidence interval generated for each observed internal consistency estimate. The projected population value was .687.

Only two procedures produced tests with internal consistency estimates significantly different from the projected population estimate when the sample size was 500, classical item analysis and factor analysis.

As sample size decreased, in most cases, the internal consistency for each method tended to decrease (Table 11). An exception was noted for the Rasch tests, when the sample size decreased from 500 to 250, internal consistency improved slightly.

The data reported in Table 11 indicated that the standard error of measurement for the 15 item tests based on the Rasch model were

consistently larger than the standard error of measurement of the tests developed from classical item analysis and factor analysis for each sample size. However, these differences were not meaningful in that the difference did not equal or exceed 1.00 for any of the three procedures.

The differences in mean item difficulties and discriminations were tested for statistical significance to determine whether there were differences in the types of items retained by each item analytic method. In this study item discriminations were measured by biserial correlations. A two-way analysis of variance (fixed effects model) was performed separately for the two dependent variables of item difficulty and item discrimination. A check was made on the assumptions for the analysis of variance to ensure that they were met. In these analyses, item analytic technique and sample size were the two independent factors, each with three levels.

For item difficulty, no significant differences were found for item analytic technique, sample size, or their interaction, F (2,126) = 2.57, $p > .05$; F (2,126) = .45, $p > .05$; F (4,126) = .33, $p > .05$ respectively.

The means, standard deviations, and ranges of the item difficulties based upon the 15 item tests have been reported in Table 13.

For the analysis of variance performed on the transformed biserial correlations a significant F ratio was observed for the factor of item analytic technique, F (2,126) = 14.862, $p < .05$. No significant differences were observed for sample size or the interaction of sample size and item analytic technique for the transformed biserial

correlations [9] F (2,126) = .30, $p > .05$; F (4,126) = 1.16, $p > .05$ respectively. The means, standard deviations, and ranges of the transformed biserial correlations based upon the 15 item tests have been presented in Table 14.

TABLE 13

15 ITEM TESTS:
DESCRIPTIVE STATISTICS FOR ITEM DIFFICULTY
BY PROCEDURE AND SAMPLE SIZE

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	.65	.67	.60	.66	.63	.63
Standard Deviation	.15	.14	.16	.15	.15	.16
Range	.31-.91	.35-.92	.27-.90	.31-.92	.32-.92	.27-.91

Post hoc comparisons were made to determine which of the three item analytic procedures based upon their means contributed to the significant F ratio for the transformed biserial correlations. Tukey's HSD (honestly significant difference) test for multiple comparisons was employed (Kirk, 1968, p. 88). The HSD value ($\alpha = .01$), was 6.88. Therefore, a difference between means had to exceed this value to be significantly different. The results of the post hoc comparisons

[9] When the actual biserial correlations were tested in the two-way analysis of variance design similar F ratios were observed.

between the mean item discriminations have been reported in Table 15.

TABLE 14

15 ITEM TESTS:
DESCRIPTIVE STATISTICS FOR ITEM DISCRIMINATIONS[a]
BY PROCEDURE AND SAMPLE SIZE

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	53.60	54.49	43.51	51.20	49.47	50.98
Standard Deviation	12.18	11.98	8.21	14.17	11.20	10.36
Range	29-82	34-91	28-64	34-91	28-78	30-73

[a]Based on transformed biserial correlations. The transformation was a linear transformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946).

From Table 15, it is apparent that the mean transformed biserial correlation from the Rasch developed test was significantly lower than the mean biserial correlations from the tests developed by classical item analysis and factor analysis.

Comparison of the 30 Item Tests on Precision

The descriptive statistics based on the double cross-validation of the 30 items selected by each procedure, according to sample size, have been presented in Table 16.

TABLE 15

POST HOC COMPARISONS OF THE DIFFERENCES BETWEEN THE MEAN ITEM DISCRIMINATIONS[a] FOR THE 15 ITEM TESTS

	Means		
	54.49	53.60	43.51
Factor Analysis (54.49)	-----	.889	10.98**
Classical Item Analysis (53.60)		-----	10.09**
Rasch Analysis (43.51)			-----

[a]Based on transformed biserial correlations. The transformation was a linear transformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1846).

** $p < .01$, HSD = 6.88.

By increasing the test length to 30 items, the internal consistency estimate was increased across each method and sample size, but a pattern similar to that for the 15 item test emerged. The internal consistency estimates from the test based on the Rasch model were slightly lower than the internal consistency estimates for the tests based on classical item analysis and factor analysis. The observed internal consistency estimates were tested for significance, using the confidence intervals described in the previous section, to see if they were statistically different from the internal consistency estimate for the population.

The projected population internal consistency estimate for a 30 item test was found to be .814 (via the Spearman-Brown Prophecy Formula).

TABLE 16

DESCRIPTIVE STATISTICS FOR THE TEST COMPOSED OF THE 30 BEST ITEMS SELECTED BY EACH ITEM ANALYTIC PROCEDURE ACCORDING TO SAMPLE SIZE

	N = 250			N = 500			N = 995		
	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch	Classical	Factor Analysis	Rasch
Test:									
Mean	17.12	17.98	16.99	17.69	17.73	18.24	17.31	18.51	17.54
Standard Deviation	5.73	6.05	5.68	6.30	6.06	5.80	6.22	6.22	6.11
Internal Consistency[a]	.834	.854	.828	.863	.853	.838	.862	.861	.851
Standard Error of Measurement	2.29	2.27	2.32	2.29	2.27	2.30	2.27	2.28	2.31
Difficulty:									
Mean	.57	.60	.57	.59	.59	.61	.58	.62	.59
Standard Deviation	.19	.17	.18	.16	.17	.16	.17	.15	.16
Range	.23-.89	.29-.92	.23-.90	.32-.89	.31-.92	.24-.89	.21-.91	.35-.90	.27-.85
Biserial[b]:									
Mean	38.67	42.17	37.43	42.30	40.63	38.17	42.10	42.03	40.07
Standard Deviation	13.30	13.13	10.67	11.19	9.33	8.82	10.85	10.03	10.34
Range	9-70	10-68	13-57	24-68	21-58	23-55	25-66	28-61	24-64

[a] Internal consistencies reported to three digits to observe slight differences.

[b] Based on linear transformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946).

The value of .814 indicated the expected internal consistency if 20 of the 50 items were randomly deleted.

Based on the observed internal consistency estimates reported in Table 16, confidence intervals were generated for each item analytic procedure and have been presented in Table 17.

TABLE 17

CONFIDENCE INTERVALS[a] FOR THE OBSERVED INTERNAL CONSISTENCY ESTIMATES BASED ON THE 30 ITEM TESTS ACCORDING TO SAMPLE SIZE

	95% Confidence Interval		
	N = 250	N = 500	N = 955
Classical	.800-.865	.845-.880*	.850-.874*
Factor Analysis	.825-.881*	.834-.871*	.848-.873*
Rasch	.794-.860	.817-.858*	.838-.863*

[a]The F values used in calculating the confidence intervals were obtained from Marisculo (1971).

*Statistical significance is observed when the population internal consistency estimate is not included in the confidence interval generated for each observed internal consistency estimate. The projected population value was .814.

For the sample of 250 examinees, only one item analytic technique (factor analysis) produced an internal consistency estimate that was statistically different from the projected population estimate at a confidence level of 95 percent.

However, all three techniques produced tests with internal consistency estimates significantly different from the projected population

estimate when the sample was increased to 500, and 995. Thus, when the number of examinees was large, each of the three techniques procedures tests with higher internal consistency estimates than if the test were produced by randomly deleting items.

For the 30 item tests, the effect of decreasing the sample size tended to decrease internal consistency for each method (Table 16). But the decrease was very slight.

The standard error of measurement was essentially the same for the three methods of item analysis across the varying sample sizes.

Two-way analyses of variance were run on item difficulties and item discriminations for the 30 item tests, similar to those run for the 15 item tests. Again, the independent variables were item analytic technique and sample size, each containing three levels.

No significant differences were observed for item difficulty for the independent variables of item analytic technique, sample size, or their interaction, F (2,261) = .46, $p > .05$; F (2,261 = .27, $p > .05$; F (4,261) = .24, $p > .05$ respectively.

No significant differences were observed for the transformed biserial correlations[10] for the independent variables of item analytic technqiue, sample size, or their interaction, F (2,261) = 1.97, $p > .05$; F (2,261) = .74, $p > .05$; F (4,261) = .48, $p > .05$ respectively.

---

[10]When the actual biserial correlations were tested in the two-way analysis of variance design similar F values were observed.

The means, standard deviations, and ranges of the item difficulties and transformed biserial correlations based upon the 30 item tests have been presented in Tables 18 and 19 respectively.

TABLE 18

30 ITEM TESTS: DESCRIPTIVE STATISTICS FOR ITEM DIFFICULTY BY PROCEDURE AND SAMPLE SIZE

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	.58	.60	.59	.58	.60	.59
Standard Deviation	.17	.16	.17	.18	.16	.16
Range	.21-.91	.29-.92	.23-.90	.23-.92	.24-.92	.21-.91

TABLE 19

30 ITEM TESTS: DESCRIPTIVE STATISTICS FOR ITEM DISCRIMINATIONS[a] BY PROCEDURE AND SAMPLE SIZE

	Procedure			Sample Size		
	Classical	Factor Analysis	Rasch	250	500	995
Mean	41.02	41.61	38.56	39.42	40.37	41.40
Standard Deviation	11.82	10.85	9.93	12.45	9.87	10.34
Range	9-70	19-68	13-64	9-70	21-68	24-66

[a]Based on transformed biserial correlations. The transformation was a linear transformation of the Fisher z statistic and multiplication of the constant 60.241 providing a range of 0-100 for the biserial correlation (Davis, 1946).

Comparison of the 30 Item Tests on Efficiency

Lord (1974a, 1974b) proposed the formula used for approximating the relative efficiency for two tests, stated previously in equation 13 as:

$$\text{R.E. } (y,x) = \frac{n_y}{n_x} \cdot \frac{x(n_x - x)f_x^2}{y(n_y - y)f_y^2},$$

where R.E. (y,x) denotes the relative efficiency of y compared to x, n_x and n_y are the numbers of items in the two tests, x and y are the number-right scores having the same percentile rank, and f_x^2 and f_y^2 are the squared observed frequencies of x and y obtained from frequency distributions for similar groups of examinees. A careful examination of the formula for relative efficiency indicated that when $n_x = n_y$ and $x = y$, that it was the number of examinees at the specified ability level (f_x^2 and f_y^2) that determined the efficiency of the test. That is, the fewer examinees observed at a particular percentile rank, the better the test discriminates at that percentile rank. Therefore, test efficiency was equated with the level of discrimination the test was able to make between examinees, at various scores or percentile ranks.

Three relative efficiency comparisons were made using the 30 item tests based on the sample of 995 examinees. The three comparisons were: (a) the test developed from factor analysis was compared to the test developed by classical item analysis (b) the test developed by Rasch analysis was compared to the test developed by classical item analysis, and (c) the test developed from the Rasch analysis was compared to the factor analytically developed test.

The efficiency curves for the three comparisons were shown in Figure 2. The relative efficiency value was plotted on the ordinate, while the percentile rank (student ability level) was plotted along the abscissa. Computed values for the relative efficiency comparisons have been reported in Appendix B. A relative efficiency of 1.00 would indicate that the tests are equally efficient.

The test developed by factor analysis was more efficient for the lower tenth of the pupils when compared to the test developed from classical item analysis. Both the tests were about equally efficient for the middle ability groups and high ability groups.

The Rasch developed test was more efficient than the test based on classical item analysis for average to high ability students (40th-90th percentile rank). However, it was less efficient than the classical item analysis test for students with very low or very high abilities (1st-20th percentile rank and 98th percentile rank).

When compared to the factorially developed test, the Rasch test was again more efficient for students of average to high abilities (50th-90th percentile rank). The factorially developed test appeared more efficient for the very low and very high ability students (1st-20th percentile rank and 98th percentile rank).

Summary

The results reported in this chapter are summarized for each of the five hypotheses.

Hypothesis 1. There are no significant differences in the internal consistency estimates of the tests produced by the three methods, as the number of items decreases, when compared to the

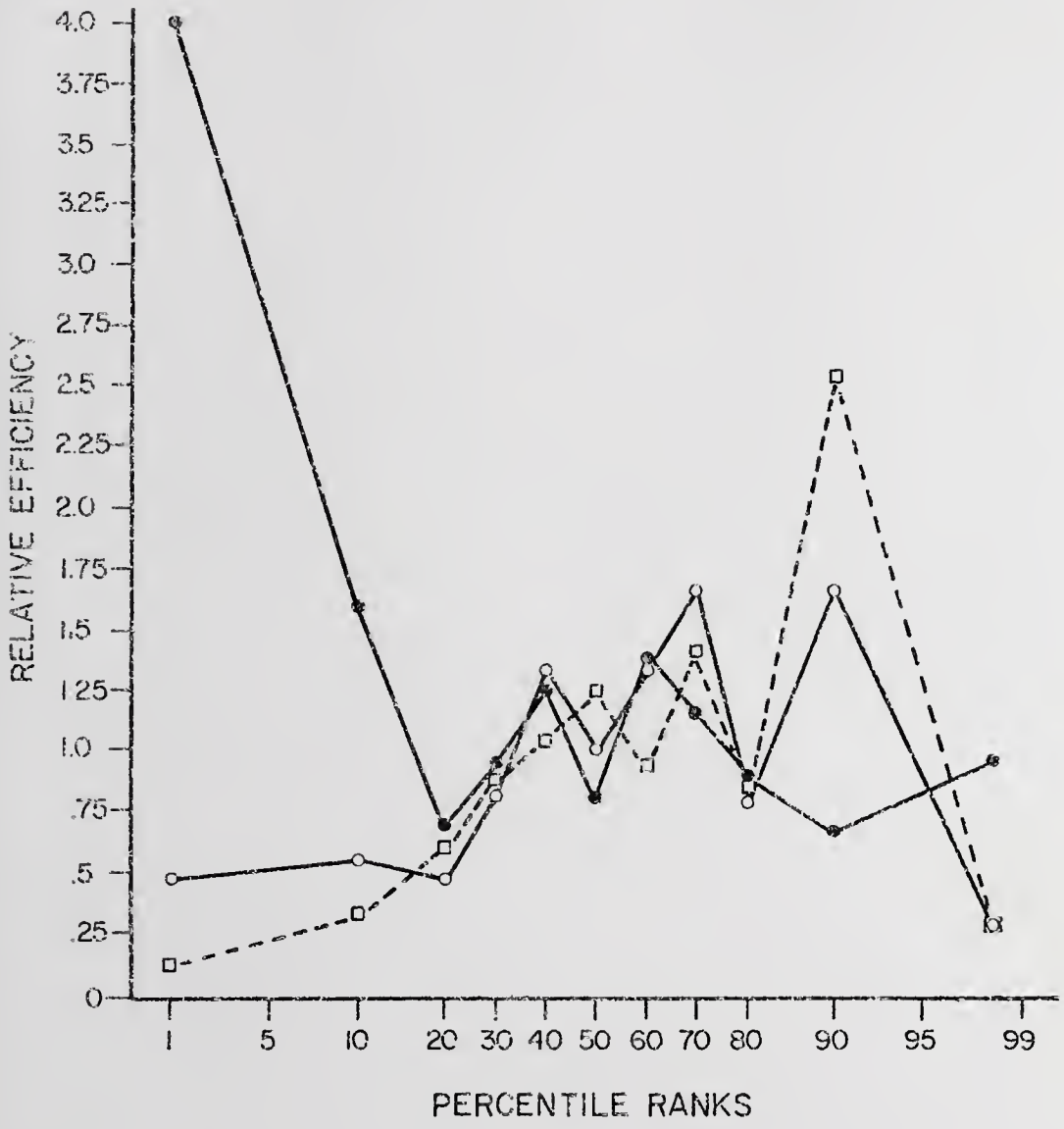


FIGURE 2. RELATIVE EFFICIENCY COMPARISONS FOR THE THREE 30 ITEM TESTS N=995.

projected internal consistency estimates for the population for tests of similar length.

Confidence intervals were calculated to test for differences between the observed internal consistency estimates and the internal consistency estimate for the population. As reported in Tables 12 and 17, for the 15 and 30 item tests, 15 of the 18 confidence intervals (at the 95 percent level) generated around the sample estimate did not contain the population value. This means that 15 of the observed internal consistency estimates were superior to the population values projected for subtests of similar length created by random deletion of items. Therefore, hypothesis one was not supported. The procedures that produced the three observed internal consistency estimates that were not significantly different from the population value, and hence no different than would be expected by random item deletion, were the Rasch procedure (15 item test, N = 500; 30 item test, N = 250) and the classical item analysis procedure (30 item test, N = 250).

Hypothesis 2. There are no differences in the internal consistency estimates of the tests produced by the three methods when the number of examinees is decreased.

Hypothesis two was supported for the 15 and 30 item tests. Slight decreases in internal consistency estimates were noted for the 15 item test (Table 11) as sample size decreased, but only decreases of one or two one-hundreths of a point. Even smaller decreases were observed on the 30 item test (Table 16).

Hypothesis 3. There are no meaningful differences in the magnitude of the standard error of measurement of the tests produced by the three methods.

Hypothesis three was supported for the 15 and 30 item tests. Meaningful differences were defined to be ≥ 1.00, but none of the three methods produced tests with standard errors of measurement that differed by that much. In each case, the difference was approximately one-tenth of a point or less (Tables 11 and 16).

Hypothesis 4. There are no differences in the difficulties or discriminations of the items selected by the three methods.

Hypothesis four was supported for the 15 and 30 item tests with respect to item difficulty. That is, the two-way analysis of variance revealed no significant differences for either the 15 or 30 item tests with regard to item difficulty.

Hypothesis four was also supported for item discrimination, but only for the 30 item tests. The two-way analysis of variance for item discrimination indicated no significant differences for the 30 item tests; however, on the 15 item tests, a significant F ratio ($p < .05$) for item analytic procedure was observed for item discrimination. Tukey's HSD test revealed that items selected by the Rasch procedure had significantly lower average biserial correlations than the items selected by factor analysis and classical item analysis (Table 15). This could have been expected because the range of the biserial correlation was restricted when the items were originally selected for the Rasch model. This procedure was necessary to meet one of the assumptions for the Rasch model.

Hypothesis 5. There are no differences across ability levels in the efficiency of the tests produced by the three methods.

Hypothesis five was not supported. The efficiency curves illustrated in Figure 2, generally indicated that the tests based on classical test theory were more effective for measuring students with very low ability (20th percentile rank or less) and students with very high abilities (98th percentile rank). The Rasch developed test was most efficient for assessing average and high ability students (40th-90th percentile rank).

CHAPTER V

DISCUSSION AND CONCLUSIONS

This study was conducted to determine which of the three item analytic procedures (classical item analysis, factor analysis, and the Rasch model) might produce the superior test in terms of the precision and the efficiency of measurement. A common item and examinee population was used to test five hypotheses. Of the five hypotheses, three dealt with elements of test precision as measured by internal consistency estimates. Another hypothesis treated the issue of item discriminations. Thus, it too was related to internal consistency. The fifth hypothesis focused on the relative efficiency of the tests produced by three item analytic techniques. This hypothesis altered the emphasis of the study from one overall specific measure of a test's accuracy, in terms of internal consistency, to a general comparison of each method as a function of ability level. The discussion of the results then has been focused in two major areas: (a) the precision of the tests, and (b) the efficiency of the tests produced by the three methods of item analysis.

The Precision of the Tests Produced by the Three Methods of Item Analysis

Each of the three item analytic techniques was applied to an independent sample to select the best 15 and 30 items. The stability of the summary statistics across each sample size for the three item analytic techniques indicated a tendency for the nine samples to be very homogeneous.

The similarity of the means, standard deviations, and percentages of variance accounted for were noted on Tables 5-8, with the exception of the mean square fit statistic (Table 7) which increased with sample size. (This exception is discussed later in this chapter.) From these samples, items were selected by each item analytic technique to maximize internal consistency.

The data reported in Tables 11 and 16 indicated the effectiveness of each item analytic technique in producing internally consistent tests. Before an overall decision can be made as to the superiority of one technique over another, each of the hypotheses relating to precision must be considered.

Internal Consistency

Data in Tables 11 and 16 indicate that the two tests based on classical test theory (factor analysis and classical item analysis) appeared superior in terms of internal consistency when compared to the tests developed by the Rasch model.

To test whether any of the three methods produced tests with greater internal consistency than a test created by random item deletion, the internal consistency estimates were compared to the projected internal consistency value for the population by using confidence intervals as suggested by Feldt (1965). In order for a given sample internal consistency estimate to be significant, the population value could not be included in the confidence interval generated around that sample value. For the 15 item tests, nine confidence intervals were calculated for the nine estimates of internal consistency, one for each method at each sample size. Eight of the nine sample values were shown to be significantly greater than the population estimate at the 95 percent confidence level (Table 12). Only the internal consistency estimate of the Rasch test,

based on the sample of 500 examinees, failed to reach a level significantly greater than would have been expected by chance.

For the 30 item tests, nine confidence intervals were also calculated for the nine estimates of internal consistency, one for each method at each sample size. Seven of the nine sample internal consistency estimates were shown to be significantly greater than the population estimate at the 95 percent confidence level (Table 17). The tests based on classical item analysis and Rasch analysis, for the sample of 250 examinees, were not significantly different from the projected population value for a 30 item test created by random item selection. Therefore, for smaller samples (N = 250) factor analysis appeared to be superior to classical item analysis and the Rasch analysis in producing the most precise test.

Generally, as the number of examinees decreased so did the internal consistency estimates. However, the tests based on factor analysis were least affected by decreasing the sample sizes used in the cross-validation for the 15 and 30 item tests (Tables 11 and 16).

Standard Error of Measurement

The standard error of measurement is the standard deviation of the distribution of errors surrounding an individual's observed score on an infinite number of parallel tests. Hence the smaller the standard error of measurement, the greater the precision of the measurement. This statistic is often considered a more meaningful measure of an instrument's reliability than the reliability coefficient itself (Magnusson, 1966, p. 82). Based on the data for this study, the standard errors of measurement were consistently smaller for both the 15 and 30 item tests

produced by classical test theory as compared to the 15 and 30 item tests based on the Rasch model; however, the differences in the standard errors of measurement did not equal or exceed 1.00 for any of the methods.

Types of Items Retained

Item difficulty. Item difficulties of the 15 and 30 item tests were analyzed in separate two-way analyses of variance. The two independent variables were sample size and item analytic technique. No significant F ratios were observed for either the 15 or 30 item tests on item difficulty. Therefore, each item analytic technique tended to select items which had similar item difficulties on the average.

Item discrimination. In this study, the item discriminations were measured by biserial correlation. Transformed biserial correlations for the 15 and 30 item tests were analyzed in separate two-way analyses of variance. The two independent variables were sample size and item analytic technique. For the 15 item tests, a significant F ratio ($p < .05$) was observed for the main effect of item analytic technique. The mean biserial correlation for each 15 item tests were 44 for the Rasch test, 54 for the classical item analysis test, and 54 for the factorially developed test.[11] Tukey's HSD post hoc comparison indicated that the items selected by the Rasch procedure had lower biserial correlations, on the average, than items selected on the basis

---

[11]The actual mean biserial correlations for the three tests corresponding to the transformed biserial correlations were: .62, .71, .71 respectively.

of factor analysis or classical item analysis. It should be noted that this finding was due to the fact that the range of the biserial correlations was restricted to .39 - .79 on the items selected for the Rasch calibration. This was necessary to meet the assumption of equal item discriminations.

However, when the test length was increased to 30 items, no significant F ratios were observed for the variable of item discrimination. The difference in these two findings for the 15 and 30 item tests can be explained by the way the 15 and 30 item tests were constructed. The 15 item test was made up of the 15 items with the highest biserial correlations. The 30 item test was made up of the above 15 items and an additional set of 15 items with the next highest biserial correlations. The addition of 15 more items meant that their average biserial correlation was something less than the original 15 items. Hence, the mean biserial correlations were reduced for the longer 30 item tests (Table 19).

Conclusions

From the data presented for each of the four areas above, it was concluded that each of the three item analytic techniques tended to produce tests that were really no different in terms of the precision of measurement.

Thus, the question to consider now is: Should practitioners in the field of measurement spend their time learning to use the Rasch model to develop cognitive norm-referenced tests[12] knowing the extra

---

[12]The field of test development is limited to cognitive norm-referenced tests because that was the type of instrument used in this study.

work and sophistication of knowledge required to effectively use the Rasch procedures? With the criterion of internal consistency as a measure of test superiority, it appeared from this study that time spent factorially developing tests, or if computer facilities were not available, the use of classical item analysis procedures seem more than adequate for good test construction.

However, it must be remembered that internal consistency may not be a fair and sufficient criterion. Internal consistency is an integral part of classical test theory and may be biased since it was derived from the classical model. Following that reasoning, Whitely and Dawis (1974) have commented on the precision of tests developed using classical item analysis and Rasch analysis. They stated that if the goal of item selection was to develop fixed-content tests, then the classical techniques of item selection will yield the more precise test since precision is specific to the trait distribution in a given test. Whitely and Dawis indicated that the strength of the Rasch analysis was in the individualized selection of items, as in tailored testing, rather than the construction of fixed-content tests.

In considering the above situation, Lord (1974b) has stated that internal consistency is an overall estimate of a test's homogeneity, but provides no information on how the test as a whole discriminates for the various ability groups taking the test. Thus, the three techniques of test development were compared using an additional criterion, relative efficiency.

The Efficiency of the Tests Produced by the Three Methods of Item Analysis

The review of the literature concerning the relative efficiency of a test presented in Chapter II, cited studies mainly dealing with latent trait theory (Birnbaum, 1968; Hambleton and Traub, 1971, 1973). Studies comparing the relative efficiency of tests developed by latent trait theory to classical test theory appear to be missing from the literature on test efficiency.

The relative efficiency formula (Lord, 1974a, 1974b) was not derived for any specific test development theory; therefore, relative efficiency estimates should be applicable to any test development technique. Lord (1974b) suggested his formula may not work well for extremely short tests and that it should only be used on large samples of examinees. Thus, in the present study, only the 30 item tests were compared using the sample of 995 examinees. The three comparisons of relative efficiency were: (a) the test based on factor analysis was compared to the test based on classical item analysis, (b) the test based on the Rasch analysis was compared to the test based on classical item analysis, and (c) the test based on the Rasch analysis was compared to the test based on factor analysis.

Generally, the results indicated that the Rasch test was superior to the two tests based on classical test theory for students of average and high ability. The two tests based on classical test theory, however, were superior in efficiency to the Rasch developed test for very low and very high ability students (Figure 2). Test efficiency has been defined as a measure of how well a test is able to discriminate between examinees of varying abilities. Therefore, the test constructor must

ask himself, for which segment(s) of the examinee population is the test intended to discriminate?

In this study, the test under consideration was a verbal aptitude college admissions test. Usually college admissions officers are interested in selecting students who will be successful once admitted to college. The examinees who score very high on college admissions tests will generally be admitted to college without any question. Thus, it is less important to be able to discriminate among the very high scoring examinees than to discriminate among the students who score near the mean or in the upper middle range on a college admissions test. For these students it is difficult to decide who should be admitted and who should be denied admittance. If it is known that the admissions test discriminates very well for average to high ability students, then the reliability of the selection process based on test scores should be increased. The data in this study, therefore, indicate that the test based on the Rasch analysis would be most efficient for selecting the average to high ability students for admission to college.

Lord (1968) has illustrated a very important feature of test information and relative efficiency curves. He has shown that the contribution of each item to a test is independent of all other items. Thus, when information curves are available on a pool of items they can be added to a test to achieve a prespecified information or relative efficiency curve for any subpopulation of examinees (Lord, 1968). Therefore, by using measures such as relative efficiency and information curves psychometricians are able to develop very discriminating tests for any segment of the population.

Conclusions

It has been suggested that because efficiency and test information curves are a function of ability, these estimates ought to replace the use of classical reliability estimates and the standard error of measurement in test score information (Hambleton et al., 1977). This suggestion certainly deserves some consideration in light of the present study where it was shown that the three methods of item analysis produced similar tests in terms of precision, but the three methods produced very different tests in terms of efficiency. Today, with the increasing use of computers in test construction, perhaps the more meaningful question to be asked by psychometricians is: For which ability group is the test superior? Only test information curves and measures of relative efficiency can answer that question.

Implications for Future Research

The results of this empirical study revealed that tests developed using classical test theory, in spite of its inherent weaknesses, were no different with respect to precision of measurement than tests developed using one of the latent trait models, the one-parameter Rasch model. Comparisons of relative efficiency for the 30 item tests showed that the tests based on classical test theory were superior to the Rasch developed test for very low and very high scoring examinees, and the Rasch developed test was more efficient for average to high scoring examinees on the verbal aptitude college admissions subtest used in this study.

Only one of the four latent trait models, the Rasch model, was used in this comparative study of test development techniques. Perhaps it was the very nature of this simple model that resulted in the

development of equivalent tests when compared to the tests developed by classical item analysis and factor analysis in terms of precision. It may be that the more technical two- and three-parameter logistic models would have produced tests comparable or superior to those developed by classical test theory in terms of precision of measurement and overall relative efficiency.

The two-parameter model allows for varying item discriminations so that the initial selection of items would not have to have been restricted to a prespecified range. The three-parameter model not only allows for varying item discriminations, but also for the effects of guessing on the test. It is reasonable to suspect that guessing may have been a factor in the item scores for the type of cognitive test used in the present study. Thus, before the findings of the study can be generally accepted, replication is needed using not only other populations and other instruments, but also other latent trait models.

If other latent trait models are to be considered in addition to the Rasch model, several points need to be evaluated. The Rasch model is the only latent trait model that provides for the direct calibration of items and abilities based on unweighted "number right" scoring (Wright, 1977). The two- and three-parameter models require a more complex scoring system where the item response is weighted in order to estimate the discrimination and guessing parameters. The weighting is an iterative process that may never converge or stablize unless arbitrary boundaries are established (Wright, 1977, p. 104). Because of this complex scoring system, the two- and three-parameter logistic models are less efficient for parameter estimation than the one-parameter model in terms of computer time.

A further criticism of the two- and three-parameter logistic models has been offered by Wright (1977) concerning the additional item parameters. If item discrimination parameters and item guessing parameters are introduced into a theory of measurement, why not person parameters for sensitivity to difficult items, and inclination toward guessing? Wright questions whether psychometricians really need to make measurement theory so complex.

A final point to consider when comparing the latent trait models is that only the one-parameter Rasch model provides a ratio scale of measurement in terms of the calibrated item and ability scores (Hambleton et al., 1977). Success on a particular item is given by the product of the person's ability and the item's easiness. Thus, the person with no ability will have zero odds or probability of success on any item. The same logic applies to items with no easiness (zero difficulty) they cannot be solved. Thus, measurements made with Rasch calibrated items are on a ratio scale; and it is the ratio scale of measurement that leads to the concept of specific objectivity.

Therefore, it seems that each latent trait model has its own set of advantages and disadvantages and should be considered if comparisons are to be made to the classical models for test development purposes.

Additional areas for future research may lead to actual comparisons of the content of the items selected by each of the three item analytic techniques. Davis (1951) and Cox (1965) have criticized the selection of items solely on statistical criteria. The use of statistical criteria alone, may result in changing the nature of the trait being measured by deleting the items essential to adequate content coverage. Whitely

and Dawis (1976) found this to be true in a study of verbal analogy items where the type of relationship and content of specific analogy items proved to be quite significant when studied in isolation.

A restricting factor in this study was that a prespecified number of test items were selected, e.g., 15 and 30, by each item analytic procedure. For the Rasch procedure, perhaps some number of items less than 30, but greater than 15 may have provided a better fit to the model. This procedure might have produced mean square fit statistics that were equivalent across the sample sizes rather than increase with sample size as found in this study. Thus, selecting the number of items precisely fitting the Rasch model and comparing that number of items with classical item analysis and factor analysis might have led to different conclusions than those made in the present study with regard to the precision and relative efficiency of measurement.

CHAPTER VI

SUMMARY

The quality of the items in a test determine its validity and reliability. Through the application of item analysis procedures, test constructors are able to obtain quantitative objective information useful in developing and judging the quality of a test and its items.

Classical test theory forms the basis for one method of test development. An integral part of the development of tests based on the classical model is the utilization of classical item analysis or factor analysis. Classical item analysis is a procedure to obtain a description of the statistical characteristics of each item in the test. This approach requires identification of single items which provide maximum discrimination between individuals on the latent trait being measured. Theoretically, selecting items which have a high correlation with total test score will result in a discriminating test which is homogeneous with respect to the latent trait. Therefore, classical item analysis is an aid to developing internally consistent tests.

An alternative method of test development, but based on the classical model, is factor analysis. Factor analysis is a more complex test development procedure than classical item analysis. It is a statistical technique that takes into account the item correlation with

all other individual items in the test simultaneously. Groups of similar items tend to cluster together and comprise the latent traits (factors) underlying the test. Thus, under the classical model then, classical item analysis can be viewed as a unidimensional basis for item analysis, less sophisticated than the multidimensional procedure of factor analysis.

Classical item analysis and factor analysis have long been the only techniques described in measurement texts for use in test development (Baker, 1977). However, with the publication of Lord and Novick's Statistical Theories of Mental Test Scores, (1968) considerable attention is being directed now toward the field of latent trait theory as a new area in test development. Proponents of this approach claim that the advantages of latent trait theory over classical test theory are twofold: (a) theoretically it provides item parameters that are invariant across examinee samples which will differ with respect to the latent trait, and (b) it provides item characteristic curves that give insight into how specific items discriminate between students of varying abilities.

Four latent trait models have been developed for use with dichotomously scored data: The normal ogive, and the one-, two-, and three-parameter logistic model (Hambleton and Cook, 1977; Lord and Novick, 1968). This study was concerned with the one-parameter logistic Rasch model because it is the simplest of the four models.

A review of the literature revealed numerous studies conducted in each of the three areas of item analysis, but relatively few comparative studies were reported between the three methods. Missing

from the review were comparative studies among all three, item analytic techniques. Therefore, the present study was designed to compare the methods of classical item analysis, factor analysis, and the Rasch model on measures of precision and relative efficiency in test development.

An empirical study was designed to compare the effects of the three methods of item analysis on test development across different sample sizes. Item response data were obtained from a sample of 5,235 high school seniors on a cognitive test of verbal aptitude.

The subjects were divided into 9 samples: three independent groups of 250 subjects each, three independent groups of 500 subjects each, and three independent groups of 995 subjects each. The independent groups were obtained so that tests of statistical significance could be performed.

The item response data were then analyzed in three phases. First, the "best" 15 and 30 items were selected using each item analytic technique. Under classical item analysis, the best 15 and 30 items were selected based on the highest biserial correlations. For factor analysis, the best 15 and 30 items were selected based on the highest item loadings on the first (unrotated) principal component. The selections of the best 15 and 30 items using the Rasch model were based upon the mean square fit of the items to the model. These procedures were used for each group of subjects. Second, a double cross-validation design was employed to obtain estimates on the item and test parameters for the best 15 and 30 items. The items selected from the three item analytic techniques were scored for different samples of subjects by randomly reassigning the samples which had been used in the original item analysis.

Then, the best 15 and 30 items chosen by each method were submitted to a common item analytic procedure in order to obtain estimates for comparing the three item analytic methods. Third, a two-way analysis of variance and a Tukey HSD post hoc comparison test, when indicated, were used to test for differences in the properties of items selected by each item analytic procedure. Also confidence intervals were calculated to compare the internal consistency estimates to a population value. In addition, the relative efficiencies of the 30 item tests developed by each item analytic technique were compared for the sample of 995 subjects.

The results of the analysis showed that there were no apparent differences in the types of tests produced by the three methods of item analysis in terms of the precision of measurement. The three methods were compared on measures of internal consistency, the standard error of measurement, mean item difficulty, and mean item discrimination.

Confidence intervals were generated around the observed internal consistency estimates for the 15 and 30 item tests produced by each method. The confidence intervals were obtained to compare the observed internal consistency estimates from each test to the project population internal consistency for tests of a similar length as suggested in Feldt (1965). The projected population value (obtained via Spearman-Brown Prophecy Formula) represented what the test's internal consistency would have been for a test created by deleting items at random. Of the 18 confidence intervals calculated at a 95 percent level of confidence, 15 did not contain the projected population value. Therefore, it was concluded that the three item analytic techniques were significantly different from random item deletion in producing tests with higher internal consistency estimates. It was also noted that as the

number of examinees decreased so did the internal consistency estimates.

The standard error of measurement was consistently smaller for both the 15 and 30 item tests produced by classical test theory when compared to the 15 and 30 item tests based on the Rasch model. However, the differences in the standard errors of measurement did not equal or exceed 1.00 for any of the methods.

No significant F ratios were observed for either 15 of 30 item tests on item difficulty. Therefore, each item analytic technique tended to select items which had similar item difficulties on the average.

For the variable, item discrimination on the 15 item tests, a significant F ratio ($p < .05$) was observed for the main effect of the item analytic technique. Tukey's HSD post hoc analysis indicated that the Rasch test tended to contain items with lower biserial correlations, on the average, than tests procedured by factor analysis and classical item analysis. This finding was probably due to the fact that the range of the biserial correlations was restricted to .39 - .79 for items retained in the Rasch analysis. However, when the test length was increased to 30 items, no significant F ratios were observed for the variable of item discrimination.

In terms of the test efficiency, the results indicated substantive differences in the tests produced by the three methods of item analysis. The 30 item test for the sample of 995 examinees was used in this comparison. It was found that the Rasch developed test was superior to the two tests based on classical test theory for students of average to high ability. The tests based on classical test theory, however, were superior in efficiency to the Rasch test for students of very low

or very high ability. In light of these findings, it was suggested that measures of test efficiency ought to be incorporated into the test development procedures, as it provides much more detailed information on how the test discriminates for various ability groups than does a single overall estimate of a test's homogeneity.

REFERENCES

Anderson, E. B. Goodness of fit for the Rasch model. Psychometrika, 1973, 38, 123-140.

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

Anderson, L. W. A comparison of classical item analytic precedures with affective data. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1976.

Baker, F. B. Empirical comparison of the item parameters based on the logistic and normal functions. Psychometrika, 1961, 26, 235-246.

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

Baker, F. B., and Martin, T. Fortap: A Fortran test analysis package. Occasional Paper No. 10, Office of Research Consultation, College of Education, Michigan State University, 1970.

Bedell, B. J. Determination of the optimum number of items to retain in a test measuring a single ability. Psychometrika, 1950, 15, 419-430.

Benson, I. G. The Florida Twelfth Grade Testing Program: A factor analytic study of the aptitude and achievement subtests. Unpublished Master of arts in education, thesis, University of Florida, 1975.

Binet, A. and Simon, T. (The development of intelligence in children) (E. S. Kite, trans.). Baltimore, Md.: Williams & Wilkins, 1916.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Ma.: Adison-Wesley, 1968.

Bock, R. D. and Wood, R. Test theory in P. H. Mussen and M. R. Rosenweig (Eds.), Annual review of psychology (Vol. 22). Palo Alto, Ca.: Annual Reviews, Inc., 1971

Brogden, H. E. Variation in test validity with variation in the test distribution of item difficulty, number of items, and degree of their intercorrelation. Psychometrika, 1946, 11, 197-214.

Burt, C. and John, E. A factorial analysis of the Terman-Binet tests. British Journal of Educational Psychology, 1943, 12, 156-161.

Carroll, J. B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Cattell, R. B. Personality and motivation, structure and measurement. New York: World Book, Inc., 1957.

Cattell, R. B. and Tsujioka, A. The importance of factor trueness and validity, versus homogeneity and orthogonality in test scales. Educational and Psychological Measurement, 1964, 24, 3-30.

Cochran, W. G. and Cox, G. M. Experimental Designs. (2nd ed.) New York: John Wiley and Sons, Inc., 1957.

Cox, R. C. Item selection techniques and evaluation of instructional objectives. Journal of Educational Measurement, 1965, 2, 181-185.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Davis, E. Item Analysis Data: Their computation, interpretation and use in test construction. Cambridge, Ma.: Harvard University, 1946.

Davis, E. Item selection techniques. In E. F. Lindquist (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1951.

Dinero, T. E. and Haertel, E. A computer simulation investigating the applicability of the Rasch model with varying item discrimination. A paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Fan, C. T. Item analysis table. Princeton, N. J.: Educational Testing Service, 1952.

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. Psychometrika, 1965, 30, 357-365.

Flanagan, J. General considerations in selection of test items and a short method of estimating the product moment coefficient from data at the tails of the distribution. Journal of Educational Psychology, 1939, 30, 674-680.

Florida Twelfth Grade Testing Program. (Report No. 1-75). Gainesville, Florida: Office of Instructional Resources, 1975.

Fruchter, B. Introduction to factor analysis. New York: D. Van Nostrand Co., Inc., 1954.

Guertin, W. and Bailey J. Introduction of modern factor analysis. Ann Arbor, Mich.: Edward Bros., Inc., 1970.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79-91

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, Inc., 1950.

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

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

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

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D., and Gifford, J. Developments in latent trait theory: A review of models, technical issues, and applications. A paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Harman, H. H. Modern factor analysis. (2nd ed.) Chicago: University of Chicago Press, 1967.

Henrysson, S. The relationship between factor loadings and biserial correlations in item analysis. Psychometrika, 1962, 27, 419-424.

Henrysson, S. Gathering, analyzing, and using data on test items. In E. L. Thorndike (Ed.), Educational Measurement. (2nd ed.) Washington, D. C.: American Council on Education, 1971.

Hotelling, H. Analysis of a complex statistical variables into principal components. Journal of Educational Psychology, 1933, 24, 417-441, 498-520.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Kelley, T. L. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 1939, 30, 17-24.

Kirk, R. Experimental design: Procedures for the behavioral sciences. Belmont, Ca.: Wadsworth, 1968.

Lange, A., Lehmann, I., and Meherens, W. Using item analysis to improve tests. Journal of Educational Measurement, 1967, 4, 65-68.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al., Measurement and prediction. Princeton, N. J.: Princeton University Press, 1950.

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

Lord, F. M. The relationship of the reliability of multiple-choice tests to the distribtuion of item difficulties. Psychometrika, 1952, 18, 181-194 (b).

Lord, F. M. An application of confidence intervals and of maximum likehood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-75 (a).

Lord, F. M. The relations of test scores to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-548 (b).

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

Lord, F. M. The relative efficiency of two tests as a function of ability level. Psychometrika, 1974, 39, 351-358 (a).

Lord, F. M. Quick estimates of the relative efficiency of two tests as a function of ability level. Journal of Educational Measurement, 1974, 11, 247-254 (b).

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

Lord, F. M. and Novick, M. Statistical theories of mental test scores. Reading, Ma.: Addison-Wesley, 1968.

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

McNemar, Q. The revision of the Stanford-Binet scale. Boston: Houghlin-Mifflin, 1942.

McNemar, Q. Psychological Statistics. New York: John Wiley and Sons, Inc., 1962.

Magnusson, D. Test theory. Reading, Ma.: Addison-Wesley, 1966.

Mandeville, G. K. and Smarr, A. M. Rasch model analysis of three types of cognitive data. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Marscuilo, L. A. Statistical methods for behavioral science research. New York: McGraw-Hill, 1971.

Mehrens, W. and Lehmann, I. Measurement and evaluation in education and psychology. New York: Holt, Rinehart, and Winston, Inc., 1973.

Mendenhall, W., Ott, L., and Schaffer, R. Elementary survey sampling. Belmont, Ca.: Wadsworth Publishing Co., 1971.

Mosier, C. I. Problems and designs of cross validation. Educational and Psychological Measurement, 1951, 11, 5-11.

Myers, C. T. The relationship between item difficulty and test validity and reliability. Educational and Psychological Measurement, 1962, 22, 565-57.

Pearson, K. On the correlation of characters not quantatively measurable. Royal Society Philosophical Transactions, Series A, 1900, 195, 1-47.

Pearson, K. On a new method of determining a correlation between a measured character of A, and a character of B, of which only the percentage of cases wherein B exceeds (or fall short of) intensity is recorded for each grade of A. Biometrika, 1909, 7, 96-105.

Rasch, G. Probablistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedagogiske Institute, 1960.

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

Rentz, C. Rasch model invariance as a function of the shape of the sample distribution and degree of model-data fit. A paper presented at the annual meeting of the Florida Educational Research Association, January 1976.

Rentz, R. R. and Bashaw, W. L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Richardson, M. W. Notes on the rational of item analysis. Psychometrika, 1936, 1, 69-76.

Ryan, J. P. The rationale for the Rasch model. A paper presented at the pre-convention training session for the Rasch model, the annual meeting of the American Educational Research Association, New York, April 1977.

Spearman, C. General intelligence objectively determined and measured. American Journal of Psychology, 1904, 15, 201-293.

Swineford, F. Biserial r versus Pearson r as measures of test-item validity. Journal of Educational Psychology, 1936, 27, 471-472.

Tinsley, H. and Dawis, R. An investigation of the Rasch simple logistic model: Sample-free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Thurstone, L. L. Multiple factor analysis. Chicago: University of Chicago Press, 1947.

Ware, W. B. and Benson, J. Appropriate statistics and measurement scales. Science Education, 1975, 59, 575-582.

Webster, H. Maximimizing test validity by item selection. Psychometrika, 1956, 21, 153-164.

Wherry, R. J. and Winer, B. J. A method for factoring large numbers of items. Psychometrika, 1953, 18, 161-179.

Whitely, S. and Dawis, R. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Whitely, S. and Dawis, R. The influence of the context on item difficulty. Educational and Psychological Measurement, 1976, 36, 329-337.

Woodcock, R. W. Woodcock Reading Mastery Test. Circle Pines, Minn.: American Guidance Service, 1974.

Wright, B. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Wright, B. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D. and Mead, R. J. CALFIT: Sample-free item calibration with a Rasch measurement model. Research Memorandum No. 18. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1975.

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

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

APPENDIX A

MATHEMATICAL DERIVATION OF THE RASCH MODEL[1]

[1]Summarized from Ryan (1977)

If ability is equal to θ and item difficulty is equal to ζ then the odds of correctly solving an item is given by

$$\text{odds} = \frac{\theta}{\zeta} \quad , \tag{1}$$

When $\theta > \zeta$ the person will get the item right, when $\theta < \zeta$ the person will get the item wrong, and when $\theta = \zeta$ the person will have a 50-50 chance for success on the item.

The probability of a correct response can readily be derived from the statement of the odds. In general,

$$\text{Probability} = \frac{\text{odds}}{1 + \text{odds}} \quad , \tag{2}$$

substituting (1) into (2)

$$P = \frac{\theta/\zeta}{1 + \theta/\zeta} \quad . \tag{3}$$

which is equivalent to

$$P = \frac{\theta/\zeta}{\zeta/\zeta + \theta/\zeta} = \frac{\theta/\zeta}{\zeta + \theta/\zeta} = \frac{\theta}{\zeta + \theta} \quad . \tag{4}$$

Formula 4 is the probability of a correct response.

The formula of an incorrect response is represented by Q, which is 1-P. Thus,

$$Q = (1-P) = 1 - \frac{\theta}{\zeta + \theta} \quad , \tag{5}$$

which is equivalent to

$$Q = \frac{\zeta + \theta}{\zeta + \theta} - \frac{\theta}{\zeta + \theta} = \frac{\zeta}{\zeta + \theta} \quad . \tag{6}$$

Formula 6 is the probability of an incorrect response.

The relationship of these probability statements to the statement of the odds is given by

$$\frac{P}{Q} = \frac{\theta/\zeta + \theta}{\zeta/\zeta + \theta} = \frac{\theta}{\zeta} \quad . \qquad (7)$$

The right side of equation 7 is simply the odds as defined in equation 1. The left side of the equation is the probability of a correct response divided by the probability of an incorrect response. The probability of a correct response, P, is estimated by the proportion of examinees in a sample who correctly answer an item. On a test of k items, the probability of a person with score of j (where $1 \leq j < k$), correctly answering a particular item is simply the proportion of people with the raw score j who correctly answered the item. This is nothing more than the item difficulty for all the people who have a raw score of j. The probability can be calculated in terms of the item difficulty for all raw score groups from 1 to (k-1) and across all items.

The probability of an incorrect response, Q, is the proportion of people answering an item incorrectly. This is simply one minus the proportion answering it correctly (1-P). Both P and Q are easily calculated from a set of data hence the value of P/Q in formula 7 is an easily derived statistic which estimates the odds.

Separating the Parameters

Consider again equation 7 and take the natural log (ln) of both sides of the equation. This gives

$$\ln\left(\frac{P}{Q}\right) = \ln\left(\frac{\theta}{\zeta}\right) \quad . \qquad (8)$$

Since the ln of a ratio is the same as the difference of the ln. 8 becomes

$$\ln \left(\frac{P}{Q}\right) = \ln \theta - \ln \zeta \, . \qquad (9)$$

Person Free Item Difficulties

Next consider a score group with ability θ_1 and two test items with difficulties ζ_1 and ζ_2 respectively. The probability of a person with ability θ_1 correctly answering an item of difficulty ζ_1 is P_{11}. The probability of an incorrect response is Q_{11}. The probability of a person with ability θ_1 correctly answering an item of difficulty ζ_2 is P_{12} and the probability of an incorrect response is Q_{12}. By equation 9,

$$\ln \left(\frac{P_{11}}{Q_{11}}\right) = \ln \theta_1 - \ln \zeta_1 \text{ and} \qquad (10)$$

$$\ln \left(\frac{P_{12}}{Q_{12}}\right) = \ln \theta_1 - \ln \zeta_2 . \qquad (11)$$

If equation 11 is subtracted from equation 10, term by term, the result is

$$\ln \left(\frac{P_{11}}{Q_{11}}\right) - \ln \left(\frac{P_{12}}{Q_{12}}\right) = (\ln \theta_1 - \ln \zeta_1) - (\ln \theta_1 - \ln \zeta_2) \, . \qquad (12)$$

This is the same as

$$\ln \left(\frac{P_{11}}{Q_{11}}\right) - \ln \left(\frac{P_{12}}{Q_{12}}\right) = \ln \theta_1 - \ln \zeta_1 - \ln \theta_1 + \ln \zeta_2 \, , \qquad (13)$$

$$\text{or,} \quad \ln \left(\frac{P_{11}}{Q_{11}}\right) - \ln \left(\frac{P_{12}}{Q12}\right) = \ln \zeta_2 - \ln \zeta_1 . \qquad (14)$$

Equation 14 should be examined very carefully. On the left side of the equation is an easily calculated statistic: The difference between the ln odds of a correct response on item 1 compared to item 2.

The right side is significant for what it does not contain. There is no parameter for the person's ability on the right side of the equation. This same result occurs regardless of the ability of the person or group examined. The difference between the difficulty of item 1 and the difficulty of item 2 can be calculated independently of the subject or group of subjects involved. In general, for any two items of difficulty ζ_1 and ζ_m, the difference between the difficulty of the two items is given by

$$\ln \zeta_m - \ln \zeta_1 = \ln \left(\frac{P_{i1}}{Q_{i1}}\right) - \ln \left(\frac{P_{im}}{Q_{im}}\right) . \qquad (15)$$

Item Free Person Abilities

The discussion of person ability estimates is an exact parallel to the discussion of item difficulty estimates. Instead of comparing two items across any group of examinees the discussion of ability proceeds by comparing any two groups on any test item. Consider score group $_1$ with ability θ_1, score group $_2$ with ability θ_2, and item $_1$ with difficulty ζ_1. From equation 9,

$$\ln \left(\frac{P_{11}}{Q_{11}}\right) = \ln \theta_1 - \ln \zeta_1, \text{ and} \qquad (16)$$

$$\ln \left(\frac{P_{21}}{Q_{21}}\right) = \ln \theta_2 - \ln \zeta_1 . \qquad (17)$$

Subtracting equation 17 from equation 16 will yield

$$\ln \left(\frac{P_{11}}{Q_{11}}\right) - \ln \left(\frac{P_{21}}{Q_{21}}\right) = \ln \theta_2 - \ln \theta_1. \qquad (18)$$

The difference between the abilities of the examinees in score group $_1$ and score group $_2$ (the right side of equation 18) is described without reference to the item involved. In general, for any two groups with abilities θ_i and θ_j,

$$\ln \theta_j - \ln \theta_i = \ln \left(\frac{P_{ik}}{Q_{ik}}\right) - \ln \left(\frac{P_{jk}}{Q_{jk}}\right) . \qquad (19)$$

for any item with difficulty ζ_k. In this case the abilities are being compared independently of the difficulty of the item used to compare them. This is often referred to as item free person ability estimation.

Formalizing the Model

To describe the Rasch model let $\ln\theta$ and $\ln\zeta$ be re-defined. Specifically, let,

$$\beta = \ln \theta, \text{ and} \qquad (20)$$

$$\delta = \ln \zeta. \qquad (21)$$

Equations 20 and 21 simply define the ln ability as β and the ln difficulty as δ.

If both sides of equation 20 and 21 are raised to the base of the natural log system, e, we get

$$e^{\beta} = e^{\ln \theta}, \text{ and} \qquad (22)$$

$$e^{\delta} = e^{\ln \zeta}. \qquad (23)$$

Recall equation 3

$$P = \frac{\theta/\zeta}{1 + \theta/\zeta} , \qquad (24)$$

and substitute the equivalent terms for θ and ζ as defined in equations 22 and 23. This gives

$$P = \frac{e^{\beta} / e^{\delta}}{1 + e^{\beta} / e^{\delta}} \text{ or} \qquad (25)$$

$$P = \frac{e^{(\beta-\delta)}}{1 + e^{(\beta-\delta)}} . \qquad (26)$$

More formally this is

$$P\ (X_{ki} = 1 \mid \beta_k, \delta_i) = \frac{e^{(\beta_k - \delta_i)}}{1 + e^{(\beta_k - \delta_i)}} \qquad (27)$$

Equation 27 is the Rasch model.

APPENDIX B

RELATIVE EFFICIENCY VALUES USED IN FIGURE 2 FOR THE COMPARISONS AMONG ITEM ANALYTIC METHODS

TABLE B.1

RELATIVE EFFICIENCY VALUES USED IN FIGURE 2 FOR THE COMPARISONS AMONG ITEM ANALYTIC METHODS

Percentile Rank	Factor Analysis to Classical	Rasch to Classical	Rasch to Factor Analysis
1	4.00	.49	.12
10	1.62	.55	.34
20	.72	.49	.63
30	.95	.84	.88
40	1.25	1.33	1.06
50	.80	1.00	1.25
60	1.39	1.30	.94
70	1.17	1.68	1.43
80	.91	.78	.86
90	.67	1.69	2.54
98	.98	.29	.29

BIOGRAPHICAL SKETCH

Iris Benson was born January 19, 1946, in Charleston, South Carolina. She graduated from Hialeah High School, Hialeah, Florida, in June 1964.

She began attending college part-time in 1968, and graduated in August 1971 from Santa Fe Junior College in Gainesville, Florida. Iris then attended the University of Florida, and received a Bachelor of Arts degree with honors in March 1973 with a major in Psychology.

In the fall of 1973, she began her graduate studies in the College of Education at the University of Florida. Upon completing her master's thesis entitled, The Florida Twelfth Grade Testing Program: A factor analysis of the aptitude and achievement subtests, she received the degree Master of Arts in Education in June 1975.

Iris began her doctoral program at the University of Florida in the fall of 1975. While working on her graduate degrees she has held, at various times, a graduate assistantship in the area of evaluation and test development, and teaching assistantships in graduate level courses in educational measurement and statistics. She was also an evaluation intern at the Northwest Regional Educational Laboratory in Portland, Oregon from September 1976 to March 1977.

Iris is a member of the American Educational Research Association, and the National Council on Measurement in Education. She has co-authored several articles and papers on the topics of examinee test-taking behavior, appropriate statistics for different measurement scales, and the adversary evaluation model.

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

William B. Ware, Chairman
Professor of Foundations of Education

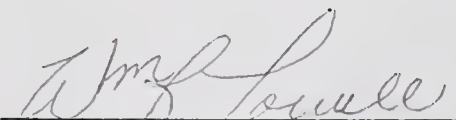
I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda M. Crocker
Associate Professor of Foundations of Education

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

John M. Newell
Professor of Foundations of Education

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



William R. Powell
Professor of Instructional
Leadership and Support

This dissertation was submitted to the Graduate Faculty of the Department of Foundations of Education in the College of Education and to the Graduate Council, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December 1977

Chairman, Foundations of
Education

Dean, Graduate School