AN INVESTIGATION OF THE EFFECTS OF TEST LENGTH ON SHORT-FORM BASIC SKILLS COMPETENCY TESTS DEVELOPED BY USING THE ONE-PARAMETER ITEM RESPONSE MODEL

By

Leroy J. Tompkins

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF EDUCATION

in

Curriculum and Instruction

Ronald L. McKeen, Co-Chairman

Barbara A. Hutson

Jim C. Fortune, Co-Chairman

Thomas E. Gatewood

Paul L. Williams, Cognate Professor

March, 1984

Blacksburg, Virginia

ACKNOWLEDGEMENTS

Rarely does anyone attain any worthwhile goal in life without help and support from the people around him. The attainment of my doctorate degree is no exception. I am extremely grateful to the many individuals who encouraged me throughout my growth and development in attaining this goal. While, the number of individuals to whom I am indebted are too numerous to name individually in this document, my gratefulness is no less sincere. Several individuals I feel are due special tribute because of the special contributions they made to my modest success.

First and foremost I am eternally grateful to my wife, Bettye, and my sons, Leroy and Derek, for their love, compassion, and understanding while I took time from their lives to complete this project. My fondest wish is that my achievement serves as a model for them and future generations that any goal can be attained through perserverance and hard work.

To my mother, Ellawese, whose patience, confidence, and love laid the foundation for all that I am and all that I shall become, I am forever thankful.

I would like to thank all of the members of the staff at Virginia Tech who have contributed to my academic and personal growth. However, there are several members to whom I am especially appreciative. It should be noted that the order in which they are mentioned is not indicative of the contributions they made to my efforts.

I would like to thank Dr. Ron McKeen for helping me design and negogiate my program of studies while attending the university. I would also like to thank him for orchestrating my committee and taking care of administrative details in my preparations for this presentation.

I am especially grateful to Dr. Paul Williams and his staff at the Maryland State Department of Education, for planting the seed of item response theory in my mind and for staying around to help me nurture and cultivate it to the point that I can independently continue my study in the field.

I would like to thank Dr. Jim Fortune for his genius in helping me develop the analytic design which was used in this investigation.

To Dr. Barbara Hutson, who's forethought and wisdom helped to mold my somtimes disparate thoughts and ideas into a coherent study, I offer a very special thank you.

I would like to thank Dr. Tom Gatewood for participating on my committee and helping me develop the applied component on the study.

Finally, I would like to thank Steve and Joy for their understanding, support, and compassion while I pursued my studies. Also thanks to my co-workers for their "ears" and encouragement while I conducted this investigation.

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Background

Efficient and effective use of available time for instructional purposes is almost always a primary concern for educational practitioners. One of the areas of the instructional process which can be more efficiently utilized is that of testing students to determine whether or not they possess the minimum skills for such requirments as for promotion and graduation. Presently, a considerable amount of the time available for instruction is consumed by testing. When state-mandated testing, local school district-mandated testing, and departmental and individual teacher testing are all taken into account, a substantial amount of time which could be used to provide other instructional activities directed toward increasing student learning is consumed.

Generally, the purpose of testing is to assess the relative effectiveness of other instructional activities which are designed to promote learning or for diagnosis of areas in which additional instruction is needed. It is also generally recognized that testing is an integral part of the teaching-learning process and that there is a definite need for much of the information obtained from the administration of these tests in making program decisions regarding students. However, it is very likely that an increase in the amount of time teachers have to engage in those activities which are specifically directed toward student acquisition of information is the place

where the greatest potential payoff lies with respect to pupil learning.

Consequently, the questions arise as to how can the time presently consumed by testing be minimized without any substantial loss in the information needed by decision makers and how can more time be allowed for teachers to teach? In other words, how can the time which is presently available for instructional activities, which include testing, be optimized? One possible means of using instructional time more efficiently is through short-form tests. This is especially true if these tests are capable of providing sufficient information to decision makers and require less time for administration. In addition, short-form tests could substantially reduce the testing burden currently placed on students. This particular approach can be accomplished by revising currently used tests with the use of classical test development procedures and/or with more contemporary test development procedures such as those made possible with Item Response Theory (IRT).

Classical testing and measurement procedures have proved to be inadequate in resolving several problems in some very crucial areas of test design and analyses which are anticipated in reducing the length of some of the tests presently used by school districts to determine student competency levels (Hambleton, et al., 1978). Those problems include variation of classically obtained item difficulty and discrimination indices across examinees of different abilities, incomparability of examinees who have been administered different subsets of test items, and the absence of a theoretical basis for

predicting the interaction between an examinee and a test item (Hambleton, 1978). Thurstone (1928) recognized these needs in measurement during the first quarter of the century:

> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is affected, the validity of the instrument is impaired or limited. If a yardstick measures differently because of the fact that it is a rug, a picture, or a piece of paper that is being measured, then to that extent the trustworthiness of that yardstick as a measurement device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement. (p. 547)

Further, it is common knowledge that the educational community has been sorely in need of psychometric methods and statistical procedures which are oriented toward measurement of individual pupil abilities and competencies. Zubin (1955) supports this notion as is apparent by these remarks:

> Recourse must be had to individual statistics, treating each patient [student] as a separate universe ... present day statistical methods are entirely group-centered so that there is a real need for developing individual-centered statistics. (p. 2-28)

B. F. Skinner (1956) reiterated the position taken by Zubin regarding classical testing procedures:

> The order to be found in human and animal behavior should be extracted from investigations into the individual, and

psychometric methods [classical] are inadequate for such purposes since they deal with groups of individuals. (p. 221)

During the last four decades, many testing specialists and psychometricians have been engaged in investigations in the field of latent trait theory, also referred to as item response theory or item characteristic curve theory. Many of these investigations have been developed around the one-, two-, and three-parameter models of item response theory. These models differ primarily in the number of item attributes which are used in determining the ability of an examinee. The one-parameter model uses only the difficulty of the item, the two-parameter model uses the difficulty of the item and the descrimination of the item, and the three-parameter model uses item difficulty, item descrimination, and a third factor for guessing.

Results have been obtained from these investigations which address many of these issues associated with measurement (Hambleton, et al., 1978). According to Hambleton much of the work in this area was begun by Lawley (1943, 1944) and Lazarsfeld (1950). Since that time major contributions to the field have been made by Frederick Lord, Benjamin Wright, Allen Birnbaum, Ronald Hambleton, and several other testing experts. The successful application of this theory has produced some very promising results in resolution of some of the problems in test design and evaluation, particularly in the development of parallel form criterion-referenced tests, equating tests (Hambleton and Novick, 1973; Rentz and Bashaw, 1975), and construction of situation specific tests (Wright and Douglas, 1975).

Statement of the Problem

Based on the findings from previous research in the field of item response theory, a definite potential exists for minimizing the time presently consumed in testing pupils for the purpose of ascertaining their level of competence or mastery with regard to some specific skill or trait. Further the potential exists for accomplishing this task without any substantial loss of information (e.g., the location of an examinee on the variable defined by the items comprising the test). Propositions for reducing the time presently consumed by testing are inherent in the one-parameter item response model and were examined empirically in this study. If these propositions are shown to be valid, the results could lead to a substantial decrease in both the time presently consumed by testing and the testing burden on students. Specific application of using the one-parameter model to this end will be investigated by using the Maryland Functional Reading Test (MFRT), which is one of the tests used in the Maryland State Department's annual plan to assess the competency of students in functional reading.

Guided by this potential the primary purpose of this study is the development of a model for constructing short-form criterion-referenced tests by using test items which have been calibrated with the one-parameter model. The major factor impacting on this process which will be investigated in this study is test length (e.g., the number of items on the test). The present study was designed to examine the following problem:

What are the effects of test length on the estimation of functional reading ability levels and the pass/fail classifica-

classifications of ninth grade pupils in the state of Maryland when test items are used which have been calibrated with the one-parameter item response model?

The test score attributes most important in this investigation were invariance of pupil ability estimates between short-form tests which varied in length and the original form of the MFRT and the classification consistency between the original form of the MFRT and different short forms of the test in terms of assigning pupils to mastery/non-mastery states. The specific research questions examined in this investigation were as follows:

1. What are the effects of test length on estimating the functional reading ability of ninth grade examinees in the state of Maryland?
2. What are the effects of test length on classification of students to mastery/nonmastery states of functional reading when the same cut-score is used on test forms of different lengths?

Hypotheses

Based on the preceding discussions, problem statement, and research questions, the following hypotheses were examined in this study:

1. The correlation between each of the short forms of the test and the original form of the test will be greater than or equal to the reliability of the original form of the test.

2. The classification decisions made by using the short forms of the MFRT developed in this investigation will be the same, within the standard error of measurement, as the classification decisions made by using the original form of the MFRT.

The data analyzed in this study were obtained from the actual administration of the Maryland Functional Reading Test to all eligible ninth grade students in the state of Maryland during the fall of 1982. The situations relevant to this study (e.g., pupil responses to items on the short forms of the test) were simulated by using subsets of the item response data from the actual administration of the test.

Definition of Terms

Latent Trait

Latent traits are unobservable explanatory abstractions which manifest themselves through observable behaviors or surface traits. Latent trait theories are theories which are based on the assumption that performance on a test (an observable behavior) is explained by the amount of the unobservable trait possessed by the examinee. According to Lord and Novick (1968), any theory of latent traits supposes that an individual's behavior can be accounted for by defining the traits, quantitatively estimating the individual's standing on the traits by using test scores, and then using the numerical results from the test to explain the performance of the examinee.

The set of all of the psychological dimensions which are necessary to explain the performance of the examinee or population of examinees on a given set of items is called the latent space. The dimensionality

of this latent space refers to the number of traits which are required to explain the performance of an individual on a given task. The assumption of unidimensionality of the test data which fit the one-parameter model is the assumption that a single psychological dimension or latent ability accounts for or is sufficient to explain the observed performance of an examinee on a set of test items. It should be noted that rarely is this assumption totally true. However, Lord (1968) has indicated that the extent to which it is true can be closely approximated by using factor analysis techniques.

Rasch's Model of Latent Trait Theory is one of several probabilistic models which purport to describe this relationship between the observed test score of an examinee and the underlying trait prompting that score.

Local Independence

Local independence means that within any sample of "k" examinees each characterized by the same latent variable 0_1, 0_2, 0_3,...0_k, the conditional distribution of the item scores are independent of each other, except through the latent trait they possess (Lord and Novick, 1968).

Hambleton and Cook (1977) describe two notions of the local independence concept in terms of a strong form of the assumption and a weak form of the assumption. According to these researchers, the strong form of the assumption is met when the items comprising the test are statistically independent. That is to say that the performance of an examinee on any single item of the test is not affected by his/her

performance on any other item on the test. The weak form of the assumption is met if pairs of items of the test are simply uncorrelated for examinees of the same ability level.

Error

In the context of latent trait theory, error refers to the observed responses to a test item which are not a function of the trait assessed by the item nor the difficulty of the test item. These are responses generated by some random event such as guessing.

Item Characteristic Curve

An item characteristic curve is a mathematical function which relates the probability of a correct response to a test item to the trait which is measured by the item (Hambleton and Cook, 1977, p. 78). In other words, it is the nonlinear regression function of an item score on the trait measured by the item. This curve is one of a family of curves describing different latent trait or IRT models. Hambleton and Cook (1977) provide a brief description and illustrations of several of these curves on page 79 of their publication. A more detailed description of item characteristic curves is provided by Lord and Novick (1968).

According to Hambleton and Cook (1977), item characteristic curves for the one-parameter model are non-intersecting curves that differ only by a translation along the horizontal axis representing the difficulty scale. Theoretically, each curve has the same slope indicative of equal discrimination indices. The curves representative of the two-parameter model differ both in slope and position relative to the ability axis. This indicates that the items differ in

difficulty and discriminating capability as well. The three-parameter model is symbolized by item characteristic curves which embody the characteristics described for the two-parameter model above, in addition to the different lower asymptotes. The probability of a correct response varies from zero to one when the one- and two-parameter models. However, because the three-parameter model accounts for guessing, the lower asymptote will be greater than zero.

Test Characteristic Curve

A test characteristic curve is a mathematical function which shows the relationship between the observed (raw) score of an examinee and his/her ability score. Latent trait models allow the transformation of examinee performance on a given test (subset of items) onto an ability scale defined for a larger pool of items.

Logit

A logit is a unit of measurement used in measuring persons and calibrating items in Rasch data analyses. According to Wright (1977, p. 99), a person's ability in logits is equal to the natural log odds of his/her succeeding on an item which has a difficulty of zero. That is, the probability P of an examinee correctly answering an item with a difficulty of zero is $e^B/(1 + e^B)$ from which his/her odds of success are $P/(1 - P) = e^B$. The natural log of e^B is B, the individual's estimated ability.

Conversely the difficulty of an item is the natural log odds of failure on the item by a person whose ability is zero. The probability P of an examinee with an ability of zero correctly answering an item

with a difficulty of δ is $e^{-\delta}/(1 + e^{-\delta})$. The odds of incorrectly answering the item are $(1 - P)/P = e^{\delta}$ whose natural log is δ (Wright, 1977).

Specific Objectivity

Specific objectivity refers to the person-free test calibrations and test-free person estimates which result from use of the Rasch Model of IRT. The four criteria suggested by Rasch (1960) as necessary for specific objectivity are as follows:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should be independent of which other stimuli within the class were or might also have been compared. Symmetrically, a comparison between individuals should also be independent of which stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion. (p. vii)

That is, the item parameter estimates are independent of the mean ability and variance of the sample used to calibrate the items. Further, examinee ability estimates are independent of the subset of calibrated items used to measure the examinee.

It is this property which is alluded to in discussions of invariance of item and person parameter estimates when IRT models are used to analyze test data.

Classification Errors

Classification errors are errors resulting from incorrect classification of examinees to a state of mastery or nonmastery. False-positive classification errors are those which occur when examinees are classified as masters of specified subject content when in fact they are nonmasters. False-negative classifications are the result of classifying examinees as nonmasters when in reality they are true masters.

Summary

In summary, the successful completion of this study has the potential benefit of reducing pupil test burden by providing a plan for developing short-form competency tests which will substantially reduce the time consumed by these activities without any loss in information for teachers or other decision makers regarding pupil achievement levels. Further, this study will provide a needed contribution to the literature regarding the applicability of the item response models to resolving practical educational problems in measurement by using empirical data. An additional benefit of the study includes the potential for reducing the cost of testing to state and local school districts by making the testing process more efficient.

Organization of the Research Report

Chapter I presented the introduction and the statement of the problem which guides this investigation. Also presented in chapter I are the research questions and hypotheses addressed by the study. The chapter concludes with the definition of technical terms which are

inherent in discussions regarding item response theory and which are used throughout the report.

Chapter II describes the literature reviewed in pursuit of this study. This section of the report is divided into four major areas. The first section presents an explanation of item response models and the context of the literature in which they can be found. Also included in this first section is a description of the basic assumptions underlying the theory of latent trait models and the purported advantages of these models over traditional methods of test development and analysis. The second section of this chapter summarizes some of the studies which are similar to the one presented in this report in which the one-parameter model of item response theory has been utilized in the development of short-form tests, item selection strategies, and test equating. The third section included in this chapter discusses some of the research on the assessment of the reliability of criterion-referenced tests. The final section of this chapter describes the results of other studies which investigated the effects of test length and item selection strategies on estimates of student ability.

The third chapter presents the methodology used in conducting this investigation, with a detailed description of the procedures utilized. The sections presented in this chapter include the following: (a) Samples, (b) Instrumentation, (c) Calibration of Item Parameters, (d) Development of Short-Form Tests, (e) Estimation of Pupil Functional Reading Ability Levels, (f) Assessment of the Relationship Between Short- and Long-Form Ability Estimates, and (g) Assessment of the Classification Consistency of Short-Form Tests.

The results obtained from implementation of the procedures described in the preceding chapter (Chapter III) are presented in chapter 4 of this report. The results are presented in five major sections: (a) Estimation of Pupil Ability, (b) Test Reliability, (c) Correlational Analysis, (d) Classification Consistency, and (e) Analysis of Pass/Fail Misclassifications. The chapter is concluded with a summary of the overall findings.

Chapter V presents the implications and significance of the findings resulting from this investigation in terms of the feasibility of using short-form tests to estimate pupil ability levels and make decisions regarding student pass/fail classifications. Also presented in this chapter is a proposal for empirically validating the results obtained in this study and testing the new hypotheses developed from this study in future research. The chapter concludes with the presentation of a preliminary proposal for utilizing the reduced test forms to maximize the time available for other instructional activities.

CHAPTER 2

REVIEW OF THE LITERATURE

Scope

A comprehensive review of the literature related to item response theory in general and the Rasch Model in particular has been undertaken in the process of preparing for and conducting this investigation. One of the major factors guiding this review was the need to place latent trait or item response models into some perspective with regard to more traditional techniques of item and test analysis. The review was also conducted to appraise the worth of the one-parameter model in resolving the problem posed by this investigation.

The literature reviewed in this study has been organized into four major sections and is presented in that format. The first section presents an explanation of item response models and the major advantages derived from using these models over classical test analysis models. The second section describes some of the applications of IRT models in resolution of practical educational measurement problems. The third section presents a summary of studies which have investigated methods for assessing the reliabilty and validity of criterion- referenced tests. The fourth section provides a review of other studies which have been conducted regarding the effects of test length and item selection procedures on measurement of student achievement.

Explanations of the Model

The literature reviewed for presentation in this section reflects considerable variation in terms of scope and level of technical sophistication. Most notable has been the range in the technical level and detail of expositions explaining how the models work, thus making the reading palatable for readers of varying levels of interest and expertise in test theory. The theoretical or mathematical derivation of the one-parameter logistical model which is frequently referred to as the Rasch Model of Latent Trait Theory can be found in Rasch's 1960 or 1981 expanded edition of Probabilistic Models for Some Intelligence and Attainment Tests. The purpose for developing the model, according to Rasch, was to discount the role played by the examinees in the analysis of test results. Rasch (1960) prefaces his text by the following remark:

> Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments - tests or items or other stimuli - within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class - "measuring the same thing" - independent of which particular individuals within a class considered were instrumental for comparison. (p. VII)

Obvious similarities can be observed between these remarks and those made by Thurston at the beginning of the century (see p. 3 of this presentation).

Other explanations of the model developed by Rasch are contained in works by Lord (1968 & 1974); Wright (1977); Hambleton (1977); Phillips (1981); Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978); Ryan (1977 & 1982); Rentz (1969); and many others. These expositions vary in level of sophistication from very complex mathematical derivations of the model by the likes of Lord and Birnbaum (1968) to a more easily understandable, less mathematically cumbersome descriptive explanations of the model by Hambleton and Cook (1977), Ryan (1977 & 1982), Wright (1968 & 1977), and Wright and Panchapakesan (1969).

Two of the basic assumptions underlying item response theory are unidimensionality of the latent space and local independence of the items on the measurement instrument (Hambleton & Cook, 1976). Briefly defined, unidimensionality of the latent space refers to the assumption that examinee performance on the test is accounted for by a single latent ability (Hambleton, et al., 1978). Local independence of test items is the assumption that the probability of an individual answering an item correctly on a test is not affected by his/her performance on any other item on that test (Hambleton 1978). Note: The assumption of local independence does not mean that the item scores are unrelated to each other for the total group of examinees but that the items scores are related only through the latent ability being measured (Lord & Novick, 1974, p.361).

The specific model used in this study was the one-parameter logistic model. This model represents one of several item response models which propose to explain the observed responses of an individual to a test item in terms of very specific underlying psychometric theory regarding person-item interactions (Hambleton & Cook, 1977). A basic assumption of all item response models is that the performance of an examinee can be explained by unobservable traits or abilities inherent in the examinee. According to Lord (1968), these theories suppose that:

> An individual's behavior can be accounted for, to a substantial degree, by defining certain human characteristics called traits, quantitatively estimating the individuals standing on each of these traits, and then using the numerical values obtained to predict or explain performance in relevant situations. (p. 358)

Models representing this theory are mathematical descriptions of the relationship between the amount of the trait possessed by the examinee and his/her observed score on the test (Hambleton, et al., 1978).

Advantages of Item Response Models

When the test data conform to the basic assumptions of the model, there are several very important advantages which justify the use of these models over classical psychometric procedures in the development and analysis of tests and test results (Hambleton, et al., 1978; Hambleton & Cook, 1977, 1976; Engelhard, 1980; Slinde & Linn, 1979; Doughlas, Khavari, and Farber, 1979; and Forster, 1977). One of the most important of these advantages and the one most critical to this

investigation is that a person's ability with regard to some specified trait can be estimated by using any subset of items which have been calibrated to that trait. Assuming the truth of the model, regardless of how many items comprise the test or what the statistical characteristics of the items are, an asymptotically unbiased estimate of an examinee's true score with regard to the trait measured can be obtained by using those items which have been calibrated to that trait (Hambleton & Cook, 1977). Parenthetically, it was noted by Hambleton that the above proposition is valid provided that the number of items is not too small. His inclusion of this parenthetical note is suggestive of the notion that there is some lower limit with regard to the number of items the test can contain and still produce the purported estimates. Hence an important question raised is how few items might the test contain and still provide a reliable estimate of the examinee's ability. The present investigation should provide some additional insights into the nature of the relationship between test length and the estimation of examinee ability.

A second advantage which is especially pertinent to school districts in making the individual decisions which are made with the current use of competency testing is the increased precision of measuring pupil ability at all ability levels. This advantage results from the fact that item response models provide a measure of the precision of the estimated ability at each level of the ability range. When a decision is made regarding the competency classification of a given examinee, the models allow estimates of measurement error to be

computed specifically for that individual or group of individuals. According to Hambleton and Cook (1977):

> Instead of providing a single standard measurement error that applies to all examinees, regardless of their test scores, latent trait models make it possible to provide separate estimates of error for each examinee or for each ability level. (p. 91)

A third advantage resulting from the use of item response models over classical models of test analyses is the possibility of comparing the abilities of examinees when they have been administered different subsets of items. This property permits the use of tailored testing for assessment purposes which is virtually impossible by using classical test procedures. By using test items selected from an item bank consisting of items calibrated with item response models, individuals can be compared even when they have been administered tests containing completely different items. This advantage is possible because all of the items in the item bank are on the same ability scale.

The invariance of item difficulty estimates across examinees of different ability levels is a fourth very important advantage of using item response models. It is common knowledge that classically obtained item statistics of item difficulty and item discrimination will vary based on the ability of the sample taking the test. This particular advantage brought about with IRT models resolves the issue raised by Thurston (1925) and discussed in chapter I of this presentation. That is, regardless of the ability of the group tested, the characteristics of the test are not affected. Research by Rentz (1975), Willmot and

Fowles (1973), and Anderson et al. (1968) supports the validity of this assumption.

The advantages discussed above have definite implications for resolving the problem of reducing the time consumed by testing and the testing burden on students, while providing the information needed by school system decision makers. Based on the propositions of IRT models, once a set of items has been calibrated to a well-defined domain of behaviors, the development of short-form tests appears to be almost routine. It should be noted, however, that with regard to selecting items to shorten a given test, other theoretical propositions must also be considered. Further, the question of which set of procedures or combination of procedures will produce the most accurate and reliable results remains unanswered.

The one-parameter model was chosen for this investigation over the two- and three-parameter models because of the interpretability of the results produced by a test which fits the model. The one-parameter model follows from the assumption that the unweighted number of correct responses on a test is sufficient to measure the ability of an examinee and that the unweighted number of correct responses to an individual item is sufficient for measuring the difficulty of the item. Other models utilizing two or more parameters require more complicated measurement schemes which are dependent upon parameters which cannot satisfactorily be determined (Wright, 1977, p. 97).

Derivation of the One-Parameter Model

The mathematical derivation of the one-parameter model is based on two very elementary propositions, the truth of which is tautological (Ryan 1977). The first of these is that the more of an ability or trait being measured the individual has, the more likely he/she is to answer a test item measuring that trait correctly. The second of these propositions is that every examinee, regardless of ability, is more likely to answer an easy item correctly than a more difficult item. The mathematical representation of these propositions in terms of odds is as follows:

Odds of a correct Response = a/b

where "a" represents the ability of the person measured and "b" represents the difficulty of the item encountered. As "a" increases relative to "b" the odds of the individual correctly answering the item increase. In terms of probabilities the equation is as follows:

Probability (of a correct response) = a/(a + b)

Following several mathematical manipulative steps utilizing the laws of logarithms, the formal statement of the Rasch Model can be derived. See Phillips (1981) for a detailed presentation of the mathematical derivations. The generalized statement of the one-parameter item response model is as follows:

$$Pr(X_{vi} | \beta_v, \delta_i) = e^{X_{vi}(\beta_v - \delta_i)} / [1 + e^{(\beta_v - \delta_i)}]$$

Where,

β = the natural log of person ability or ln a

δ = the natural log of the difficulty of the item or ln d

X_{vi} = 1 represents a correct response by person v on item i. (Wright, 1977, p. 99)

This equation states that the probability of a correct response by subject "v" with ability "β" to item "i" with difficulty "δ" is a logarithmic function of the difference between the person's ability and the item's difficulty. The graph of this mathematical function is referred to as an item characteristic curve which depicts the relationship between the probability of a correct response to a test item and the ability measured by the item set containing it (Hambleton 1978).

Much of the literature which describes the Rasch Model also provides general explanations of other item characteristic curve models (i.e., the two-parameter model, three-parameter model, etc.) which purport to perform the same function of estimating person ability. However, proponents of these models make the claim that the estimates provided by the two- and three-parameter models are more accurate because of adjustments made in examinee responses for other factors potentially affecting examinee interactions with test items such as guessing and item descrimination capabilities (Lord, 1968; Wright, 1977; Hambleton & Cook, 1976 & 1977).

It should also be noted, however, that considerable disagreement exists among psychometricians regarding the need for and applicability of the additional parameters for discrimination and guessing. The parameter for item discrimination in Rasch analyses is the variation in

the slope of the item characteristic curve to allow items to discriminate among persons (Wright 1977, p. 104). Lord (1968 & 1975) and Wright (1977) both agree that it is impossible to estimate the item discrimination parameters efficiently or accurately. Birmbaum (1968) concluded that when the number of test items is very large the differences between the ability estimates obtained by using the one-parameter model and the two-parameter model (which includes a parameter for discrimination) were negligible. Panchapakesan (1969) found that the benefits of using the Rasch one-parameter model greatly outweighed any losses in measurement accuracy when differences between discrimination indices were small.

Ross (1966) and Ryan (1982) have shown that guessing may only be a marginal factor in basic skills test analysis. Further, guessing is a person attribute and not an item attribute because people guess, not test items. With the one-parameter model, where guessing does occur, it is easily detectable by using procedures to identify systematic differences between observed responses and expected responses (Ryan, 1982, p.9). He does not, however, describe what should be done with the results once guessing has been detected.

Application of the Rasch Model

Psychometricians and educational measurement specialists have been aware of latent trait models for a number of years; however, acceptance and application of the models has, until very recently, been relatively slow. Several factors may be responsible for the apparent slow growth in applications of the models to psychometric problems in educational measurement. Two of those factors are the tremendous amount of

mathematical manipulations required to operate the model in a large scale testing program and the mathematical knowledge required to understand the model. The development of commercially available computer programs such as BICAL, LISEREL, LOGIST, and others over the last decade has made use of these models more readily available to a wider population of researchers interested in applying them to practical problems in educational measurement. As such, considerable research has now been completed which provides some very promising results, as well as some potential problem areas to be resolved in applying these models to real testing situations.

Arguments reviewed in favor of latent trait models and the potential benefits of these models suggest a tremendous potential for making the assessment component of the teaching and learning process more efficient and manageable. This section will review some of the research which has been done regarding fundamental propositions of Item Response Models and their application in resolving practical educational measurement problems.

Some of the areas in which item response theory have proved to be most viable include test development, test equating, tailored testing, and item banking. A brief summary of some of the research in these areas is presented in the sections that follow.

Test Development

In the development of measurement instruments, it is frequently important for test developers to be able to predict how a sample of examinees might be expected to interact with the items comprising a test prior to the administration of the test (Lord, 1977). When

classical test development procedures are used, this convenience is only available when the specific items in question have been administered to a sample of examinees having the same characteristics as the sample the psychometrician wishes to test. Lord (1977) describes the situation as follows:

> In practical test development work, we often need to predict the statistical properties of a test composed of items for a target group of examinees that is somewhat different from the groups to which the seperate items have previously been administered. We need to be able to describe the items by using item parameters, and the examinees by using examinee parameters, in such a way that we can predict probabilistically the response of an examinee to any item, even if similar examinees have never taken similar items before. (p. 117)

Lord further notes that the assumptions required to make the necessary predictions regarding the relationship between the ability of the examinee and his/her performance on the items that measure that ability are inherent in item response theory. According to Lord the property of item response theory which makes the above prediction possible is that given a set of items pretested using any sample of examinees, the descriptive statistics and the raw score distribution of scores for any subset of these items can be predicted when the general level of the sample of examinees is known.

Wright (1977) contends that the ability level of the potential examinee can be estimated by using almost any information available and the items selected from the test which are targeted to that ability.

When this information is used with item response models, a "best test" can be designed by using the following steps:

1. Guess target location (mean ability, M) and dispersion (standard deviation of ability) as well as possible. If outer boundaries are used to specify the target's location and dispersion, relate them to M and S by letting the lower boundary define M - 2S and the upper boundary define M + 2S.
2. Design a test with item difficulties centerd at M and spread evenly over the range M - 2S to M + 2S, with enough items in between to produce a test of length of $L = 6/SEM^2$, where SEM is the desired standard error of measurement in logits.
3. Select from the item bank the best available items to fulfill this design, and use the range and mean of the obtained item difficulties to describe the "height" h and "width" w of the resulting test. (pp. 32-35)

When the method described by Wright and Douglas is used, it is obvious that the length of the test is determined by the degree of precision demanded by the test developer.

Item Banking

A test item bank consists of a set of test items which are representative of a prespecified domain of behaviors from which a subset of the items can be drawn to measure the amount of the trait possessed by an examinee. An important attribute of item banking made possible by item response models is that the item characteristics are known prior to their administration to the sample to be tested and can therefore be used to develop a test with almost any characteristics

desired (Wright, 1977). Further, because the items share a common calibration, the scores obtained from any subset of items from the item bank are automatically equated (Wright, 1977; Rentz and Bashaw, 1975, 1977).

Test Equating

Test equating was reviewed in considerable detail for this study because these procedures are very important when comparisons are made between tests consisting of items which do not share a common calibration. The use of IRT procedures in test equating has been described by a number of authors: Lord (1977), Wright (1977), Hambleton (1976), Ryan (1981), and several others. Application of IRT models in equating tests in real testing situations has also been studied by several researchers: Engelhard (1980), Ryan and Sanders (1982), Ryan (1981), Slinde and Linn (1979), Loyd and Hoover (1980), Gustafsson (1979), Forster (1977), and Cook and Eignor (1981).

Test equating is generally considered as the process of placing test scores from two or more different tests on the same scale for the purpose of comparing the results obtained from the different test instruments. Lord (1977) noted that the tests which are to be equated should measure the same trait or ability and succinctly describes the process as follows:

> Scores y on test Y are equated to scores x on test X by a transformation y*=y*y which transforms 'raw' y scores on to the x scale. (p. 128)

Angoff (1971) described four criteria for test equating. In summary those criteria include the following:

1. The two tests must measure the same ability.
2. The equating conversions and the groups of examinees used to obtain them must be independent.
3. The results obtained from the equating procedures mustbe interchangeable for each of the tests used.
4. The equated scores must be equal regardless of which test, X or Y, is used as the base.

This process is generally one of two types: horizontal test equating and vertical test equating. Horizontal equating of two or more tests is usually performed when the two instruments have approximately the same difficulty level and are intended for use with similar samples of examinees. Situations in which this type of equating is used are described by Cook and Eignor (1981). In large-scale testing programs several forms of a test which is designed to measure some prescribed attribute (aptitude, math achievement, etc.) are usually developed for security purposes. Given recent disclosure laws regarding test results, the need for alternate forms of the same test is obvious. However, where the results from these tests are used for selection into special programs, graduation certification decisions, or even pass-fail decisions, it is critical that scores obtained on these instruments be comparable.

Vertical equating of two or more tests is used when the difficulty levels of the test differ significantly from each other, usually by design, and are intended for use with populations which differ one from

the other in the amount of the ability or attribute measured by the test (Ryan & Saunders 1982). A situation in which this type of equating would be necessary would be when educators would care to compare performance over several grade levels (e.g., grade 4 to grade 8) by using commercially prepared tests which were developed for and administered at each of those grade levels.

The literature identifies several procedures for equating test results by using both classical procedures or IRT procedures. However, given the restrictions denoted by Lord and Angoff, classical test equating procedures (e.g., equipercentile methods) may have serious limitations in test equating, particularly when applied to tests differing in difficulty level and/or using examinee populations differing in their levels of ability (e.g., vertical equating). Research conclusions reached by Slinde and Linn (1977) suggest corroboration with this opinion noting inconsistencies in both the magnitude and directions of change when classical procedures of equating were used.

The property of "specific objectivity" inherent in the one-parameter model makes the need for test equating unnecessary when the log ability estimates of examinee ability are used. Where certain assumptions have been met, the use of IRT methods of equating has proven to be an extremely viable means for equating tests both vertically and horizontally. Aside from the theoretical advantages of using IRT equating procedures, Cook and Eignor (1981) outlined several extremely important practical advantages for using these models:

1. Improved equating, including better equating at the end of the scale where important decisions are often made.
2. Greater test security through less dependence on items in common with a single form.
3. Easier re-equating should items be revised or deleted.
4. The possibility of pre-equating, or deriving the relationship between the test forms before they are administered operationally. (p. 16)

Based on the literature reviewed, there are three IRT procedures which are generally used to equate tests. According to Ryan (1981 and 1982), the three most commonly used approaches are the common-item linking approach, the common subject (examinee)/separate calibration approach, and the common calibration approach. An excellent discussion of IRT equating has been presented by Cook & Eignor (1981). This particular discussion also includes some of the practical considerations which should be used in carrying out those procedures.

The common-item linking method of equating tests has been used successfully by Cook and Douglas (1982) with the three-parameter model in equating forms of the scholastic Aptitudes Test with forms of the PSAT/NMSQT. Ryan (1982) replicated the study by Cook and Doughlas by using both the one-parameter Rasch model and the three-parameter model, with similar results being obtained for each of the two models. His conclusion from the study states:

> The two methods produced sets of scaled scores that are very nearly identical. This similarity is especially striking in light of the differences, both theoretical and practical between the two methods. (p. 10)

He goes on to say, however, that several factors recommend the use of the Rasch model. Among those were the simplicity of the model leading to greater efficiency and less cost for calibrations and the invariance of item characteristics across samples of examinees.

A prerequisite for the common-item linking method is that the tests which are to be equated have a set of common items with which a linking constant can be computed. When the two instruments meet this prerequisite, in addition to other conditions discussed by Lord (1977) and Angoff (1971), the tests can be calibrated to place the item parameters on an interval scale with an arbitrary origin (Saunders & Ryan 1982). The distance between the origins of item parameters of the common item sets in each of the two test is the additive constant which represents the linking constant. When the linking constant is added to the item difficulties of the second test, they are transformed onto the same scale as that of the first test with the same origin (Saunders & Ryan 1982, p. 5). By utilizing those item difficulty estimates of the person ability measured by the test is obtainable directly from the Rasch formula.

The second method of common subject/separate calibration is operationally the same as the common-item method. The only difference in the linking constant is derived from the ability estimates of the common subjects. This procedure would be used in situations where the

two tests which are to be equated have no items in common but do have a subset of the examinees who have measures obtained from both tests. In order to use these procedures, examinee responses from both tests are mixed together to form a single large test. The items are then calibrated as if they were from the same test. Once the item parameters have been estimated from the joint calibration, the two tests are separated and test characteristic curves developed for each test. Since all of the items are on the same scale, the abilty estimates obtained for each examinee should be the same (Ryan, 1981).

Cook and Eignor (1981) discuss in considerable detail other test equating designs which are applicable to the situations discussed above and other dissimilar situations as well. These methods are not discussed further here because they are only marginally related to the situations designed in this study.

Test Reliability

In competency assessment programs such as the one used by the state of Maryland, criterion-referenced tests are used primarily to discriminate among examinees only in terms of those who exceed the minimum competency level and those who do not. There is little concern regarding degrees of competency or incompetency. Given this purpose for testing, reliability is most important at the point in the score distribution where decisions for certification are made. The remainder of this unit describes research regarding the assessment of criterion-referenced test reliability and the effects of test length on the indices used to judge reliability.

A review of the literature reveals the existence of several statistical indices for measuring the reliability of criterion-referenced tests. Detailed summaries and critiques of some of these can be observed in studies by Eignor and Hambleton (1979), Swaminathan et a.l (1974), Fitzpatrick (1981), Berk (1980), Phillips (1983), Wilcox (1980), and Subkoviak (1976).

Berk (1980) has conducted an exhaustive review of the literature and compiled a comparison of no less than 12 approaches to assessing the reliability of criterion-referenced tests. Among those reviewed by Berk are two of the indices which were used in the present study (viz., P_o and K). The former of these indices, P_o , reflects the consistency of the decision-making process across repeated administrations of the measuring instrument (Swaminathan, Hambleton, and Algina, 1974, p. 263). The formula for computing this index is as follows:

$$P_o = P_{ii},$$

where P_{ii} is the proportion of examinees observed in the ith mastery state on each of the instruments used to classify examinees. (Subkoviak, 1980, p. 152)

Factors affecting this consistency include the mastery/nonmastery composition of the sample tested and the measurement precision of the test instrument itself (Subkoviak, 1980, p. 152). Berk indicates that this index is also influenced by the length of the measurement instrument, the variability of scores on the instrument, and cut-score used for making the mastery/nonmastery classifications.

It should be noted that the index computed for P_o does not account

for the proportion of examinees who would be classified to the same mastery state on the second administration of the test due to chance alone. For this reason coefficient Kappa (K) , developed by Cohen (1960), is preferred by many researchers in assessing the classification consistency of two test measures and is also used in this investigation. Swaminathan et al. (1974) define Kappa as follows:

$$K = (P_o - P_c)/(1 - P_c),$$

where P_o is the observed proportion of examinees consistently classified by the two measurements and P_c is the proportion of examinees expected to be consistently classified due to chance alone.

The proportion of examinees expected to be classified due chance is obtained by using the following formula:

$$P_c = \sum_{i=0}^{1} P_{i.}(P_{.i}),$$

where $P_{i.}$ and $P_{.i}$ represents the proportion of examinees assigned to the mastery state of i on the first and second administration of the test, respectively.

It should be noted that while K as defined above represents the proportion of agreement between the two administrations of the test with an adjustment for the classification attributable to chance, it is this component which is questionable according to Berk (1980). He proposes that the correction in K for chance agreement is restricted by the marginals of the contingency table used to compute K. To support his contention, he quotes an argument proposed by Livingston and Wingersky which has been reviewed and quoted below:

> Applying such a correction to the pass/fail contingency table is equivalent to assuming that the proportion of examinees passing the test could not be anything but what it happened to be. For example, if 87% of the examinees passed the test, Kappa will "correct for chance" under the assumption that "chance" would result in exactly 87% of the examinees passing the test. This assumption makes sense when a pass/fail cut-off is chosen on the basis of scores to which it will be applied, so as to pass a specified proportion of examinees. It does not make sense when the pass/fail cut-off represents an absolute standard that is applied to each examinee. (p. 250)

It would appear from these remarks that Kappa would only be appropriate for assessing classification consistency when the cut-score for making the mastery/nonmastery decisions is not constant. Berk suggests the the use of Kappa is appropriate only when the cut-score for classification decisions is determined, at least in part, by the consequences of passing or failing a predetermined proportion of examinees.

The other disadvantages noted by Berk include the sensitivity of Kappa to test length and test score variability. The first of these concerns regarding test length is related to one of the questions addressed by the present study. Berk's concerns in this area appear to be in reference to tests constructed for classroom use, where teachers frequently use fewer items to assess pupil learning gains. While issues related to subtest scores may be important in the assessments conducted by the Maryland Functional Testing Program to help individual

schools or school districts plan for remediation, these concerns are not of paramount importance in the present investigation.

The latter of the two disadvantages described by Berk, sensitivity of K to test score variability, raises some concern because in most instances criterion-referenced tests are characterized by limited test score variability. Because K varies directly with the variability of the test, it is expected that the values obtained for K in the present study may be somewhat restricted. Research by Eignor and Hambleton (1979) resulted in findings which tend to support this assumption.

Item Selection Procedures

Item selection procedures have been shown to have an effect on test score distributions and the reliability indices of criterion-referenced tests. This factor has been controlled for and is therefore not an issue in the present study. However, a brief summary of some of the literature related to item selection procedures and their effects on criterion-referenced measurement is presented here to establish a rationale for the item selection strategy employed in this investigation.

Generally, in the construction of criterion-referenced tests, the items are randomly selected from an item bank representative of some prespecified domain of behaviors (Nunnally, 1967; Lord & Novick, 1968; Popham, 1978; and Haladyna and Foid, 1981). However, other theorists, Wood (1968) and Reckase (1981), suggest that criterion-referenced tests should consist of items which are at the 50% difficulty level of the population tested. A third school of thought regarding item selection procedures for criterion-referenced test is that the items selected

from the item bank should be targeted at the cut-score of the test which is used to make the decisions for classifiying the examinees (Hambleton, 1980 and Phillips, 1982).

Each of the item selection strategies described above is based on sound theoretical assumptions. Randomly selecting the items from an existing item pool allows the examiner to generalize the results obtained from the item sample to the universe of items representing the specified domain of behavior represented by the item bank. Research conducted by Haladyna and Roid (1979) concluded that smaller errors of measurements were observed on criterion-referenced tests developed by randomly selected items from an item pool than by using item parameters to target the items on the test.

Selecting items for the test which are targeted to the mean ability of the population of examinees tested generally results in tests characterized by higher reliability and should provide estimates of examinee ability with greater precision than tests composed of randomly selected items (Haladyna and Roid, 1981). It should be noted, however, that the reliability of test scores for samples of the population at the extremes of the test score distribution will be considerably less than that for students clustered about the mean.

Where tests are used solely for the purpose of classifying examinees into two categories (e.g., pass/fail, mastery/nonmastery. etc.) based on whether their test scores are above or below some preset standard or cut-score, concern for reliability, in terms of measurment error, is directed at the error band surrounding the score about which

the classifications are made. It is obvious that by minimizing this error band the examiner can reduce the number of examinees for whom a pass/fail classification would be tentative. Phillips (1982) has shown that by targeting the items comprising the test at its cut-score, the error of measurement is minimized at that point in the score distribution.

Test Length

While the literature abounds with studies centered around the use of item response models in test development and test analysis in general, few studies have been directed toward the use of this technology for the express purpose of reducing testing time and student-testing burden. Discussed in this section are several studies which investigated the relationships between test length and student ability estimates and pass/fail classifications.

Eignor and Hambleton (1979) explored the relationship between test length and indices of criterion-referenced test reliability and validity. The specific reliability and validity indices they investigated include decision consistency, Kappa, decision accuracy, predictive validity, and efficiency. Several of these indices were also used in the present study (see the section on methodology for greater detail). By using computer-simulated item response data, five domain score distributions were generated which fit both the binomial model and the compound binomial model. These distributions differed with regard to skewness, ranging from moderate negative skewness to very high negative skewness. Their research was directed at two major

areas of interest:

1. The relationship between advancement scores and several reliability and validity indices for several test lengths in five different domain score distributions.
2. The relationship between test length and reliability and validity indices from a fixed cut-score in five domain score distributions. (p. 3)

The results from the study by Eignor and Hambleton (1979) showed that both the domain score distributions and the length of the test had considerable influence on decision consistency, decision accuracy, and efficiency. Each of the indices studied varied directly with the length of the test utilized. Kappa values were found to be highest when the advancement scores were near the middle of the distribution.

Garrison and Coggiola (1980) conducted a study to identify a subset of items from a larger test which would provide the most nonredundant information about examinee performance levels. Researchers in this particular study also used item difficulty estimates and pupil ability estimates computed by using the Rasch Model.

The results from the study revealed that the number of test items required to estimate pupil ability levels on four different tests could be reduced an average of 54 percent with only negligible effects on test reliability. Tests initially consisting of 60, 86, 60, and 50 items were reduced to 25, 42, 22, and 29 items, respectively, by selecting from the original tests those items whose difficulty

estimates approximated the mean ability of the student sample tested. Correlations between the results produced by the shortened forms of the test and those from the long forms of the test ranged form .94 to .97. The internal consistency reliability estimates of the short-form tests and the long-form tests averaged .88 and .93, respectively.

It was concluded by the investigators in this study that control over the distribution of item difficulty indices produced test instruments requiring substantially fewer stimuli than the number contained on the original test instruments. Further, over a large number of test cases, the item difficulty information obtained from the Rasch calibrated procedures could be used to determine characteristics of a "best" test for a particular population with whom the test would be used in future assessments. It should be noted, however, that the procedures used in the study by Garrison and Coggiola (1980) differed from some of the other studies reviewed and the present study in several potentially important ways: (1) the sample of examinees consisted of approximately 600 deaf students entering the National Technical Institute for the Deaf, (2) the process used to reduce the length of the test consisted of eliminating items which varied most from the ability level of the examinees (targeting the test to the ability level of the population tested), and (3) only one of the four tests used could be considered as a criterion-referenced test.

A study by Slawski and Bauer (1978) concluded that a substantial reduction in testing time and cost could be realized by using short-form tests developed by using item response procedures to make

mastery/nonmastery decisions about student performance. Results from their study show that by using a subset of 10 items, selected from a 95-item test, mastery/nonmastery decisions made by using the short-form test and the original test agreed for about 81 percent of the examinees tested. Correlation between the two measures was approximately .67. The items selected for the short-form test in this study were the 10 most difficult items on the original 95-item test. However, because the subjects included in the sample of examinees were high ability students, this procedure approximates that of targeting the items to the ability of the sample of students tested. On both the short and long forms .pa of the test, a score of 80 percent correct was used as the cut-score for mastery classifications.

The study by Slawski and Bauer (1978) also revealed that by using the 95-item test as the criterion, the short-form test produced 14.9% fewer false-negative classifications and 1.6% more false-positive classifications. According to the investigators in the study, the short-form test imposed a more rigid criterion than the longer version of the test. This prognosis was anticipated, however, given that the overall difficulty of the short-form test was higher. It was concluded that the short form performed adequately for making student mastery/nonmastery decisions regarding student competency levels in reading.

Haladyna and Roid (1981) conducted a study to examine the relationship between test length and measurement error. The dependent variables in this study were the average absolute differences conceived by Hambleton, Hutton, and Swaminathan (1976) and the ratio of error

variation and true score variation (Haladyna and Roid, 1976, p. 10). The researchers in this study developed tests varying in content, sensitivity to instruction, and length. The initial item subsets consisted of items representing objectives for first-year dental students and items representing objectives taught in elementary school. The instructional sensitivity of the tests was based on the difference between pretest scores and posttest scores of examinees used in the study. Short-form tests of 10, 20, 30, and 40 items were selected in two ways: randomly sampling items from the available item pools and selecting items targeted to the average ability of the group tested. Test scores on the shortened versions of the test were derived from student responses to items on the original form of the test.

Results from the study revealed a relationship between test length and measurement error, with smaller errors being associated with the longer forms of the test. Further, the study revealed that test length accounted for as much as 50% of the variance observed between the measures analyzed.

Summary

The review of the literature conducted in this study suggests that the potential benefits which may be possible as a result of using item response theory in test development and analysis have yet to be fully realized. While the use of this technology appears to be feasible in reducing testing time, cost, and pupil burden, it is but one of the many areas of measurement in which the theory may be applicable.

Each of the studies reviewed in this investigation regarding test length has contributed to the knowledge of test development using item response theory. However, several relevant questions still remain unanswered. The conclusions rendered in the study by Eignor and Hambleton (1979) were the result of using simulated data which allowed the investigators to manipulate and control all of the variables related to the distribution of test scores. These results still need to be validated in a large-scale testing program using empirical data where such controls are not possible.

In the studies conducted by Garrison and Coggiola (1980), Slawski and Bauer (1978), and Haladyna and Roid (1976), items were selected for the short-form tests which were targeted to the ability of the sample tested. Haladyna and Roid (1976) also randomly selected items from the existing item pool. It is acknowledged that the use of these procedures generally leads to higher total test reliabilities. However, where the purpose of the test is to make pass/fail decisions based on whether students score above or below some preselected cut-score, the greatest reliability, in terms of minimizing the error of the test, should be sought at the point where decisions are made. Remaining unanswered is the question of whether similar results might be obtained if the short-form tests were developed by using items which have been targeted at the cut-score of the test as opposed to items targeted at the perceived ability of the sample tested or items which have been randomly selected from the item selection pool.

The limitations inherent in the aforementioned studies suggest the need for a study of the type presented here. Unlike some of the

studies observed in the literature, the present study addresses questions regarding the effects of test length on pupil ability estimates, classification consistency, and test reliability indices by using empirical data obtained in a large-scale testing program where absolute control of test score distributions is not possible. Further, the present study will provide answers to questions of the relationship between test length and pupil ability estimation and test length and pass/fail classification by using tests which have been developed with items targeted at the cut-score of the test.

CHAPTER 3

METHODOLOGY

Introduction

This chapter describes the procedures which were used in conducting this investigation. Discussed in this section are the samples used, the instrumentation used to collect the data, the procedures used to develop the simulated test situations, the procedures used to estimate pupil ability, and the statistical procedures used to analyze the data. The sequence of activities used in this investigation was organized to allow the investigator to logically render a conclusion regarding the efficacy of utilizing short-form basic skills competency tests to make certification decisions about students in the state of Maryland. The data analysis procedures used in the study consist of a combination of classical statistical techniques and those techniques inherent in Item Response Theory. The following figure depicts the sequential order of the activities used in this investigation and facilitates the subsequent discussion of the procedures used.

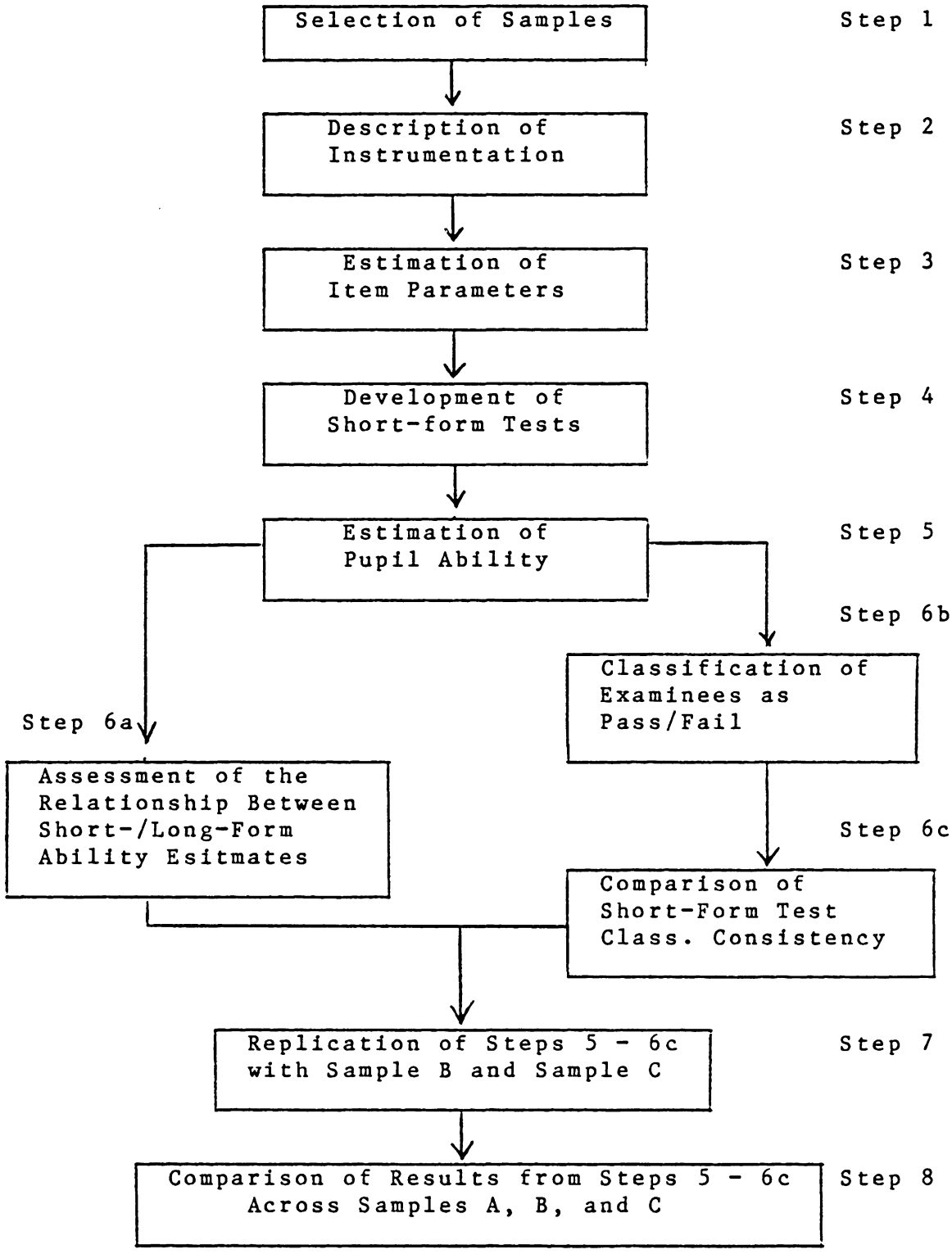


Figure 1. Procedures for conducting the study

Samples

The total sample used in this study initially consisted of 36,410 ninth grade students who were enrolled in the Maryland Public School System during the fall of 1982. Approximately 3,610 students were eliminated from this group because of missing data or lack of fit to the model. Specifically, 290 students were eliminated because they had a total fit statistic greater than 3.0, the level selected for this study. It should be noted that there are several reasons why the responses of these students did not fit the model. According to Smith and Mitchell (1980), several of the reasons student responses do not fit the model include "guessing, plodding, item-type interactions, and sloppiness." No analysis was performed to determine the nature of the misfitting students. The remaining 3,300 students eliminated from the original sample were dropped due to missing or uninterpretable item responses. The final sample in this study contained 32,800 students. The ethnic composition of the sample was approximately 67% White Americans, 30% Black Americans, 2% Asian Americans, and 1% Hispanic Americans. A breakdown of the sample by sex showed the sample to consist of 49.6% females and 50.4% males. It is assumed that the students in this sample come from a variety of social and economic backgrounds.

Three nonoverlapping samples of 5,000 examinees each, hereafter referred to as Sample A, Sample B, and Sample C, were randomly selected from the total sample. Sample A was used for the initial data analysis, and Samples B and C were used to validate the results obtained in the initial analysis. The initial data analysis activities

consisted of comparing the ability estimates and pass/fail classifications obtained by using the short-form tests with those obtained by using the original form of the test. The distribution of ability estimates resulting from the original form of the test is shown for each of the samples in Table 1. Provided in Appendix A is a histogram illustrating the shape of the distribution for each of the three samples. Superimposed on each histogram is an SPSS-generated image of the normal curve.

Table 1

Mean Ability Estimates of Students on the Original Form of the MFRT

Statistics	Sample A	Sample B	Sample C
N	5,000	5,000	5,000
Mean[a]	1.790	1.789	1.816
S. D.	1.054	1.042	1.081
SEM	0.015	0.015	0.015
Range	8.029	8.029	7.916
Skewness	-0.059	-0.105	-0.089

a/ Ability estimates are on the logit scale.

As shown in Table 1 above and the histograms in Appendix A, the distributions of scores for the three samples are almost identical. The distribution in each case is slightly negatively skewed.

Instrumentation

The test data analyzed in this study were those which resulted from the 1982 fall administration of the Maryland Functional Reading Test. The test, which is one component of the Maryland Functional Testing Program, is administerd annually to determine the extent to which pupils have achieved a predetermined level of proficiency on a set of state-mandated objectives and is used as a criteria for graduation certification.

The MFRT consists of 75 items which were selected from a bank of items which have been judged to be representative of the the functional reading objectives prescribed by the state. The items were developed to measure pupil competency in five functional reading categories: Locating References, Understanding Forms, Gaining Information Using Details, Gaining Information Through the Main Idea, and Following Directions. Items used on the test have been validated by over 15 different citizen groups, reading teachers, and curriculum specialists. As such, it is assumed that the items comprising the test are representative of the domain of behaviors required by the state. It should be noted that while the items comprising the original form of the test were selected to be representative of these five reading categories, the concerns of the present study were directed at estimation of the total score and may not contain a representative number of items for each of the categories.

A passing standard of 80% of the items correct was set by the state on the original form of the test. Subsequent forms of the test

have used scale scores for making pass/fail decisions because of the variability of percentage scores across tests which differ in overall difficulty. In the present study, a score of 80% correct was also selected as the cut-score for making pass/fail decisions. The logit score equivalent to 80% correct on this test was 1.71 logits.

The assumptions of unidimensionality and local independence of the test which are required by the model have each been tested by the Maryland Functional Testing Program by using factor analysis techniques proposed by Hambleton & Traub (1973), Lord (1968), and Hambleton & Cook (1977). The fit of the test data to the one parameter model of IRT has also been established by the department by using the Chi-Square procedures developed by Wright and Panchapakesan (1960). The assumption is therefore made that the fundamental assumptions required for use of the model are satisfied.

Estimation of Item Parameters

The items administered to students on the total test were calibrated by using Wright and Mead's Bical Computer Program in step 3 of the procedures to obtain estimates of their item difficulty parameters. BICAL is a FORTRAN computer program which calibrates the test items by using examinee responses to produce estimates of examinee ability and item difficulty parameters. The program also assesses the fit of the test items and examinees according to the one parameter model of IRT (Wright & Mead, 1977). The initial BICAL run resulted in the elimination of 290 misfitting persons (Total fit T above 3.00)

after which the items were recalibrated. See Appendix B for Fit Statistic and Distribution.

Although several procedures exist which could be used to estimate item parameters, the BICAL Program estimates were made by using Wright's corrected unconditional maximum likelihood estimation and Cohen's normal approximation procedures. Table 2 displays the item parameters obtained from the application of these procedures in the second calibration.

Table 2

Item Parameters Resulting From Calibration

of the Original Form of the Test

Item Name	Item Diff.	SEM	Disc. Index	Item Name	Item Diff.	SEM	Disc. Index
I16	-2.564	.035	1.098	I29	0.061	.014	0.925
I26	-2.353	.032	1.034	I17	0.147	.014	0.762
I38	-2.244	.030	1.191	I60	0.167	.014	1.189
I1	-2.067	.028	1.039	I20	0.176	.014	0.682
I25	-1.839	.026	0.939	I64	0.221	.014	1.218
I66	-1.705	.024	1.341	I90	0.312	.014	1.145
I2	-1.439	.022	1.072	I93	0.336	.013	1.097
I72	-1.343	.021	1.315	I89	0.376	.013	1.159
I70	-1.339	.021	1.329	I73	0.384	.013	1.074
I9	-1.314	.021	1.207	I22	0.456	.013	0.553
I6	-1.311	.021	1.088	I77	0.474	.013	0.571
I63	-1.239	.021	1.414	I85	0.488	.013	1.141
I28	-1.171	.020	0.842	I94	0.498	.013	1.051
I15	-1.130	.020	1.172	I11	0.504	.013	0.684
I24	-1.052	.019	0.842	I79	0.556	.013	0.957
I13	-0.995	.019	0.736	I44	0.572	.013	1.019
I5	-0.980	.019	1.168	I75	0.576	.013	1.136
I87	-0.899	.018	1.406	I74	0.595	.013	0.880
I21	-0.872	.018	1.203	I95	0.619	.013	1.188
I12	-0.779	.018	0.986	I84	0.663	.012	0.922
I80	-0.718	.017	1.160	I50	0.736	.012	0.954
I37	-0.689	.017	1.240	I48	0.798	.012	1.393
I3	-0.660	.017	0.622	I67	0.801	.012	1.226
I68	-0.623	.017	1.121	I58	0.829	.012	1.182
I62	-0.579	.017	1.329	I53	0.834	.012	1.251
I40	-0.508	.016	1.108	I46	0.859	.012	1.194
I18	-0.482	.016	1.020	I43	1.284	.012	0.814
I71	-0.472	.016	1.105	I32	1.388	.012	0.560
I36	-0.466	.016	1.157	I51	1.442	.012	0.781
I41	-0.421	.016	1.157	I81	1.449	.012	1.051
I30	-0.328	.016	1.154	I54	1.471	.012	0.894
I61	-0.272	.015	1.121	I55	1.522	.012	1.031
I31	-0.230	.015	1.213	I56	1.543	.012	1.237
I33	-0.149	.015	1.024	I42	1.588	.012	0.785
I8	-0.115	.015	1.037	I7	1.967	.012	0.784
I34	-0.108	.015	1.038	I78	2.658	.013	0.162
I33	0.034	.014	0.628	I92	2.788	.013	0.782
				I47	3.252	.014	0.351
				Mean	0.0000	.016	1.022

The item discrimination indices computed by the BICAL program and shown in Table 2 above represent the linear trend across score groups (Wright & Mead, 1977). According to Wright and Mead, "values larger than one indicate that the observed characteristic curve for an item is steeper than the average best fitting logistic curve for all items; values less than one indicate the curve is flatter" (p. 53).

It would appear to be obvious from the statistics presented in Table 2 that based on classical test theory, this particular test could very readily be reduced by approximately 25% by eliminating those items with very low discrimination indices. However, according to Wright and Mead, "whether discrimination indices are relevant to stable, meaningful item parameters that are useful in describing future outcomes in similar situations or whether they are only useful for diagnosing potential problems in a single set of observations remains unanswered" (p. 10). Wright and Mead further propose that the concept of specific objectivity does not apply to the item discrimination indices.

While it is conceded that items with lower discrimination indices contribute little to the measurement of the examinees upon which they are computed, the fact that they vary does not result in poorer calibrations of either persons nor items (Dinero & Haertel, 1976). Hence, item discrimination indices were not used as a criteria in the selection of items for the short-form tests developed in this study.

Development of Short-Form Tests

In step 4 of the procedures, development of short-form tests, two important issues had to be resolved. The first issue to be resolved

was how many items to include on each of the short forms, and the second issue was which item selection procedure to use in selecting the items for each of the short-form tests.

Because of the applied nature of the study, the decision was made to develop three test forms consisting of 10, 20, and 30 items selected from the 75 items comprising the original form of the test.

The rationale for developing tests of the lengths described is inherent in the applied nature of this study. It is foreseeable and therefore assumed that since the original test is designed to be completed in two class periods, a test approximately half its length could be completed in half the time or one class period. While a test reduced by any length has the potential of reducing the testing time in real time, to reduce the test by 5 minutes or 20 minutes would be of little or no practical significance in the normal flow of the school's daily program. The greatest benefit will be received if the disruption of the schools program is reduced by an entire class period unit of time, which is approximately one-half the time presently consumed by administering this test.

A 30-item test was selected as the maximum length of a short-form test for two important reasons. The first reason is related to the fact that the time allotted for administration of the test is also the time that new items are field tested to expand the existing item bank. A 30-item test would allow for the inclusion of five to ten experimental items and still permit the test to be administered within one class period. Secondly, while the other tests of 10 and 20 items

may be viewed as potentially viable as short forms themselves, they were selected for the purpose of observing the relationship between test length and the estimation of person ability. It is not likely that a test of fewer than ten items would have sufficient face validity to be viewed as valid or worthy of taking by either examinees or school system decision makers.

The second issue in this phase of the study was deciding which strategy to use to select the items for each of the tests. It is generally accepted that the greatest reliability for a test used for making pass/fail or mastery/nonmastery decisions is sought at the cut-score used for making those decisions. This is usually accomplished by targeting the items comprising the test at the cut-score established for the test. Further when the items are targeted at the cut-score, the error band surrounding the cut-score is minimized. Therefore, in the present study, items selected for the short-form tests were those which were closest to the cut-score. To prevent the potential effects of item dispersion about the cut-score and the average difficulty of the test forms from confounding the subsequent analyses of the results produced by the short forms of the tests, these variables were controlled for by constructing the tests with approximately the same average distance from the cut-score and the same average difficulty level.

It was also important in this investigation to minimize the intercorrelation among the short forms of the test due to overlapping items. Therefore, different items were selected for each of the three short forms developed, a relatively conservative procedure.

The tasks described above regarding the characteristics of the short-form tests developed in this study were accomplished in several steps. First each item's distance from the cut-score was computed by subtracting its difficulty parameter from the cut-score used on the test and the items rank ordered by the absolute value of that quantity.

In the second step, the total number of items required to produce the three nonoverlapping test forms, each consisting of the number of items required, was determined. For the present study, 60 items were required to produce three non-overlapping short forms of 10, 20, and 30 items.

The next step required selecting items so that each test would initially consist of items evenly distributed across the range of computed distances from the cut-score of the test. This was accomplished by systematically sampling from the 60 items rank ordered by their distance from the cut-score. Every other item was selected for the 30-item test and from the remaining 30 items every third item was selected for the 10-item test. The remaining 20 items comprised the 20-item test. Computation of the average item difficulty after this step revealed an average item difficulty of 0.161 logits, .0498 logits, and .417 logits for the 10-, 20-, and 30-item tests, respectively. The final step involved interchanging selected items among the short-form tests to produce test forms which were equal in average item difficulty. Table 3 shows the item and test statistics resulting from these procedures.

Table 3

Item Parameters for Short-Form Tests

10 Item Test				20 Item Test				30 Item Test			
Item Name	Item Diff.	SEM	Disc. Index	Item Name	Item Diff.	SEM	Disc. Index	Item Name	Item Diff.	SEM	Disc. Index
I37	-0.689	.017	1.248	I13	-0.995	.019	0.736	I5	-0.980	.019	1.168
I71	-0.472	.016	1.105	I87	-0.899	.018	1.406	I21	-0.872	.018	1.203
I33	-0.149	.015	1.024	I12	-0.779	.018	0.986	I80	-0.718	.017	1.160
I17	0.147	.014	0.762	I68	-0.623	.017	1.121	I3	-0.660	.017	0.622
I90	0.312	.014	1.145	I40	-0.508	.016	1.108	I62	-0.579	.017	1.329
I11	0.504	.013	0.684	I41	-0.421	.016	1.157	I18	-0.482	.016	1.020
I95	0.619	.013	1.188	I61	-0.272	.015	1.121	I36	-0.466	.016	1.157
I58	0.829	.012	1.182	I34	-0.108	.015	1.038	I30	-0.328	.016	1.154
I51	1.442	.012	0.781	I20	0.176	.014	0.682	I31	-0.230	.015	1.213
I54	1.471	.012	0.894	I93	0.336	.013	1.097	I8	-0.115	.015	1.037
				I89	0.376	.013	1.159	I83	0.034	.014	0.628
				I22	0.456	.013	0.553	I29	0.061	.014	0.925
				I85	0.488	.013	1.141	I60	0.167	.014	1.189
				I44	0.572	.013	1.019	I64	0.221	.014	1.218
				I48	0.798	.012	1.393	I73	0.384	.013	1.074
				I46	0.859	.012	1.194	I77	0.474	.013	0.571
				I56	1.543	.012	1.237	I94	0.498	.013	1.051
				I42	1.588	.012	0.785	I79	0.556	.013	0.957
				I78	2.658	.013	0.162	I75	0.576	.013	1.136
				I92	2.788	.013	0.782	I74	0.595	.013	0.880
								I84	0.663	.012	0.922
								I50	0.736	.012	0.954
								I67	0.801	.012	1.226
								I53	0.834	.012	1.251
								I43	1.284	.012	0.814
								I32	1.388	.012	0.560
								I81	1.449	.012	1.051
								I55	1.522	.012	1.031
								I7	1.967	.012	0.784
								I47	3.252	.014	0.351
Means	**0.401**	**.014**	**1.00**		**0.402**	**.014**	**.99**		**0.401**	**.014**	**0.99**

Note: Item difficulty estimates, standard error estimates, and discrimination indices are from calibration of the original form of the test.

The data in Table 3 illustrate that the test forms developed in this study were identical in terms of the attributes relevant to parallel test forms. The only obvious distinction among the test forms is test length which is the attribute investigated in this study.

Estimation of Functional Reading Ability Scores

Step 5 of the procedures in this study, estimation of person ability, was conducted in two stages. Stage 1 involved estimating the item difficulty parameters by using BICAL.

Stage 2 of this process, estimation of examinee ability by using the responses to items contained on each of the short forms of the test, was conducted by using a computer program developed by Gary Phillips and Sandy Gedieck with the Maryland State Department of Education. These procedures were used so that the ability estimates resulting from each of the test forms used in this study would be on the same scale. The program uses the item parameters obtained from the initial BICAL computer run and performs the Newton-Raphson iteration procedures on each subset of items to produce ability estimates for each of the raw scores resulting from each of the test forms and also a seperate test characteristic curve for each form of the test. Each of the resulting test characteristic curves was based only on the subset of items comprising the short form of the test analyzed. For this study four such test characteristic curves were generated, one for the 10-item test, one for the 20-item test, one for the 30-item test, and one for the total test (75 items). See Appendiices C - F for detailed illustrations. The raw scores, percent correct scores, and the equivalent logit scores which were generated with the test

characteristic curves are shown in Table 4. Also shown in the table is the standard error of measurement for each of the scores in the distribution, a feature which is unique to models of this type. The test characteristic curves derived in this step of the procedures are shown in Appendices C - F.

Table 4

Logit Ability Estimates for Short-Form Tests

10-Item-Test			20-Item-Test			30-Item-Test		
Raw Score	Domain Score	Logit Score	Raw Score	Domain Score	Logit Score	Raw Score	Domain Score	Logit Score
0	0	-3.131	0	0	-3.977	0	0	-4.327
1	10	-1.979	1	5	-2.928	1	3	-3.280
2	20	-1.123	2	10	-2.150	2	6	-2.534
3	30	-0.539	3	15	-1.655	3	10	-2.073
4	40	-0.052	4	20	-1.273	4	13	-1.728
5	50	0.399	5	25	-0.951	5	16	-1.447
6	60	0.851	6	30	-0.663	6	20	-1.205
7	70	1.339	7	35	-0.397	7	23	-0.989
8	80	1.926	8	40	-0.145	8	26	-0.791
9	90	2.785	9	45	0.101	9	30	-0.607
10	100	3.942	10	50	0.345	10	33	-0.433
			11	55	0.592	11	36	-0.266
			12	60	0.847	12	40	-0.104
			13	65	1.114	13	43	0.054
			14	70	1.401	14	46	0.211
			15	75	1.717	15	50	0.366
			16	80	2.078	16	53	0.522
			17	85	2.507	17	56	0.684
			18	90	3.061	18	60	0.842
			19	95	3.909	19	63	1.008
			20	100	5.065	20	66	1.180
						21	70	1.362
						22	73	1.556
						23	76	1.765
						24	80	1.996
						25	83	2.257
						26	86	2.562
						27	90	2.937
						28	93	3.438
						29	96	4.239
						30	100	5.347

Data Analysis

In step 6a the assessment of the relationship between the person ability estimates obtained from each of the short forms of the test and the original form of the test was conducted by using correlation analysis. Several steps were required to conduct this phase of the analyses. First, Product Moment Correlation Coefficients were computed between each of the short forms of the test and the original form of the test by using the procedures contained in SPSS 8th edition.

Second, the reliability of the test forms analyzed in this study was estimated by using two indices of test reliabilty. The first index of reliability was obtained by using the Kuder-Richardson-20 (KR-20) formula to determine the internal consistency of the test forms investigated in this study. It should be noted, however, that the Kuder-Richardson formulas, which utilize the variance of the total scores of the tests and the sum of the item variances, tend to under estimate the reliability of criterion-referenced tests due to the limited variance generally associated with these measures (Popham & Husek, 1969; Hambleton & Novick, 1973).

A second index of reliablility, purportedly more sensitive to the limited variation associated with criterion- referenced measurement, was also computed for each form of the test by using an adaptation of Subkoviak's procedures developed by Phillips and Gediek (1983). The reliability index obtained by using these procedures represents the coefficient of agreement as the probability that an examinee will be assigned to the the same pass/fail category on parallel forms of the test. However, the index is computed from a single administration of

the test and, therefore, approximates the probability that an examinee will be assigned to the the same pass/fail category on a second administration of the same test. As such, the index represents the extent to which the instrument (original form of the MFRT) is in agreement with itself. When the criterion is equal to C, 1.71 logits in this study, based on Subkoviak's method, the coefficient of agreement for person 'i', is as follows:

$$P_C^{(i)} = P(X_i \geq C, X'_i \geq C) + P(X_i < C, X'_i < C) \quad (1)$$

where,

P = Probability

C = Criterion Score

X = Pass/Fail Classification on Test 1

X'= Pass/Fail Classification on Test 2

(Subkoviak, p.267, 1976)

For a sample of examinees, the probability of consistently classifying them to the same mastery/nonmastery category is derived from the following formula:

$$P_C^{(i)} = (\sum_{i=1}^{N} P_C^i)/N \quad (2)$$

According to Subkoviak (1976), the generalization stated above holds under two fundamental assumptions. The first assumption, proposed by Lord and Novick (1968), is the requirement that the scores obtained by examinees on the parallel forms must be independent. That is to say that taking one form of the test does not affect taking the

second form of the test. Based on this assumption, equation (1) is equal to

$$P_C = P(X_i > C).P(X'_i > C) + P(X_i < C).P(X'_i < C) \quad (3)$$

The second assumption also proposed by Lord and Novick (1968) is that the binomial distribution of scores on the two test forms is the same. Hence X is equivalent to X'. Subkoviak (1976) acknowledges that the binomial distribution assumption required by the model is simplistic. However, he explains that it contains sufficient flexibility to approximate the X_i distribution. Under the two assumption stated above, equation (3) can be simplified to equal the following:

$$P_C^i = [P(X_i \geq C)]^2 + [P(X_i < C)]^2 \quad (4)$$

$$\text{Since } P(X_i < C)] = 1 - P(X_i \geq C) \quad (5)$$

$$P_C^i = [P(X_i \geq C)]^2 + [1 - P(X_i \geq C)]^2 \quad (6)$$

where,

$$P(X_i > C) = \sum_{x_i=c}^{n} \binom{N}{x_i} P_i(1 - P_i)^{n-x} \quad (7)$$

(Subkoviak, p.268, 1976)

The quantity P_i represents the true probability of a correct item response for person "i" which can be approximated from X_i, the observed score on a single test (e.g., $p_i = X_i/N$) (Subkoviak, p.268, 1976). Using equations (5) and (6), this probability can be computed for each examinee and computed for the total group by using equation (2). While Subkoviak's procedures are based on the binomial distribution model

which uses the raw score metric and assumes equal item difficulty, the present study applied these procedures based on the Rasch model by using the logit item difficulty values obtained from the BICAL program as proposed by Phillips (1983).

In step 6b examinees in sample A were assigned to pass/fail states by using the original form of the test and each of the three short forms of the test. For the purposes of this investigation, students with ability estimates greater than or equal to 1.71 logits were classified as pass; and examinees with an ability estimate less than 1.71 logits were classified as fail. It was the decision of the investigator that examinees with scores within the range described be classified as pass or fail; however for all intents and purposes, examinees within the interval of 1.71 examinees plus or minus the error of measurement at 1.71 could be classified as pass or fail, depending on the judgment of the examiner. It is also within this interval (cut-score $\pm$ sem) that misclassification of an examinee is of minimum consequence. According to Wilcox (1980) any classification within this interval is correct. It is, therefore, obvious that the classification decisions regarding examinees outside this interval that are of paramount importance.

The following figure shows how the proportion of examinees classified as pass/fail was used to assess the classification consistency of the test forms analyzed in this study.

		Original Test		
		F	P	
Short-Form	F	P_{00}	P_{01}	$P_{0\cdot}$
	P	P_{10}	P_{11}	$P_{1\cdot}$
		$P_{\cdot 0}$	$P_{\cdot 1}$	

P = Pass

F = Fail

Figure 2. Contingency table for analysis of short- and long-form classification consistency

This figure represents a 2 x 2 contingency table showing the proportion of examinees assigned to the four possible categories of pass/fail by using the results from the original form of the MFRT and one of the short-form tests developed for this study. Relative to the original form of the test, P_{01} represents the proportion of "false-negatives" resulting from classifications by using the short-form of the test. The variable P_{10} represents the proportion of "false-positives" resulting from classification decisions by using the same test. Cells containing P_{11} and P_{00} represent the proportion of examinees classified as pass or fail by each of the test instruments, respectively. By using the data in this diagram, several indices could be derived. However, to determine the extent to which the

classification decisions made by using each of the two tests differ, the indices of interest in this study and thus computed were the observed coefficient of agreement and coefficient Kappa. The observed coefficient of agreement was computed by adding the proportion of examinees consistently classified as pass or fail by both instruments (e.g., observed coefficient of agreement = $(P_{11} + P_{00})$.

Coefficient Kappa was computed by using the procedures proposed by Swaminathan, Hambleton, and Algina (1974). Those procedures are the following:

Kappa = K = (DC CA)/(1 CA)

where, DC = P_{kk},

CA = Chance Agreement = $\sum_{k=0}^{1} P_{k\cdot}\, P_{\cdot k}$

$P_{0\cdot}$, $P_{1\cdot}$, $P_{\cdot 0}$, and $P_{\cdot 1}$ are the marginal proportions for the short-forms of the test and the original-forms of the test, respectively.

The index obtained for DC ranges from 0 to +1 indicates the overall consistency of the pass/fail classifications resulting from the pairs of measures analyzed. The value of Kappa ranges from -1 to +1. A value of positive one is obtainable only when the marginals of the contingency tables are the exact same (Swaminathan et al., 1975). The extent to which examinees are classified differently is reflected in values of Kappa less than +1. Positive values indicate that the proportion of agreement is better than chance, zero shows that the proportion of agreement is exactly equal to chance, and negative values indicate that the proportion of agreement is less than chance. The extent to which the obtained Kappa was significant was determined by using the traditional Chi-Square statistic.

Preliminary Validation Procedures

The preliminary validation of the results obtained from this study was conducted by replicating the analyses performed by using Sample A with two other ramdomly selected samples of 5,000 students (Sample B and Sample C).

CHAPTER 4

RESULTS AND DISCUSSIONS

Introduction

The purpose of this study was to examine the effects of test length on estimating pupil functional reading ability and making pass/fail classifications of ninth grade students in the state of Maryland. The tests and the test items analyzed in this study were calibrated by using the one parameter model of item response theory as described by Georg Rasch (1960).

This section presents the results of the data analysis procedures described in chapter 3 of this investigation. The first section describes findings regarding the estimation of the functional reading ability of the samples investigated in this study, and the second section describes the results from the correlational analyses of those estimates. In section 3 results describing the classification consistency of the short-form tests are presented in conjunction with results obtained from analysis of students misclassified by the test forms. A summary of the overall findings resulting from this study concludes this chapter.

Estimation of Person Ability

Ability scores were calculated for each of the subjects in three nonoverlapping samples ($N_A = N_B = N_C = 5{,}000$) by using each of three short-form tests and the original form of the MFRT. The mean ability estimates, standard deviations, and sample sizes resulting from the administration of these test forms are shown in Table 5.

Table 5

Means, Standard Deviations, Range, and Sample Sizes for Short and Long Forms of the MFRT

Test Statistics	10-Item Test	20-Item Test	30-Item Test	75-Item Test
Sample A				
Mean	1.85	1.83	1.81	1.79
S. D.	1.28	1.19	1.11	1.05
Range	7.07	7.22	7.88	8.03
N	5,000	5,000	5,000	5,000
Sample B				
Mean	1.86	1.84	1.81	1.79
S. D.	1.25	1.18	1.11	1.04
Range	7.07	7.22	7.42	8.03
N	5,000	5,000	5,000	5,000
Sample C				
Mean	1.86	1.87	1.83	1.82
S. D.	1.31	1.22	1.12	1.08
Range	7.07	7.99	7.88	7.92
N	5,000	5,000	5,000	5,000

A comparison of the mean ability estimates across test forms in Table 5, contrasting the short-form test results with the results obtained from the long-form test and the results of each of the test forms across samples by using multivariate analysis of variance repeated measures procedures, shows that the mean ability estimates produced by the short forms of the test differed significantly from the results produced by the original form of the test across all samples. The mean ability estimates produced by each of the individual tests analyzed did not differ across samples. It should be noted, however, that while the observed differences between the average ability estimates produced by the short-form tests were statistically significant, they were extremely small. The results of those analyses are shown in Table 6.

Table 6

Summary Table for MANOVA Repeated Measures Analysis of Functional Reading Ability Estimates Between Short and Long Forms of the MFRT

Design on Sample	Multivariate F	Design on Measures			
		X	10 vs. 75	20 vs. 75	30 vs. 75
X	11695.05 (.000)	216.80 (.000)	-8.83 (.000)	-10.52 (.000)	-6.17 (.000)
Sample		1.21 (.286)			
A vs. B		-0.18 (.857)	0.87 (.384)	0.93 (.353)	-0.66 (.507)
A vs. C		-1.17 (.243)	-1.17 (.241)	1.55 (.120)	-0.76 (.446)

Numbers in parentheses indicate the level of significance.

The results in Table 6 also show that across samples each of the test forms used was extremely consistent in estimating the average ability estimates of what are theoretically equivalent samples. This was evidenced by the fact that none of the test differences observed between average ability estimates obtained from any of the test forms differed significantly across the three samples used in this study.

The finding that the differences between the mean ability estimates produced by the short-form measures and those produced by the-long form test were statistically significant should not be viewed as evidence of the inefficacy of the one-parameter model. On the contrary, the facts that the average reading score differences of .06, .05, .02 logits observed between 10-, 20-, and 30-item tests and the long-form test (75 items), respectively, on a scale of 7 to 8 logits (range of the distributions) are viewed as supporting the propositions of the model. It was noted that in no instance did the observed differences amount to more than seven hundredths of the standard deviation of the shorter form of the tests compared. At most these differences can only be described as trivial, given the size of the samples investigated. The position assumed in this section of the research coincides with that of Carver (1978):

> A research finding that is small and not significant from a research standpoint can be statistically significant just because enough subjects were used in the experiment to make the result statistically rare under the null hypothesis. (p. 388)

Test Reliability

The reliability of the test instruments used in this study was

computed by using two procedures, Phillips & Gediek's adaptation of Subkoviak's procedures and coefficient Alpha. The first of these procedures was used to determine the coefficient of consistent classification, and the second was used to determine the maximum likelihood estimate of the reliability of the test instruments. A random sample of 5,000 student responses was used in the computation of both indices. The results of those analyses are shown in Table 7.

Table 7

Reliability Indices for Short and

Long Forms of the MFRT

Reliability Index	10-Item Test	20-Item Test	30-Item Test	75-Item Test
PG	.73	.78	.80	.86
Alpha	.60	.72	.80	.91

PG = Probability of Consistent Classification.

The index of reliability (PG) for the forms of the tests shown in the table computed by using the adaptation of Subkoviak's procedures is an estimate of the test retest reliability of the test from a single administration of the test. The index, which is also referred to as the coefficient of stability, represents the probability of consistently classifying a sample of examinees to the same pass/fail classification on a subsequent administration of the same test. Indices of the sizes shown in the table suggest that the test forms

used were, in fact, very reliable instruments in terms of their stability. The data analyzed in this study suggest that the size of the index is a function of the length of the test. It was noted in this investigation, however, that increasing the length of the test form 30 to 75 items, more than doubling the length, only increases the reliability coefficient (stability) from .80 to .86. This represents a rather modest increase in test reliability for a 100 percent increase in testing time.

The Alpha indices reported in the table are equivalent to the Kuder-Richardson KR-20 and represent the maximum likelihood estimate of the reliability coefficient. The index squared is equivalent to the correlation coefficient one could expect between the test results obtained on successive administrations of the same test to the same sample of examinees. As defined by Winer (1962), this reliability index represent the proportion of variance attributable to true score divided by the sum of variance attributable to true score plus error variance. Although this test is not usually recommended for use with criterion-referenced tests for reasons described in chapter 3, the alpha indices obtained for the longer forms of the test were relatively high. Further, as anticipated the reliability of the tests varies as a function of test length.

Correlations

One of the primary interests in this study was the extent to which the ability estimates resulting from the short forms of the tests were related to the the ability estimates resulting from the original form of the of the test. To make this determination, Pearson's Product

Moment Correlation coefficients were computed between each of the short-form tests and the original test. The results from those analyses are shown in Table 8. A two-variable scattergram of the data points for each of the test comparisons is also provided in Appendices H - P.

Table 8

Product Moment Correlation Coefficients

Between Short- and Long- Form Tests

	20-Item	30-Item	75-Item
Sample A			
10-Item Test	0.63	0.67	0.80
20-Item Test	----	0.74	0.89
30-Item Test	----	----	0.94
Sample B			
10-Item Test	0.63	0.67	0.80
20-Item Test	----	0.72	0.88
30-Item Test	----	----	0.94
Sample C			
10-Item Test	0.65	0.67	0.81
20-Item Test	----	0.75	0.89
30-Item Test	----	----	0.94

Each of the coefficients observed in the table was significant ($p<.000$).

Analysis of the correlation coefficients between the selected pairs of tests show that the strength of the relationship between the short forms and the long form varies directly with the length of the short-form tests. The coefficients of correlation between the short-form tests and the long-form test obtained for Sample A were .80, .89, and .94 for the 10-, 20-, and 30-item tests, respectively. The coefficients of correlation between the test forms obtained from analyses of the data with Samples B and C were similar, resulting in coefficients of .80, .88, and .94 for Sample B and .81, .89, and .94 for Sample C, respectively, with the same instruments. These results are very much similar to the results observed by Garrison and Coggiola (1980) although the strategy used to select the items for the short-form tests differed. In their study the items selected for the short-form test were targeted to the average ability of the group tested, presumably to increase the variability of test score results. However, where tests are used to make pass/fail decisions such as the case in this study, unless the mean ability of the group tested approximates the cut-score used for making the pass/fail decisions, it is very likely that the error band surrounding the cut-score will be larger than it would be if the items comprising the short forms were targeted to that cut- score. While the correlation coefficients obtained in the two studies are similar, it is highly probable that the use of short-form tests developed by using the strategy proposed and implemented in the present study will lead to fewer examinee misclassifications due the smaller error band surrounding the cut-score.

A comparison of the correlation coefficients between each of the short-form tests and the original test with the alpha index of reliability of the long form of the test reveals definite implications regarding the feasibility of using short-form tests to estimate pupil ability on the MFRT. By using the results obtained from Sample A as an example, it can be observed that the correlation between the 30-item test and the original form of the test (.94) is equivalent to the square root of the reliability index obtained for the original form of the test (.94). These results suggest that the ability estimates obtained with the 30-item test account for the same proportion of the true score variance than would be accounted for if the original form of the test were administered to the same sample of examinees a second time. Further, given the relatively high reliability (.80) of the 30-item test, a short form test containing approximately 30 items might well be a very viable alternative to the 75-item test in estimating pupil function reading ability.

Results obtained from the 20-item test (r = .89) account for a slightly smaller proportion of true variance than the test does with itself. However, the somewhat moderate reliability (.72) observed for the 20-item test makes its use as an alternative to the 75-item test questionable.

Reducing the test length to 10 items results in a 10% loss of true score variance which could be accounted for on a subsequent administration of the original test to the same sample of examinees. These results, in conjunction with the low reliability (.60) observed

for the 10-item test, make it less appealing as an alternative for the original test than either the 20- or 30-item measures.

Classification Consistency

From an administrative perspective, the consistency of the pass/fail classifications when the various forms of the tests are used is of paramount importance. Therefore, one of the objectives of this study was to investigate the relationship between test length and the classification of examinees into pass/fail categories. The fact that the correlations between selected short forms of the test and the original form of the test suggest a high degree of relationship between the ability estimates produced by these pairs of measures is not sufficient to conclude that the same pass/fail classification decisions would be made regarding students in the sample. This is especially true, given the stipulation that the same cut-score would be used to make pass/fail decisions on each of the instruments investigated. Further, the correlation coefficients are not useful in identifying the the magnitude and direction of any observed biases in terms of the proportion of examinees classified as false-positives or false-negatives when the various short forms of the tests are used. This section presents the results of analyzing the extent to which the pass/fail classification decisions made by using the short-form measures developed in this study were the same as those made by using the original form of the test.

The extent to which the pass/fail decisions made by using the short forms of the test were the same as those made by using the original test was examined by using a traditional 2 x 2 contingency

table. The observed degree of agreement between the forms of the tests assessed was computed by adding the proportion of examinees classified as passed on both the long form of the test and the short form of the test to the proportion of examinees classified as failed by both measures. However, because a certain proportion of examinees would be expected to be assigned to the same classification category due to chance alone, an adjustment of the observed coefficients was necessary. This adjustment was made by using the Kappa Coefficient of decision accuracy. The use of these procedures allowed the investigator to determine the extent to which the classification decisions made by using each of the short form tests were in agreement with the classification decisions made by using the original form of the test adjusted for chance agreement. The procedures were applied to each of the short-form/long-form comparisons across the three samples used in this investigation. Both the observed coefficients and the coefficients resulting from the application of Kappa procedures are shown in Table 9.

Table 9

Coefficients of Decision Accuracy for

Short Forms of the MFRT

	Observed Coefficient	Kappa Coefficient
Sample A		
10-Item Test	.82	.64
20-Item Test	.87	.74
30-Item Test	.90	.80
Sample B		
10-Item Test	.82	.63
20-Item Test	.86	.72
30-Item Test	.91	.82
Sample C		
10-Item Test	.83	.65
20-Item Test	.88	.75
30-Item Test	.91	.81

On the average, the observed coefficients of agreement between the short-form tests and the original test were .82, .87, and .91 for the 10-, 20-, and 30-item tests, respectively. These data show that 82% of the examinees classified as passers or failers on the original-form of the test were so classified on the 10 item test. On the 20-item test, 87% were consistently classified; and on the 30-item test, 91% were classified accurately, a figure that compares quite favorably with the percentage likely to classified consistently if the original test were administered a second time to the same sample.

Adjustment of the observed coefficients for chance as reflected by the Kappa coefficients shows a reduction in the degree of agreement,

particularly for the shorter of the three test forms. By using the 10-item test, only 64% of the examinees were classified consistently with the classifications rendered by using the original form of the test, reflecting a difference of 18% from the observed results. The observed results for the 20- and 30-item tests were diminished by approximately 13% and 10%, respectively, to 74% and 81% of the examinees being consistently classified. These data suggest a very definite relationship between the length of the test and both the observed and adjusted classification indices. This relationship was also observed in the work conducted by Eignor and Hambleton (1979) in a similar study using computer generated test score distributions. The kappa agreement indices obtained in the present study appear to be somewhat larger than those obtained by Eignor & Hambleton when the distributions of test scores were moderately high in skewness.

The Kappa indices shown above are also comparable to the test reliability indices of the long form of the test obtained by using Phillips and Gedeik's adaptation of Subkoviak's procedures. As previously discussed, this index represents the probability of a classsification of students consistent with the classification expected for a second administration of the test to the same sample. Given the coefficient of agreement computed for the original test (.86), it is obvious that the short-form tests, particularly the 30-item test, produce classificatory results which are consistent with those expected if the original form of the test were administered a second time.

Analysis of Pass/Fail Misclassifications

Where disagreement between the original form of the test and the short forms of the test in assigning students to pass or fail categories was observed, the data were analyzed to determine the nature and magnitude of the misclassifications. Particularly important in the analysis were the effects of test length on the proportion of students classified as failed on the original form of the test and classified as passed on the short-form of the test (false-positives) because students truly lacking the necessary skills to pass should be provided with remediation. Equally important, however, is the the number of false-negatives (students who have the knowledge or skills for passing but were classified as failed) because of the potential for court litigation if the results were used to deny a student his/her diploma. The most appropriate short-form of the test would be the one which minimizes both the number of false-positives and the number of false-negatives resulting from use of the test as a classification instrument. The data were also analyzed to determine the effects of reducing the length of the test on the total proportion of false-negative and false-positive classifications made by using the short forms of the test. The original form of the test was used as the criterion measure in each comparison. The results of those analyses are shown in Table 10.

Table 10

Comparison of Misclassification Errors

Misclassification Categories	Proportion Misclassified		
	10-Item Test	20-Item Test	30-Item Test
Sample A			
False-Positives	.09	.08	.05
False-Negatives	.08	.05	.02
Total	.17	.13	.10
Sample B			
False-Positives	.10	.09	.05
False-Negatives	.08	.05	.04
Total	.18	.14	.09
Sample C			
False-Positives	.09	.08	.05
False-Negatives	.09	.04	.05
Total	.18	.12	.10

Overall the data shown in Table 10 suggest a very definite relationship between the total proportion of examinees misclassified and the number of items comprising the test. The data also suggest a relationship between the length of the test and the proportion of students classified as false-positives. While a similar relationship between test length and the proportion of students misclassified as false-negatives was apparent in two of the three samples investigated (Samples A & B), it was not observed in the third sample (C). Further, where an apparent relationship was observed, it appeared to be curvilinear in form. It is obvious from the results in the table that the rate of change in the total proportion of students misclassified as the length of the test is decreased from 30 items to 20 items is not the same as the rate of change when the length of the test is decreased from 20 items to 10 items. If the relationship were linear, the rate of change would be expected to be constant. The following figure shows that as the length of the test was reduced the rate of change was not constant.

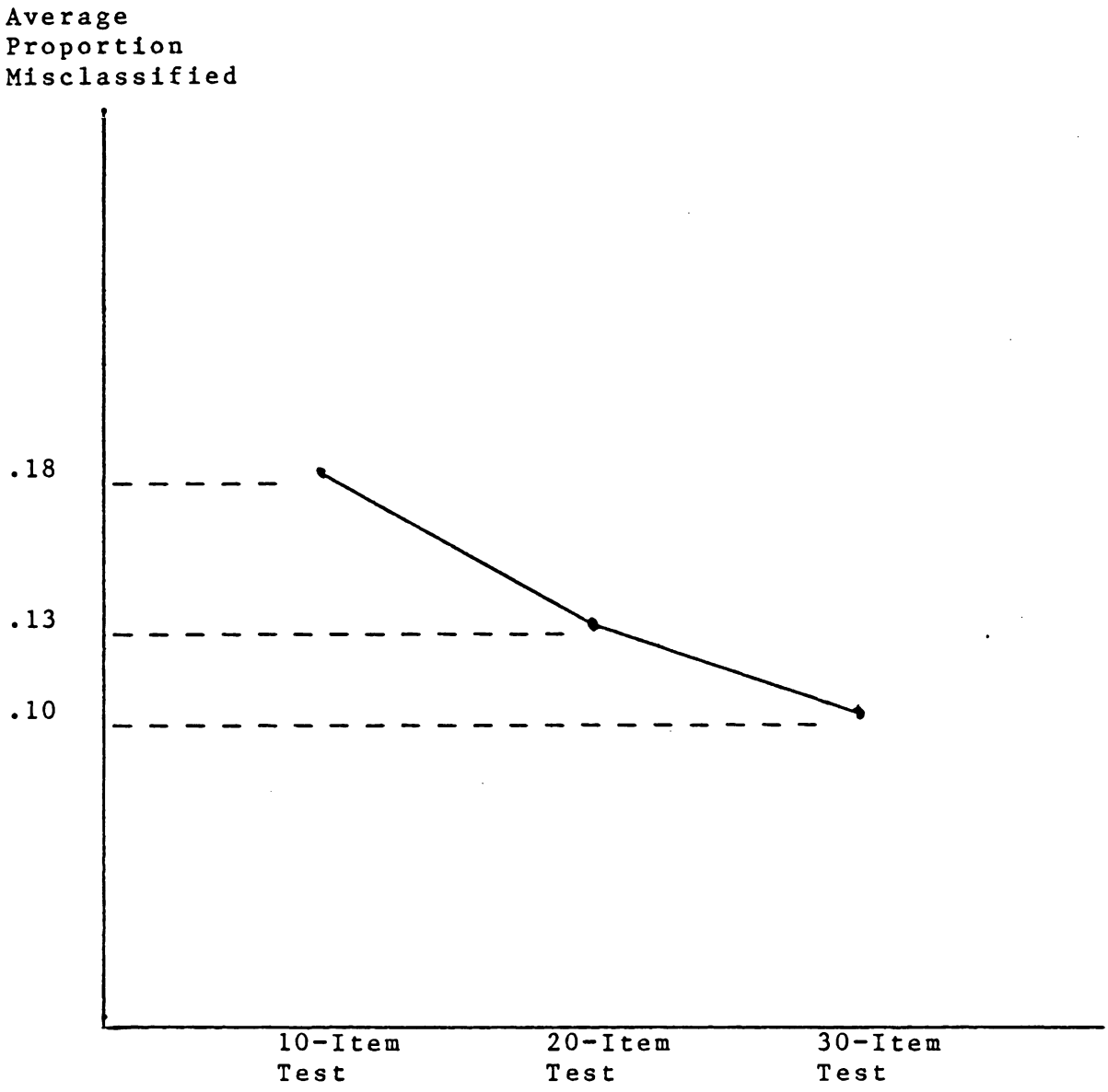


Figure 3. Total proportion of students misclassified as a function of test length

Further analysis of the data shows that more than half the students misclassified by using the short-form measures developed in this study had scores on the total test which were within the error band surrounding the cut-score of the total test. For details see table in Appendix Q. Specifically, 54%, 65%, and 80% of the examinees misclassified by using the 10-, 20-, and 30-item tests, respectively, had scores within the error band of the total test. These data suggest that the overwhelming majority of students misclassified by using any of the short-form tests were borderline cases with regard to their classification by using the total test and were students for whom any classification decision made regarding their competency would be correct or questionable. They further suggest that the least slippage is found for the 30-item test.

Analysis of the data for these same students with respect to the cut-score error bands of the short-form measures shows that no less than two-thirds of them scored within the cut-score error band on each of the tests. Virtually all of the students misclassified (98%) obtained scores within the cut-score error band of the 30-item test. This particular finding is especially important in that should future research validate these results, the cut-score error band of the short-form test can be reduced even further, thereby minimizing the number of examinees whose disposition would be tentative. Both the literature and findings from preliminary analysis of the test data used in this particular study suggest that this can very readily be accomplished by selecting the 30 items for the short-form test which

are closest to the cut-score of the test used for making the classificatory decisions.

Summary

Three different short-form tests of 10, 20, and 30 items were developed in this study to investigate the relationship between test length and the estimation of pupil functional reading ability levels and pass/fail classications in the state of Maryland. Pupil performance on the original 75-item Maryland Functional Reading Test was used as the criterion measure in each of the comparisons made in this study.

The one-parameter model, as proposed by Georg Rasch, was used to calibrate the test items and estimate pupil performance levels on the test measures investigated in this study. Pearson correlation coefficients were computed between the pupil ability estimates obtained on each of the short-form tests and the ability estimates obtained from the original form of the test to determine the strength of the relationship between the short- and long-form measures. The extent to which the pass/fail classification decisions rendered by using the short- and long-form measures were in agreement was assessed by using two classification agreement indices, the observed coefficient of agreement and Kappa coefficient of decision accuracy. The data analyses procedures were applied on three nonoverlapping samples, each consisting of 5,000 ninth grade students.

The results obtained from the correlational analyses of the data obtained in this study show a definite relationship between test length and the estimation of pupil functional reading levels. Correlation

coefficients computed between results obtained from the short- and long-form measures across the three samples averaged .80, .89, and .94 for the 10-, 20-, and 30-item tests, respectively. When compared to the reliability of the original form of the test (.91), the correlation coefficients between the 30-item test and the original test show that the 30-item test accounts for the same proportion of true score variance as the total test would be expected to account for on two successive administrations. The proportion of true score variance accounted for by the 10- and 20-item measures were somewhat less than that which would be accounted for on a second administration of the total test. It was noted that the relationships described above were observed although the mean ability estimates resulting from each of the short-form tests differed significantly ($p < .001$) from the results obtained from the total test.

Analysis of the data to determine the extent to which the classification decisions made by using the short forms of the tests were the same as those made by using the original test were in agreement shows that the observed agreement between the short and long forms was very high. The observed coefficients of agreement ranged from .80 on the 10-item test to .91 on the 30-item test. Adjusting the observed coefficients of agreement for chance by using coefficient Kappa resulted in high agreement for the 30-item test, moderate agreement for the 20-item test (.74), and low agreement for the 10-item test (.64).

Analysis of the data regarding the proportion of students

misclassified revealed a direct curvelinear relationship between test length and the total proportion of students misclassified. The data show that the observed proportion of examinees misclassified by using the 10-, 20-, and 30-item tests was 18%, 13%, and 10%, respectively. It is important to note however, that the majority of student misclassified (54% - 80%) were borderline students with regard to the total test (students scoring within the error band of the cut-score used for making pass/fail decisions).

Further analysis revealed that no less than two-thirds of the total proportion of students misclassified had scores within the error band of the short-form tests. For the 30-item test in particular, 98% of the students misclassified had scores within the error band surrounding the cut-score. Accepting the proposition by Wilcox (1980), any classification of these students could well be considered to be correct.

CHAPTER V

IMPLICATIONS AND RECOMMENDATIONS

Introduction

This chapter presents the implications and recommendations, resulting from the findings observed in this investigation, which pertain to reducing the time consumed by competency testing and to alleviating the testing burden on pupils in the state of Maryland. Further, a model describing a set of procedures which can be used to optimize the potential benefits of short-form tests to this end is also presented. Procedures for empirically cross-validating the conclusions reached in the present study conclude this chapter and the report.

Limitations

The reader is cautioned that the implications and recommendations presented in this chapter must be considered with respect to several limitations which were outside the control of the investigator in the present study. The limitations discussed here pertain to context effects of item and person parameters, limitations of empirical research, and the generalizability of the current findings.

The most salient limitation of the present study was the assumption that student responses analyzed for each of the items on each of the short-form tests were devoid of context effects. This assumption is parallel to the assumption of local independence of the items on a test fitting the model which suggests that the items can be reordered without affecting their

parameters and that items can be added to or deleted from the test without affecting the items already on the test (Yen, 1979). However, because the responses analyzed were made as a part of the total test, they might well not represent the actual responses that would have been made if the items were presented as elements of the subsets of items they are analyzed with in this study. Research by Yen (1979) has shown that the context in which an item is presented, in this case the order of the items on the test, does result in some variation of the item parameters. However, she notes that the effects were more prevalent when dealing with a single item than when dealing with a set of items. Further, her research showed that the variation in item parameters resulting from context effects had only marginal effects on the relative size of examinee ability estimates.

A second limitation surrounding the findings presented in this study is the fact that the conclusions reached are based on empirical research which is always tentative. According to Dessart (1983),

> There are no "absolute truths" in empirical research. The findings that may be appropriate for a particular set of students at a particular time and place may not be applicable to another set of students at another time and place. One must always respect the fact that the inference of findings from one sample of students to another is always subject to error. (p. 7)

While the present study attempted to diminish the phenomenon described by Dessart by using very large samples and by replicating the study with three separate samples of 5,000 examinees, it must be acknowledged that the three samples were in fact from the same population and are

therefore the same, excluding sampling error. As such, the findings are subject to the same tentativeness.

Thirdly, the generalizability of the processes and results of this investigation must be considered separately and cautiously. The processes used in the present study are generalizable to any large scale testing program which makes use of an item bank in the development of test forms for competency assessments. However, the results of this investigation might not be generalized to any item bank. It is hypothesized by the researcher in this study that the results obtained are to a great extent contingent upon the quality and integrity of the items contained in the item bank developed by the Maryland State Department of Education's Measurement, Statistics and Evaluation Section. Therefore, the results of this particular study are only generalizable to an item bank of commensurate quality.

Implications

One of the primary questions examined in this investigation was the effects of test length on estimates of functional reading ability levels of ninth grade pupils in the state of Maryland. Multivariate analysis of variance repeated measures procedures and correlational analysis were used to answer this question. The results obtained in the study showed that the mean ability estimates produced by successively shorter forms of the test differed significantly from those produced by the long form of the test. The differences were, however, trivial, with the average difference between the pairs of test forms ranging from a maximum 6/100 of a logit for the 10-item test

down to 2/100 of a logit for the 30-item test. It was concluded that differences of the size described, particularly for the 30-item test, were of no practical or educational significance at all and were statistically significant in part because of the large sample size. The bias observed was, however, systematic, varying with the length of the measures used.

Correlational analysis of the test results provided additional evidence of the relationship between test length and estimates of pupil ability levels. The relationship between the length of the test form and the magnitude of the correlation coefficients was obvious, with coefficients between each of the short-form tests and the long-form test ranging from moderately high to very high, .80, .89,and .94 for the 10-, 20-, and 30-item tests, respectively.

The second question addressed in this study pertained to the relationship between test length and the classification of examinees into pass or fail categories. Specifically, what are the effects of test length on the classification of students into mastery/nonmastery categories of functional reading when the same cut-score is used for making those decisions on each of the test forms investigated? This question was addressed by computing both the observed coefficient of agreement and kappa coefficient of agreement between the classifications made by using the short-form tests and those made by using the long-form test. Results obtained for the use of both procedures showed a distinct relationship between the length of the test and classification of examinees into pass/fail categories, with the larger

indices (greater agreement) being associated with the longer forms of the test.

At least as important as the observed relationship between test length and the classification of examinees into pass/fail categories was the observation that the pass/fail classifications made by using the 30-item test were about the same as those which could be expected from administration of the original test a second time. Further, where the 30-item short-form test and the long-form test differed in classifying pupils into pass/fail categories (10% of total sample tested), approximately 80% of them had scores on the original form of the test which were within the error band of the cut-score on the long-form test and therefore borderline to begin with. Also, virtually all (98%) of the students misclassified had scores within the error band of the cut-score of the 30-item test.

The results obtained in this study show that there is a direct relationship between test length and both estimation of pupil ability levels and the classification of students into pass/fail categories. Jointly, the findings from this investigation suggests tremendous feasibility for using short-form tests for assessing pupil competency levels in reading while simultaneously providing school system decision makers with their information needs, reducing the time consumed by testing, and reducing the testing burden on students.

This is especially true for the 30-item test developed in this investigation, which appears to be capable of producing pupil ability estimates with only trivial differences from those produced by a long-form test more than twice its length. Further, the information needed

by decision makers to determine whether or not pupils possess the minimum skills required for graduation certification is readily available and is as accurate as that provided by the longer 75-item test on a second administration to the same sample. Also, because almost all of the students likely to be wrongfully certified for graduation by using the short-form measure are within the error band of the cut-score of the test, the proportion of students that decision makers are willing to take a risk of misclassifying can readily be adjusted by an administrative decision regarding passing or failing those students with scores within that interval.

In addition to providing one means to help schools optimize the time they have available for instruction, the findings also have some potential relevance for educational cost and student test anxiety. The latter of these constructs, though not addressed beyond this point in the present study, might well be the most important of the potential benefits of reducing test length. Research by Fyans et al. (1980) has shown that test comfort (the opposite of test anxiety) is antecedent to several educationally relevant variables which are related to student achievement. Among those are student attribution of success/failure, expectancy of success, self-concept, continuing motivation, and continuing achievement. Obviously the reduction of the anxiety that many students have about tests should also result in more accurate estimates of pupil true ability levels.

Recommendations

This section of the report presents a model for making optimal use of the findings in this study. The use of this model in conjunction with a short-form test comprised of approximately 30 items can result in the school system accomplishing four major objectives: (1) drastically reducing time consumed by competency testing, (2) reducing pupil test burdens, (3) identifying potential failures at the time that it is most convenient for the school to provide them with remediation, and (4) providing decision-makers with the information they need to insure that pupils who graduate possess the minimum skills required for graduation. The use of short-form tests within the existing test administration procedures will inevitably result in at least a 50% reduction in the time and resources (materials) currently required to administer the MFRT in the state of Maryland. However, the use of the short-form test as proposed in the following set of procedures presents the greatest potential benefits in terms of reducing the time and test burden.

Procedures for Utilizing Short-Form Tests To Optimize Test Administration Time and Minimize Student Testing Burden

1. Utilize a short-form test consisting of approximately 30 items which are targeted at the cut-score used for making the classification decisions.
2. Administer the test in grade seven and classify those students with test scores greater than the cut-score,

plus or minus the error of measurement at the cut-score, as passers. Note 1: One potential advantage of administering the test in grade seven (elementary, middle, or junior high) is that there is greater accessibility of staff trained in reading at these levels than at the senior high school level. Note 2: Certifying students with test scores above the cut-score interval minimizes the chances of certifying for graduation students who do not possess the minimum skills required.

3. Subsequent to remediation in grades 7, 8, and/or 9, students classified as failures of the seventh grade administration of the test should be administered a second test (short-form). On the ninth grade administration of the test, students scoring at or above the cut-score should be classified as passed and students scoring below the cut-score classified as failed. Note: While the issues surrounding students scoring within the error band of the cut-score are still valid, these students have been provided with remediation and have also demonstrated a minimally acceptable level of competency.

4. Procedures utilized after the ninth grade administration of the test could remain as they now exist in terms of remediation and retesting with the exception that the short form should be used in the retesting and that any student scoring within or the error band of the cut-score should be classified as minimally competent for graduation.

Utilizing the procedures described above could quite easily result in time and material savings of 85 - 90% in many school districts depending on the proportion of examinees passing the initial administration of the test.

Validation Proposal

This section of the report presents a proposal for empirically validating the findings of this study. Emphasizing the fact that there is no single type or form of validity, Kerlinger (1973, p. 457) asserts that a test is valid for the scientific or practical purpose of its user. The primary purpose of the required validation research emanating from the present investigation is that of establishing the concurrent validity of the short-form test by using the long-form test as the criterion. Gay (1981) defines concurrent validity as follows:

> Concurrent validity is the degree to which scores on a test are related to the scores on another, already established, test administered at the time or to some other valid criterion available at the same time... Concurrent validity is determined by establishing that relationship or discrimination. (p. 113)

The procedures proposed by Gay (1981) to establish the relationship and discrimination of the two test measures are as follows:

1. Administer the new test to a defined group of individuals.
2. Administer a previously established, valid test (or acquire such scores if already available).
3. Correlate the two sets of scores. (p. 113)

Gay contends that the resulting correlation coefficient is indicative of the concurrent validity of the new test. The extent to which the two measures discriminate between those students who possess the prescribed level of proficiency and those who do not provides additional evidence of the concurrent validity of the measures when the tests are used for this purpose.

Based on the definition and procedures proposed by Gay, the validation of the short-form test is guided by the same questions which generated the original study: Are the ability estimates obtained from the short-form test related to those obtained from the long-form test and more importantly are the pass/fail decisions made by using the short-form test the same as those made by using the long-form test? The primary difference between the present study and the proposed validation study is that the short-form test must be administered as a separate test rather than as a subset of the long-form test.

The specific design proposed for establishing the concurrent validity of the findings in the present study is the one-group repeated trials design described by Kerlinger (1973). In this design one group of examinees is measured two different times. The major advantage of utilizing this particular design is that each examinee is matched with himself. According to Kerlinger (1973), the best possible matching of subjects is to match the subject with himself (p. 362). Kerlinger notes, however, that there are several weakenesses inherent in this design which pose a threat to its internal validity. Those noted by Kerlinger (1973) and described in detail by Campbell and Stanley (1963) include history, maturation, sensitization, and regression effects.

The procedures which are proposed here control for most of them. The specific procedures for conducting the validation study are shown in Illustration 5.

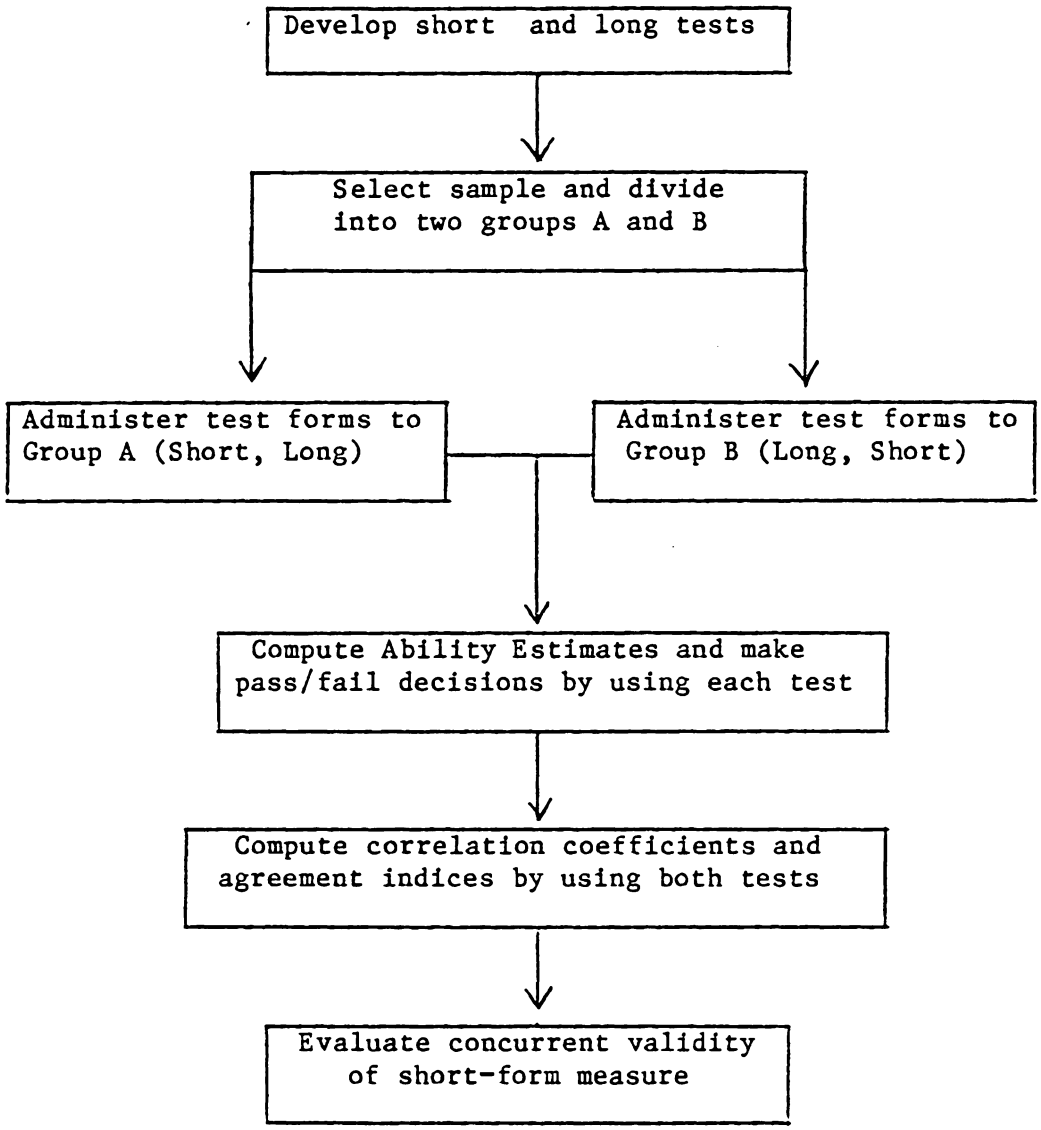


Figure 4. Procedures for assessing the concurrent validity of the short-form test

A detailed discussion of the components of this design includes the following:

1. Develop a long form of the test by using existing test development procedures. Develop a short form of the test containing 30 items from the item bank which are proportionally representative of the five functional reading categories discussed in chapter 1 of this document and which are also targeted to the cut-score used for making pass/fail decisions. The inclusion of six items from each of the categories covered by the test ensures that sufficient diagnostic information is available for decision makers to tailor their remediation activities to the needs of the students failing the test. Consideration should also be given to using the discrimination indices produced by the Bical Program. However, further research needs to be done to determine the potential benefits of doing so.
2. Randomly select a sample of examinees and randomly assign them to either of two groups (A and B).
3. Administer the two test forms (short and long) to the two groups in opposite order. Group A is administered the long-form first and Group B is administered the short-form first. On the second administration Group A is administered the short-form and Group B is administered the long-form. Alternating the order of administering the two test measures should control for or nullify any effects on the test results

by either of the threats to internal validity discussed above (e.g., history, maturation, sensitization, regression effects). Further, by limiting the amount of time between the administrations of the two measures, the effects of history and maturation can be minimized.

4. Compute the ability estimate for each of the examinees by using each of the instruments administered. For example, each student will have two test scores: one from the long-form of the test and one from the short-from of the test.
5. Assign each of the examinees to either a pass or fail category based on the results from each of the measures.
6. Compute correlations between the ability estimates obtained on each of the measures to determine the concurrent validity of the two tests from a relationship perspective. Using the same cut-score for making pass/fail decisions, compute agreement indices between the two measures to reflect the concurrent validity of the short-form measure from a discrimination perspective.

The magnitude of the two indices of concurrent validity (correlation coefficients and agreement index) should provide a basis for determining the efficacy of utilizing the short-form test as an alternative measurement instrument in assessing the competency levels of pupils in the population.

Summary

The purpose of this study was to investigate the relationship between test length on estimates of student functional reading ability and classification of students to pass/fail categories. The two research questions addressed by this study were:

(1) What are the effects of test length on estimates of functional ability of ninth grade examinees in the state of Maryland?

(2) What are the effects of test length on classification of students to mastery/nonmastery states of functional reading when the same cut-score is used on test forms of different length?

The procedures used to conduct the study involved developing several (3) nonoverlapping short-form tests, estimating pupil functional reading levels on each of the short-form tests by using IRT methods applied to their response to selected item on a longer version of the test, and classifying each student as pass or fail on each of the test forms. The item selection strategy used to develop the short-form tests involved selecting those items which were closest to the cut-score of the test (e.g., 80% correct or 1.71 logits).

Multivariate analysis of variance and Pearson's product moment correlation procedures were applied to the scores obtained on each of the short-form tests to determine the extent to which they were equal to or related to the scores obtained on the original test. Results for the analyses performed revealed a direct relationship between the mean

ability estimates produced by the short-forms of the test and also the correlations between the short-form measures and the original test. The mean ability estimates obtained on each of the short-forms of the test were found to differ significantly ($p<.001$) from the mean ability estimates obtained on the original form of the test. However, in each comparison the differences were trivial and the short-form test scores were highly related to the scores obtained on the original form of the test.

A comparison of the pass/fail classifications made by using the short-forms of the test and the original test also revealed a relationship between test length and both the observed coefficient of agreement and coefficient kappa. It was noted, however, that the indices of agreement between the 30-item test and the original test suggest that the 30-item test was about as accurate in assigning students to pass/fail categories as the original test would be if the original test were administered a second time to the same sample.

Subsequent to validation of the results obtained in this study, it is concluded that a short-form test containing approximately 30 items, targeted at the cut-score of the test, would be highly feasible in reducing test administration time and student testing burden. Specifically, using a short-form test of this length, in the manner proposed in this study, could produce highly reliable test results in addition to reducing the test administration time by up to 90% of the time presently used to conduct these activities.

BIBLIOGRAPHY

Amwick, D. J., & Walberg, H. J. (Eds.). Introductory multivariate analysis for educational, psychological, and social research. Berkley: McCuthan Publishing Corporation, 1975.

Benson, J., & Wilson, M. A comparison of three types of test development procedures using latent trait methods. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Berk, R. A. A consumers' guide to criterion-referenced test reliability. Journal of Educational Measurement, 1980, 4, 323-350.

Best, R. A. Research in education (2nd ed.). Englewood Cliffs: Prentice-Hall, Inc., 1970.

Beyer, W. H. (Ed.). Handbook of tables for probability and statistics. Cleveland: The Chemical Rubber Company, 1966.

Boyd, B. H., & Hoover, H. D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980, 17, 179-193.

Brown, F. G. Principles of educational and psychological testing. Hinsdale, Illinois: Dryden Press Inc., 1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on testing. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Canner, J. M., & Lenke, J. M. Some types of items do not fit the Rasch model: Examples and hypotheses. Paper presented at the annual meeting of the National Council of Measurement in Education, Boston, 1980.

Carver, R. P. The case against statistical significance testing. Harvard Educational Review, August, 1978, 48, 378-399.

Cook, L. L., & Eignor, D. R. Score equating and item response theory: Some practical considerations. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1982.

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

Cohen, J. Weighted chi-square: An extension of the kappa method. Educational and Psychological Measurement, 1972, 32, 61-74.

Curry, A. R. Invariance of the rasch model ability estimates over different collections of items. Unpublished doctoral dissertation, University of Georgia, 1977.

Cypress, B. K. The effects of diverse test score distribution characteristics on the ability parameter of the Rasch model. Unpublished doctoral dissertation, The University of Florida, 1972.

DeGorman, R. Coefficients of correlation and concordance for sets of triad judgement. Educational and Psychological Measurement, 1982, 42, 807-814.

Dinero, T. E., & Haertel, E. A computer simulated investigating the applicability of the Rasch model with varying item discriminations. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, 1976. (ERIC Document Reproduction Service No. ED 120 240).

Divigi, D. R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of the National Council of Measurement in Education, 1981.

Doughlass, F. M., Khavaria, K. A., & Farber, P. D., A comparison of classical and latent trait item analysis procedures. Educational and Psychological Measurement, 1979, XXXIX, 337-352.

Douglass, J. B. Applying latent trait theory to a classroom examination system: Model comparison and selection. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, 1980.

Eignor, D. R., & Hambleton, R. K. Effects of test length an advancement score on several criterion-referenced test reliability and validity indices. Laboratory of Psychometric and Evaluation Research Report Number 86, National Institute of Education, Washington, D. C., July, 1979.

Engelhard, G. Jr. An introduction to Rasch measurement and its application to test equating in the comprehensive assessment program. Paper presented at the annual meeting of the Northern Illinios Association for Educational Research, Evaluation, and Development, Bloomingdale, 1980.

Eyman, R. K., Meyer, C. E., & Bendel, R. New methods for test selection and reliability assessment using stepwise multiple regression and jack knifing. Educational and psychological Measurement, 1973, 33, 883-894.

Faggen, J. Decision reliability and classification validity for decision oriented criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Finn, J. D., & Mattson, I. Multivariate analysis in educational research: Applications of the multivariance program. Chicago: National Educational Resources, 1978.

Fitzpatrick, A. R. Validating decisions made with criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Fleiss, J. L., & Cohen, J. The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability. Educational and Psychological Measurement, 1973, 33, 613-619.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard error of kappa and weighted kappa. Psychological Bulletin, 1969, 72, 323-327.

Flexser, A. J. Homogenizing the 2 x 2 contingency table: A method for removing dependencies due to subject and item differences. Psychological Review, 1981, 88, 327-339.

Forster, F. Evaluating links through triangulation. Unpublished paper. Portland, 1976.

Forester, F. Everything you wanted to know about Rasch model (But were afraid to ask). Unpublished paper, Portland, 1977.

Fox, D. J. The research process in education. New York: Holt, Rinehart and Winston, Inc., 1969.

Fyans, L. J. Achievement related motives of educationally disadvantaged students. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, 1980.

Garrison, W. M., & Cogggiola, D. C. Practical procedures for test length reduction and item selection. Department of Health, Education and Welfare, Washington, 1980.

Gay, L. R. Educational Research: Competencies for analysis and application (2nd ed.). Columbus, Ohio: Charles E. Merrill Publishing Company, 1981.

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

Gerbing, D. W., & Hunter, J. E. The metric of the latent variables in Lisrel-IV analysis. Educational and Psychological Measurement, 1982, 423-427.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Gorman, S. A compartive evaluation of two bayesian adative ability estimation procedures with a conventional test strategy. Unpublished doctoral dissertation, The Catholic University of America, 1980.

Goldstein, H. Are there any sensible uses of latent trait models in educational assessments? How else could one measure change over time. Paper presented at the annual meeting of American Educational Research Association and the National Council of Measurement in Education, New York, 1982.

Goldstein, H. Consequences of the Rasch model for educational assessment. British Educational Research Journal, 1979, 5, 211-220.

Green, M. S. The invariance of parameter estimates in three latent trait models. Unpublished doctoral dissertation, 1981.

Guilford, J. P., & Frutcher, B. Fundamental statistics in psychology and education (5th ed.). New York: Mcgraw-Hill Book Company, 1973.

Gustasson, J. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, 1979, 16, 153-193.

Gustafsson, J. Testing and obtaining fit of data to the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979. (ERIC Document Reproduction Service No. ED 171 756).

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

Hambleton, R. K. Developments in latent trait theory: Models, technical issues, and application. Review of Educational Research, 1978, 48, 467-510.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Haladyna, T., & Roid, G. A comparison of two item selection procedures for building criterion-referenced tests. National Institute of Education, Washington, 1981.

Horn, J. L. Integration of concepts of reliability and standard error of measurement. Educational and Psychological Measurement, 1971, 31, 57-74.

Hull, H. C., & Nic, N. H. (Eds.). SPSS update 7-9: New procedures and facilities for releases 7-9. New York: McGraw-Hill Book Company, 1981.

Kane, M. T., & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, 1980, 4, 105-126.

Kerlinger, F. N. Foundations of behavioral research (2nd ed). New York: Holt, Rinehart and Winston, Inc. 1973.

Koch, W. B., & Reckase, M. D. Problems in application of latent trait models to tailored testing. Paper presented at the annual meetingg of the National Council of Measurement in Education, San Francisco, 1979. (ERIC Document Reproduction Service No. ED 177 196).

Kolen, M. J. Comparison of traditioanl and item response theory methods for equating tests. Journal of Educational Measurement, 1980, 18, 1-11.

Livingston, S. A., & Wingersky, M. S. Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement, 1979, 16, 247-260.

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

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores.Reading, Massachusettes: Addison-Wesley Publishing Company, 1974.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Hougghton-Mifflin Company, 1953.

Marshall, J. L., & Serlin, R. C. Charateristics of four mastery test reliability indices: Influence of distribution shape and cutting score. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Mckinley, R. L., & Reckase, M. D. A successful application of latent trait theory to tailored achievement testing. (Office of Naval Research). Arlington: Virginia Personnel and Training Research Programs Office, 1980. (ERIC Document Reproductions Service No. ED 190 651).

Mitchell, V. P., & Smith, R. M. Determining cut-off scores using Rasch ability estimates. Paper presented at the annual meetingg of the National Council of Measurement in Education, Boston, 1980.

Norusis, M. J. SPSS introductory guide: Basic statistics and operations. New York: McGraw-Hill Book Company, 1982.

Phillips, G. W. Applications of item response theory to criterion-referenced reliability. Paper presented at the annual meeting of the Eastern Educational Research Association, Baltimore, 1983.

Phillips, G. W. Applications of the Rasch model to criterion-referenced testing: Estimation of domain scores reliability, test length, and item selection. Paper presented at the annual meeting of the National Council of Measure- ment in Education, Montreal, 1983.

Phillips, G. W. Item response theory applications in Maryland: The integration of item response theory and criterion-referenced testing. Paper presented at the annual meeting of the National Council of Measurement in Education, Montreal, 1983.

Phillips, G. W. Rasch analysis using SPSS. Paper presented at the annual meeting of the National Council of Measurement in Education, Montreal, 1983.

Pohlman, J. T. Controllong the type I error rate in stepwise regression analysis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Popham, W. J. (Ed.). Criterion-referenced measurement: An introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Nielson & Lydiche, 1960.

Reckase, M. D. Tailored testing, measurement problems and latent trait theory. Paper presented at the annual meeting of the National Council of Measurement in Education, Los Angeles, 1981.

Reckase, M. D. An application of the Rasch simple logistic model to tailored testing. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago , 1974. (ERIC Document Reproduction Service No. ED 092 585)

Rentz, C. C. An investigation of the invariance properties of the Rasch model parameter estimates (Doctoral dissertation, University of Georgia, 1975). Dissertation Abstracts International, 1975, 36/08B, 549636. (University Microfilms No. AAD76-02258)

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work? A discussion for practitioners. (ERIC/TM Report 67). Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1978.

Ryan, J. P. The rational for the Rasch model. Paper presented at the annual meeting of the Southeastern Psychological Association, Hollywood, Florida, May, 1977.

Ryan, J. P. Testing the appropriateness of the one-parameter latent model to the analysis of basic skills assessment data. Paper presented at the annual meeting of the American Educational Research Association, New York, March, 1982.

Ryan, J. P. Equating new test forms on existing tests. Paper presented at the annual meeting of the National Council of Measurement in Education, Los Angeles, April, 1981.

Ryan, J. P., Garcia-Quintanta, R., & Hamm, D. W. Testing the fit of subjects to a latent model. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April, 1980.

Saunders, J. C., & Ryan, J. P. Test equating and analysis of fit with the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, New York, March, 1982.

Slawski, E. J., & Buaer, E. A. Reducing testing time while preserving test information: A ten item fourth grade MEAP reading test. Paper presented at the annual meeting of the Michigan Educational Research Association, Detroit, 1978.

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

Slinde, J. A., & Linn, R. L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 16, 159-165.

Subkoviak, J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic approach. Journal of Educational Measurement, 1974, 2, 111-116.

Tate, M. W. Statistics in education. New York: The Macmillan Company, 1955.

Thurston, L. L. Attitudes can be measured. American Journal of Sociology, 1928, 33, 529-554.

Tyler, R. W., & Wolf, R. M. (Eds.). Crucial issues in testing. Berkley: McCutchan Publishing Corporation, 1974.

Warm, T. A. A primer of item response theory. U. S. Coast Guard Institute, Department of Transportation, Technical Report No. 941078, October, 1978.

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

Wilcox, R. R. Determining test length to control for false-positive and false-negative error rates on criterion-referenced tests. Center for the Study of Evaluation, California University, Los Angeles, 1980.

Williams, P. L., & Moore, J. R. (Eds.) Criterion-referenced testing for the social studies. Washington: National Council for the Social Studies, Bulletin 64, 1980.

Winer, B. J. Statistical principal in experimental design (2nd ed.). New York: McGraw-Hill Book Company, 1971.

Wood, D. A. Test construction: Development and interpretation of achievement tests. Columbus, Ohio: Charles E. Merrill Books, Inc., 1961.

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

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 219-225.

Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model. Statistical Laboratory, Department of Education, The University of Chicago, 1977.

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

Wright, B. D., & Stone, M. H. Best test design. Palo Alto: Scientific Press, 1978.

APPENDIX A

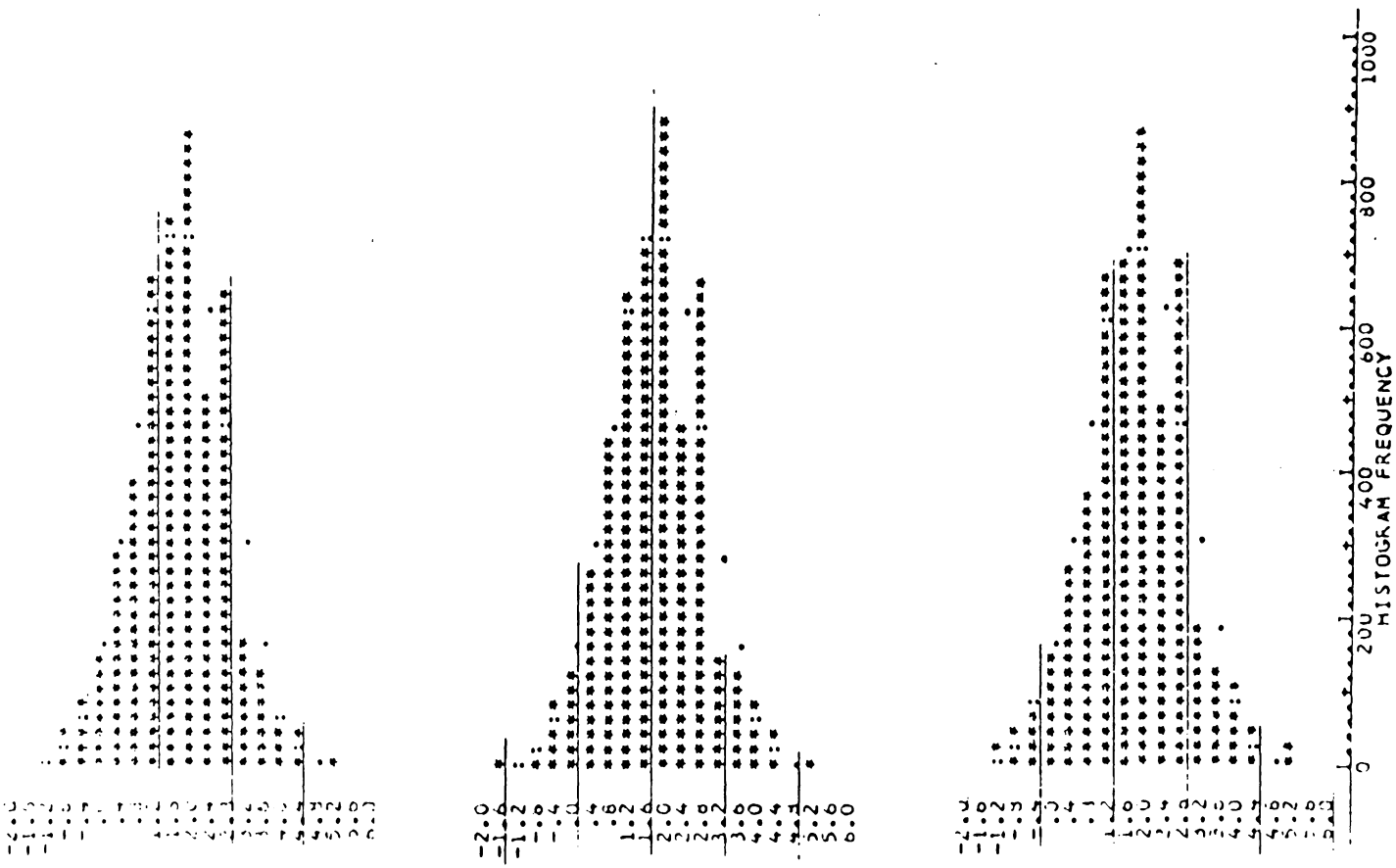
Histogram of Person Ability Estimates for each Sample

on the Original Form of the MFRT

SAMPLE A

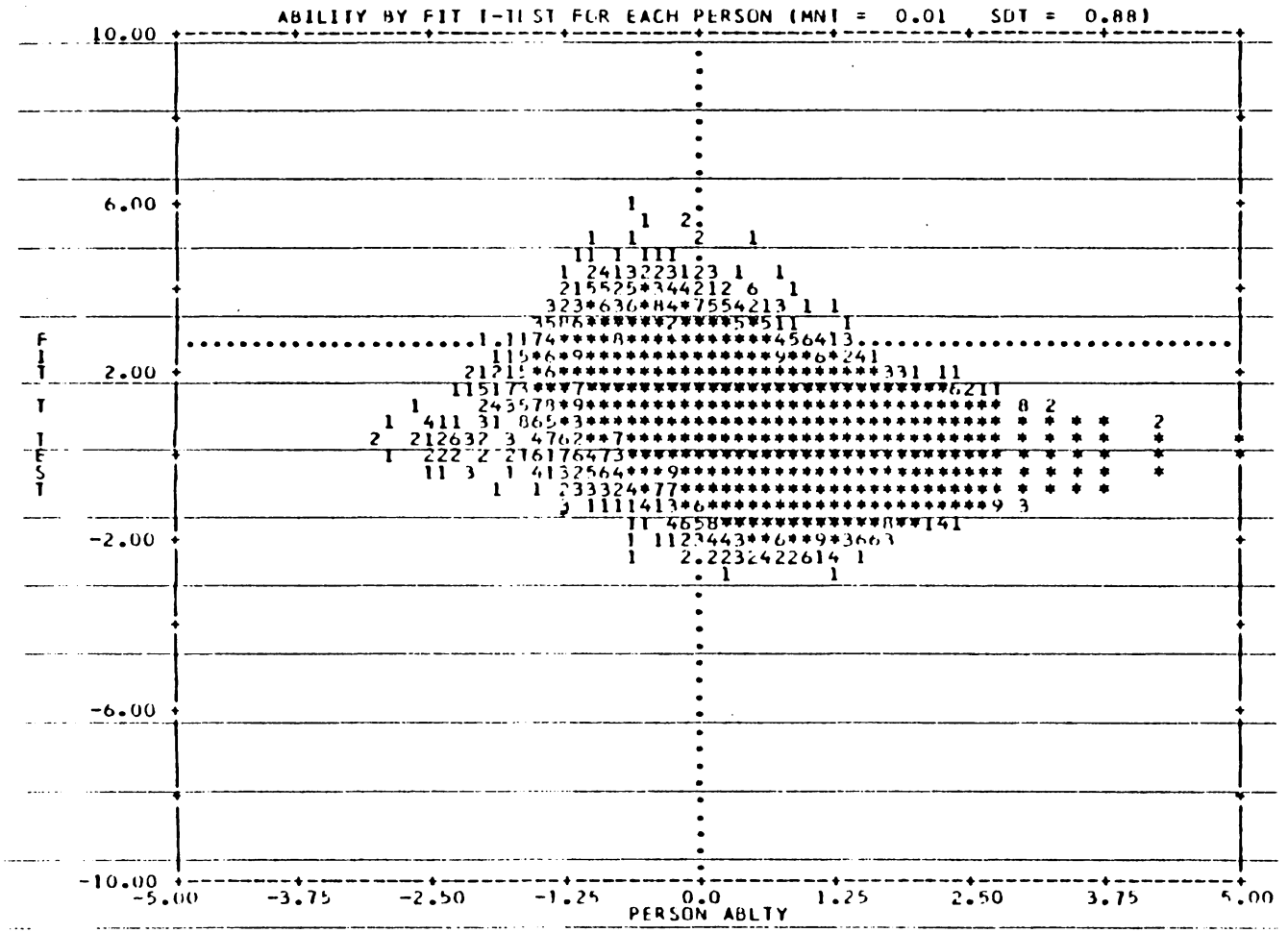
SAMPLE B

SAMPLE C



APPENDIX B

Fit Statistic and Distribution

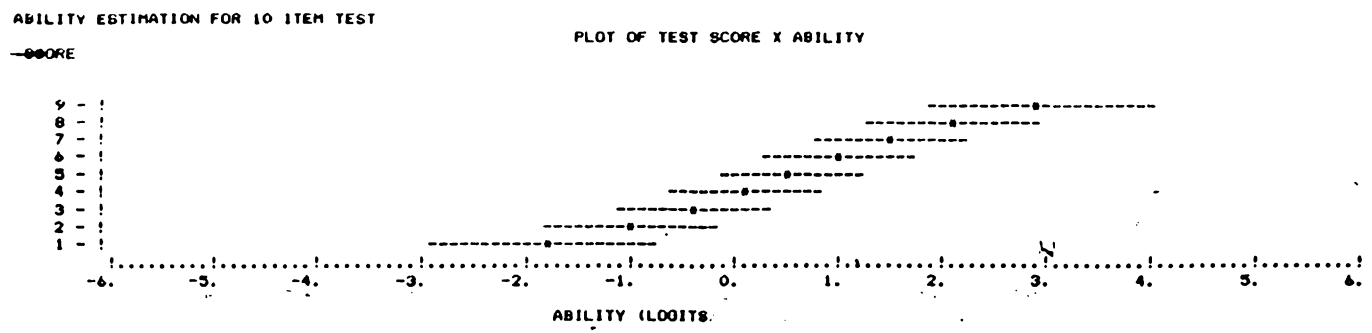


THE 290 PERSONS WITH FIT ABOVE 3.00 WILL BE OMITTED FROM RECALIBRATION

75 ITEMS CALIBRATED ON 36418 PERSONS

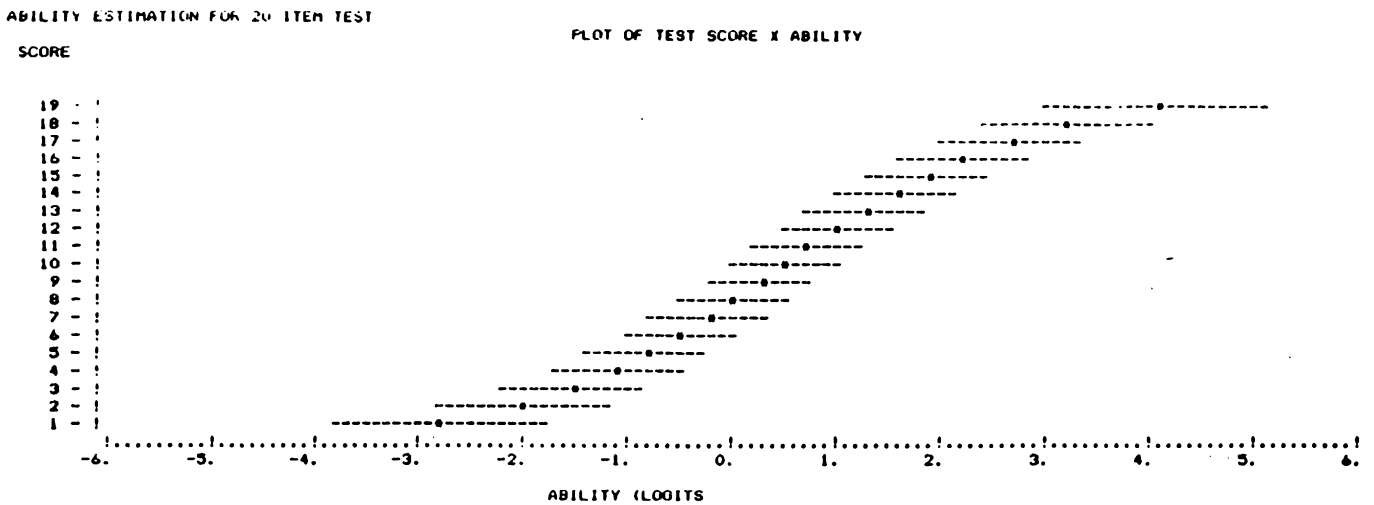
APPENDIX C

Test Characteristic Curve for the 10 Item Test



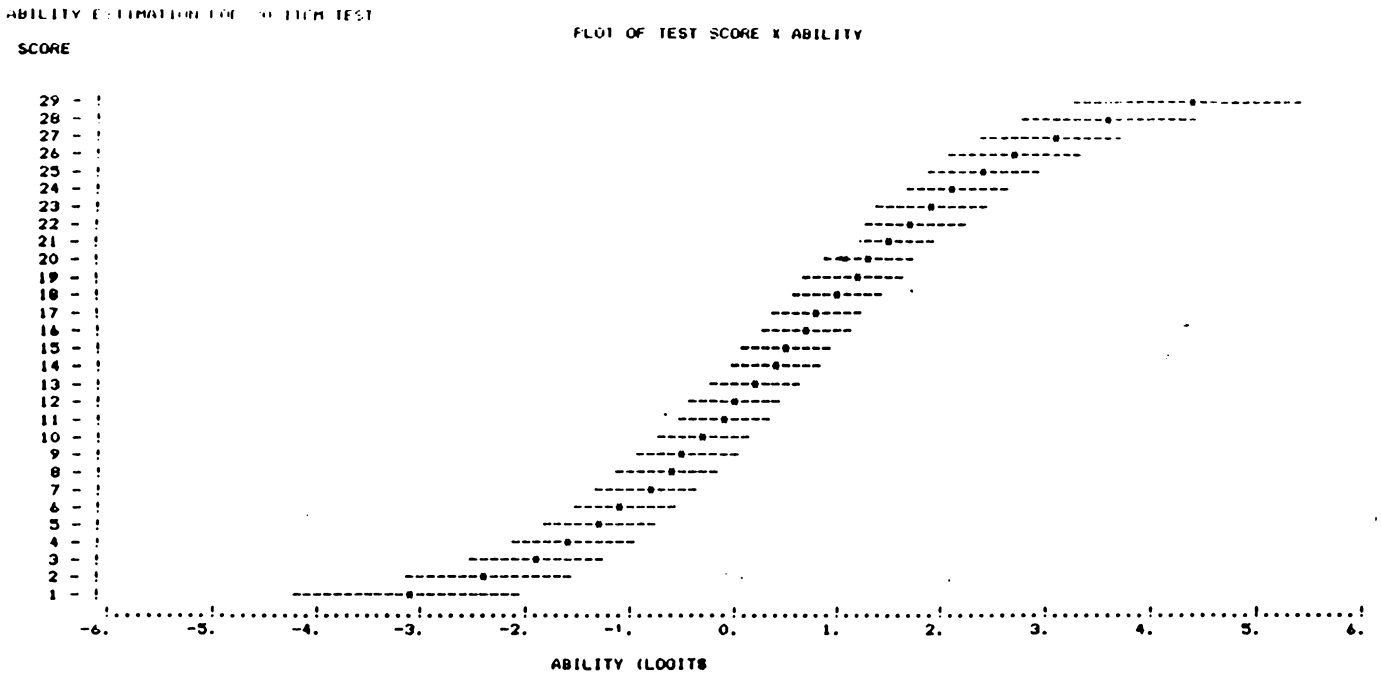
APPENDIX D

Test Characteristic Curve for the 20 Item Test



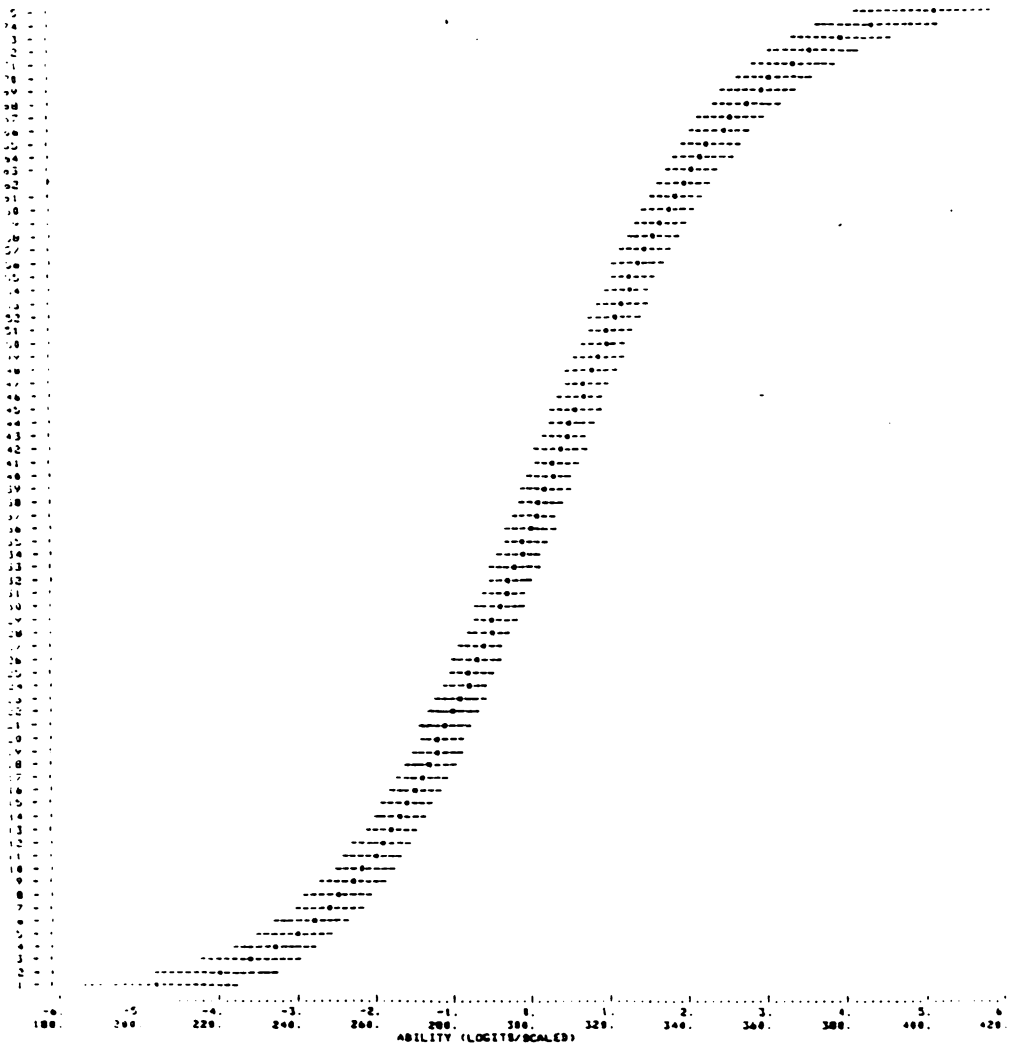
APPENDIX E

Test Characteristic Curve for the 30 Item Test



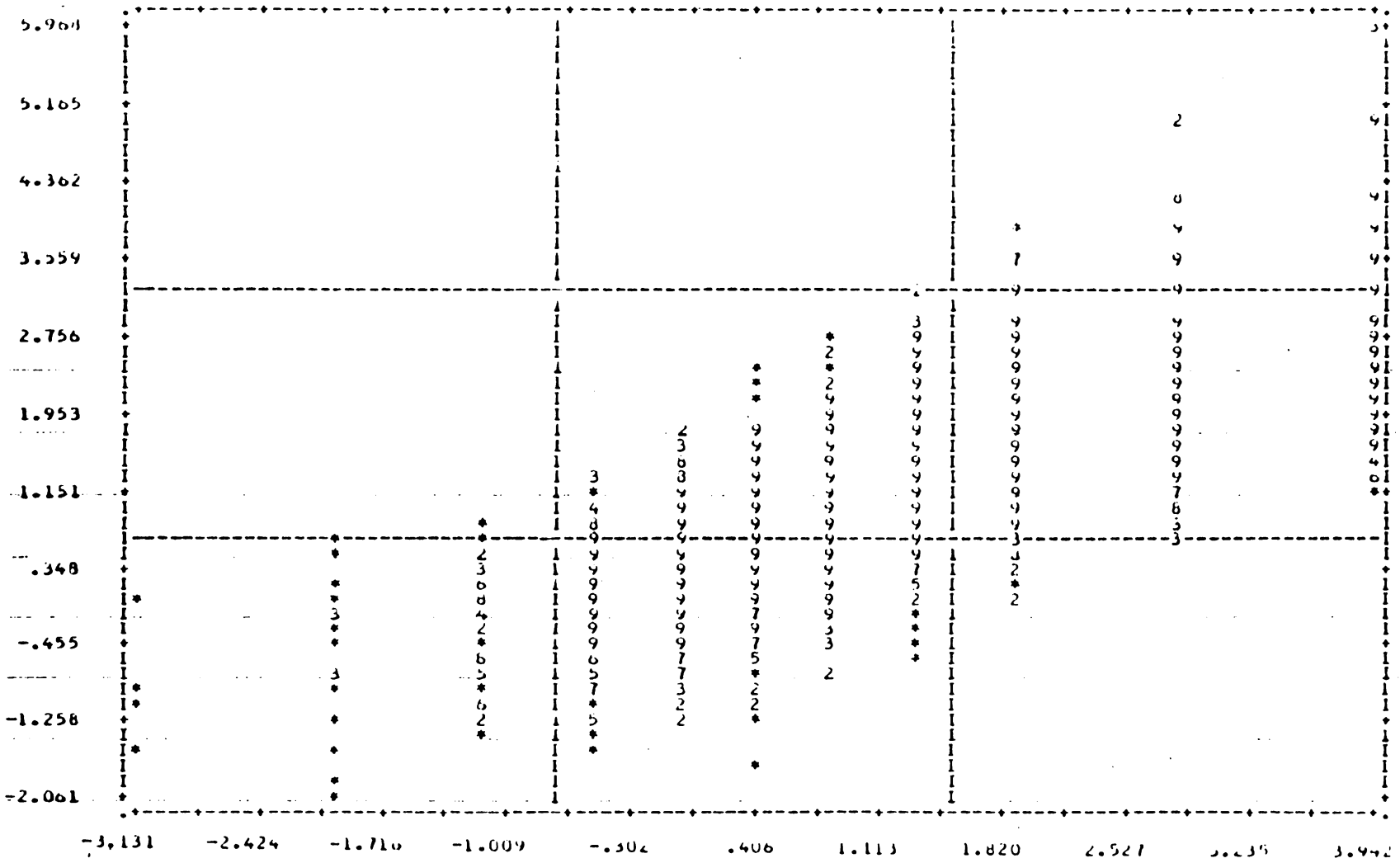
APPENDIX F

Test Characteristic Curve for the 75 Item Test



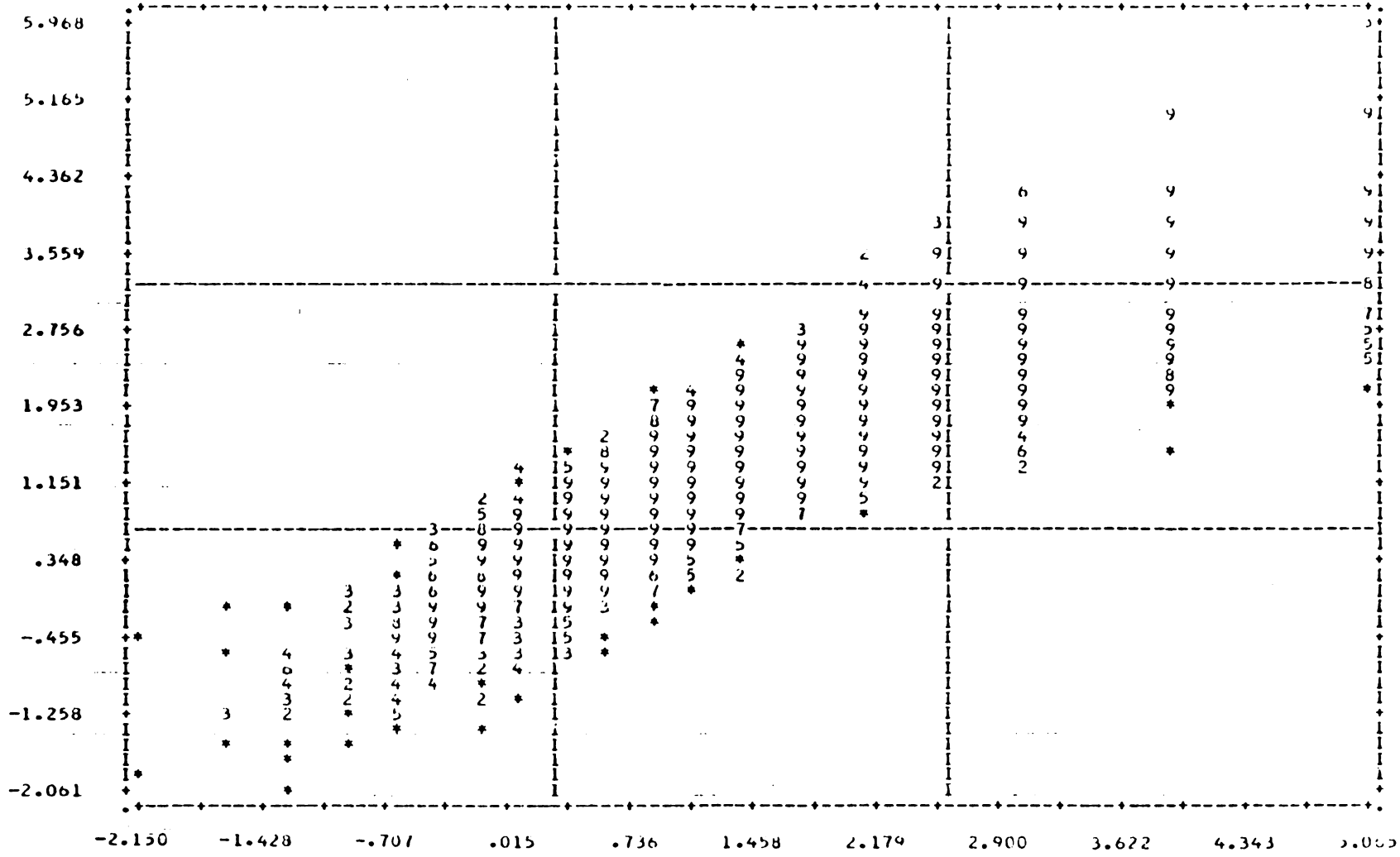
APPENDIX G

Scattergram 10 Item Test x 75 Item Test Sample A



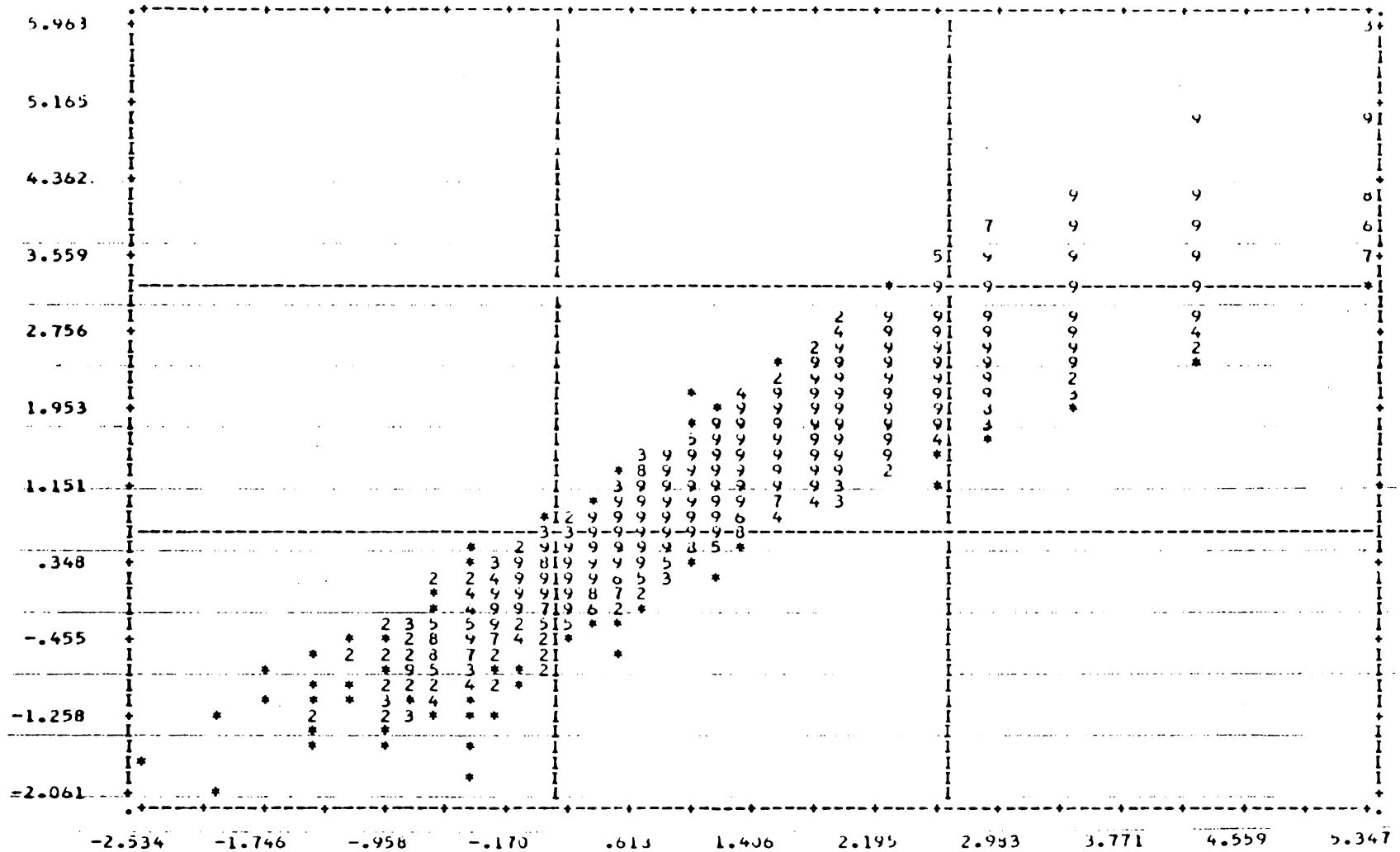
APPENDIX H

Scattergram 20 Item Test x 75 Item Test Sample A



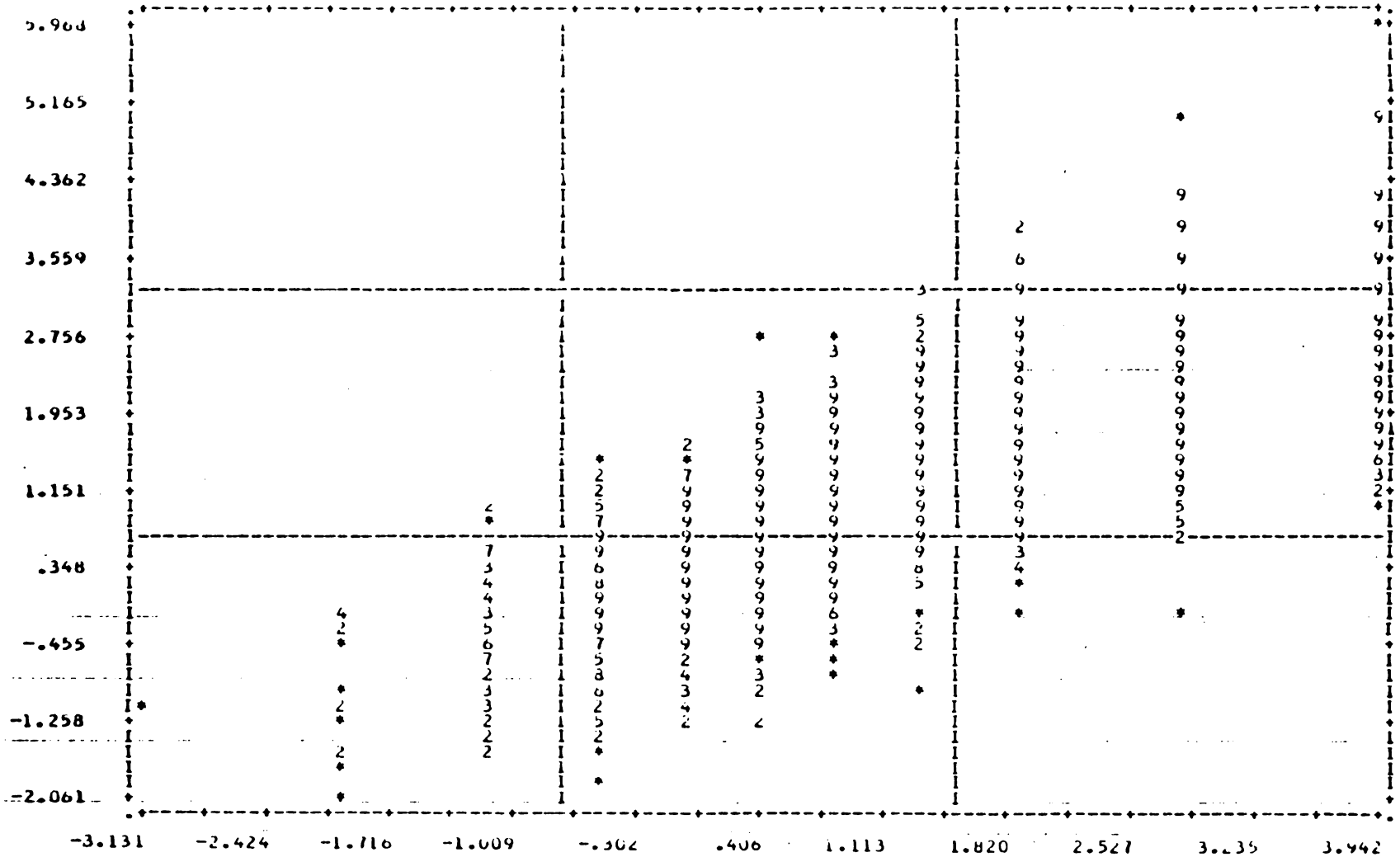
APPENDIX I

Scattergram 30 Item Test x 75 Item Test Sample A



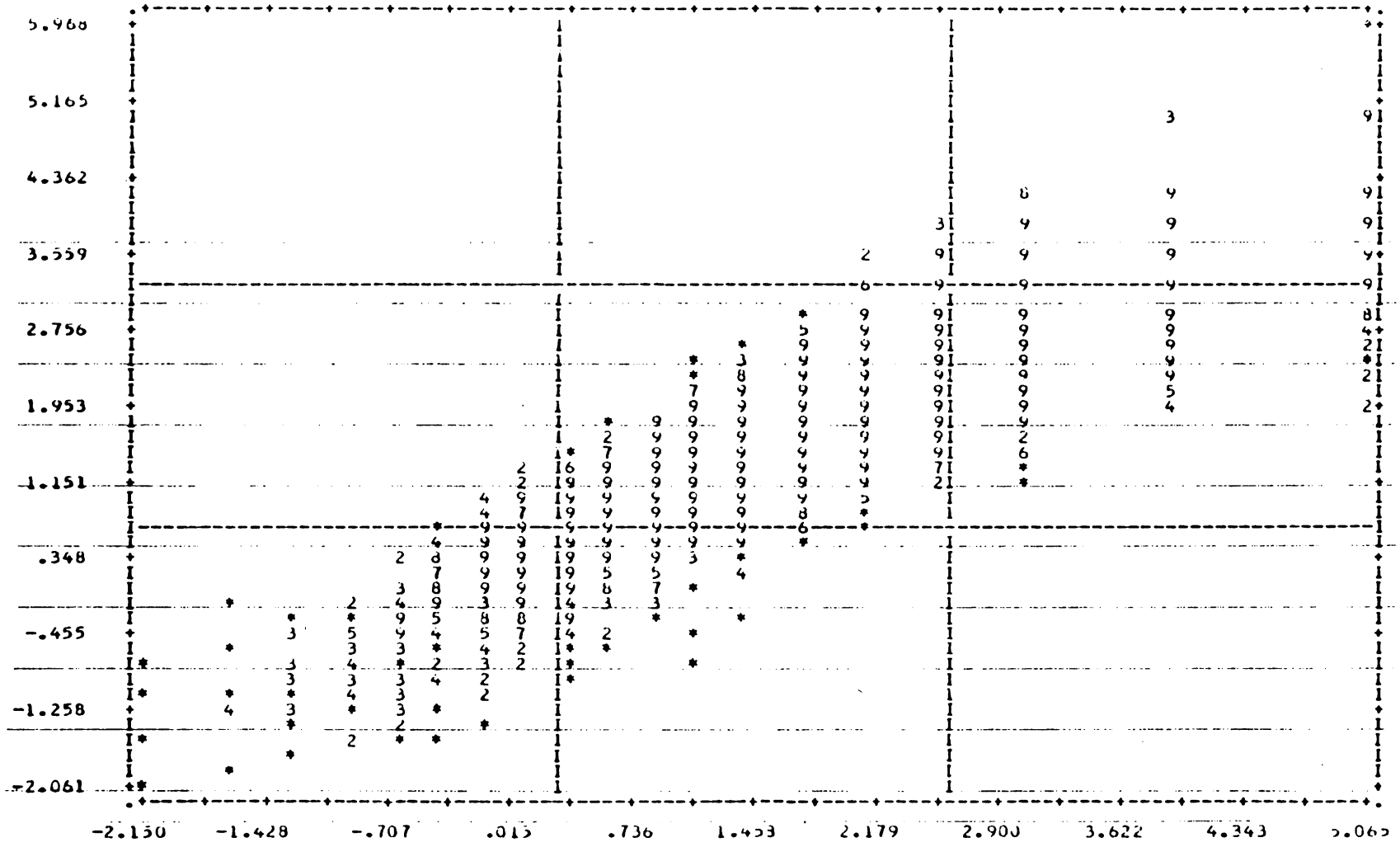
APPENDIX J

Scattergram 10 Item Test x 75 Item Test Sample B



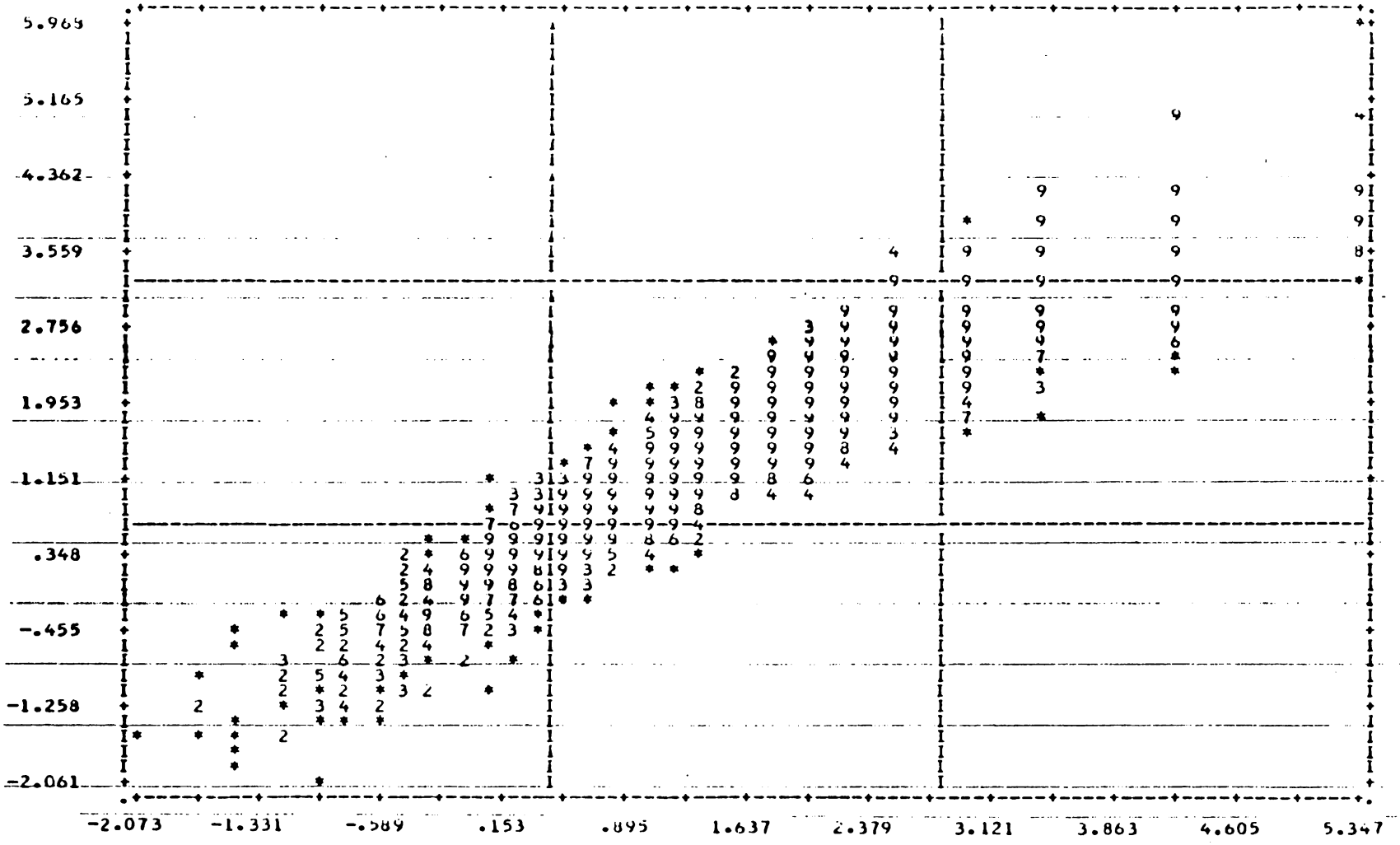
APPENDIX K

Scattergram 20 Item Test x 75 Item Test Sample B



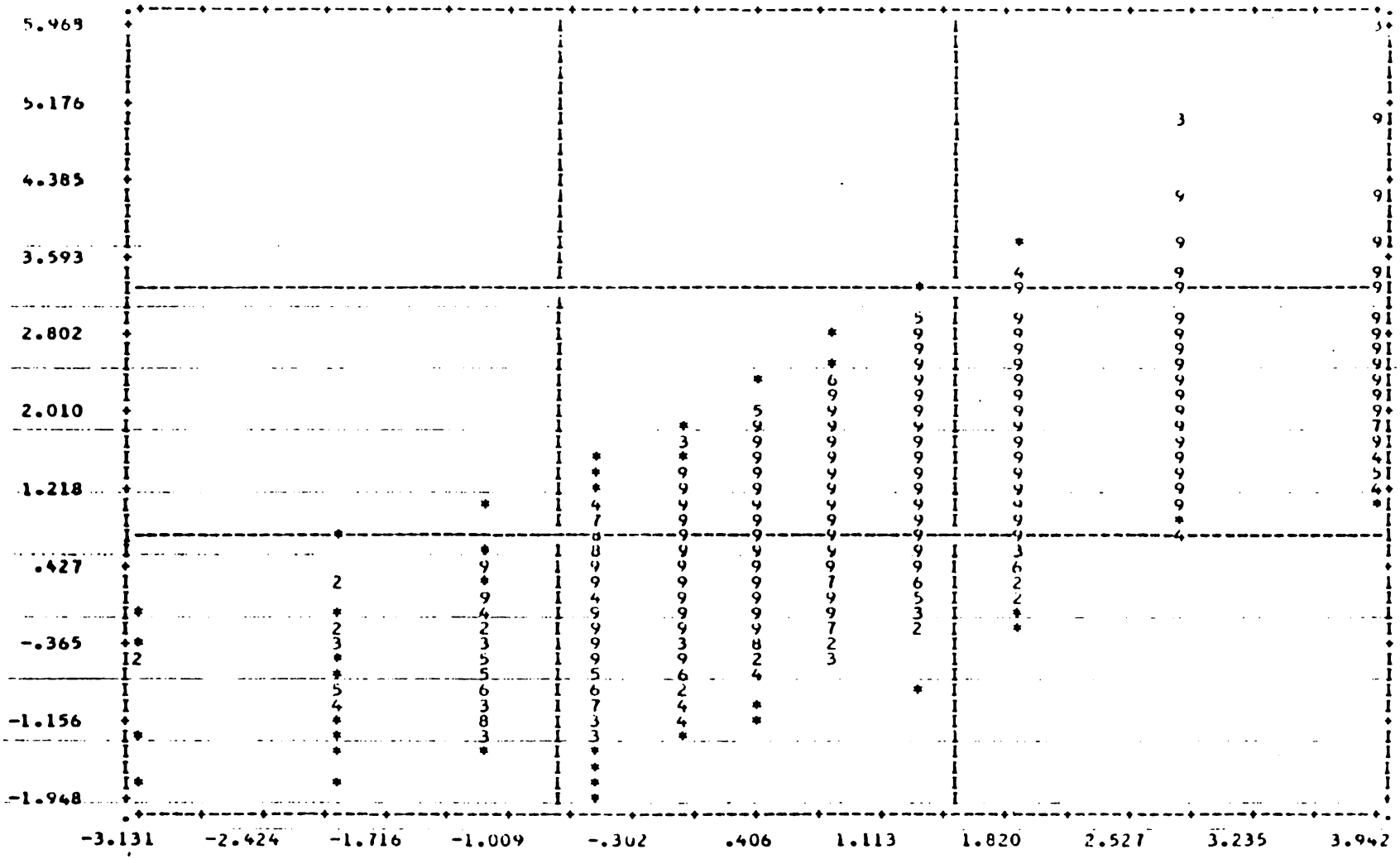
APPENDIX L

Scattergram 30 Item Test x 75 Item Test Sample B



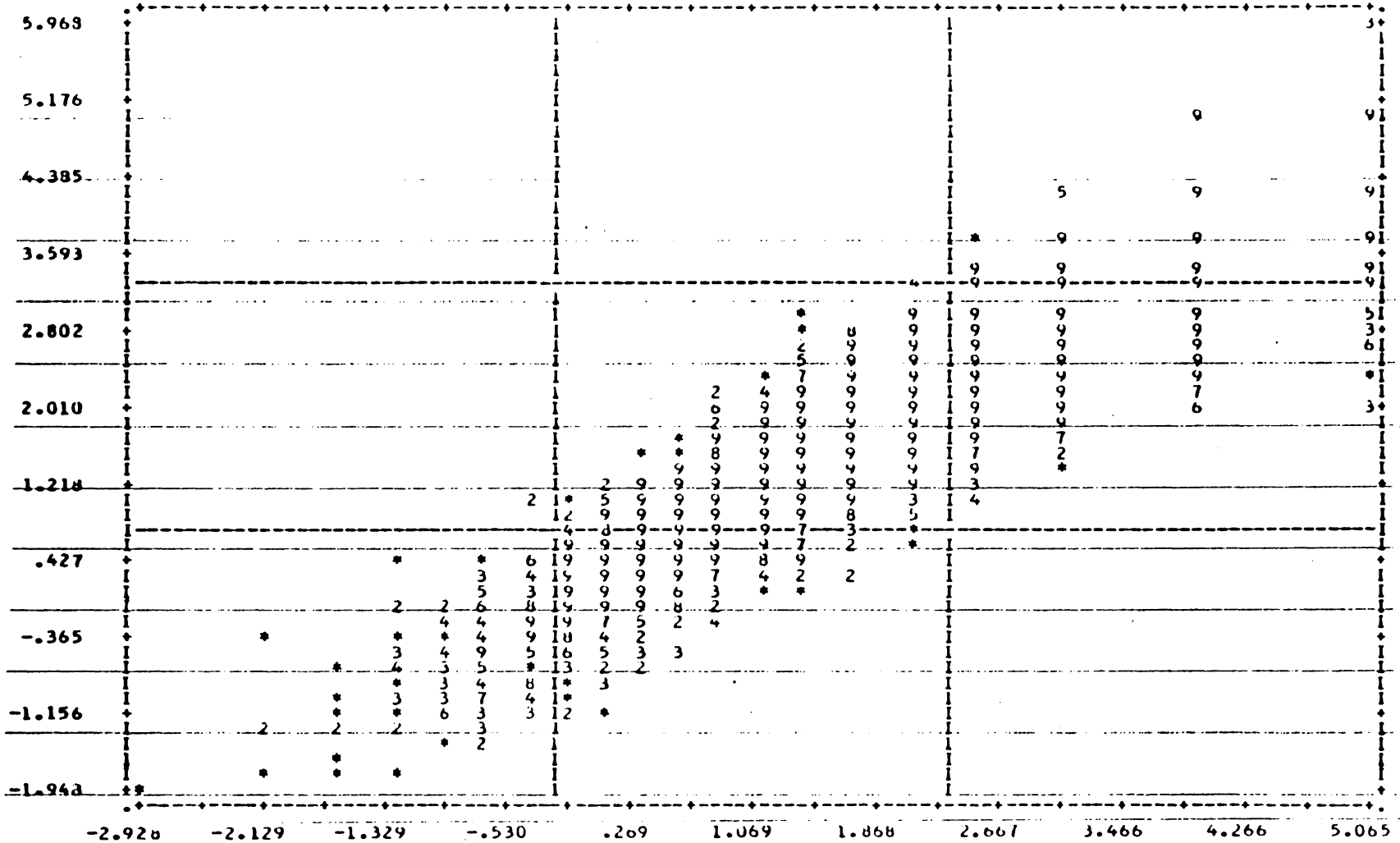
APPENDIX M

Scattergram 10 Item Test x 75 Item Test Sample C



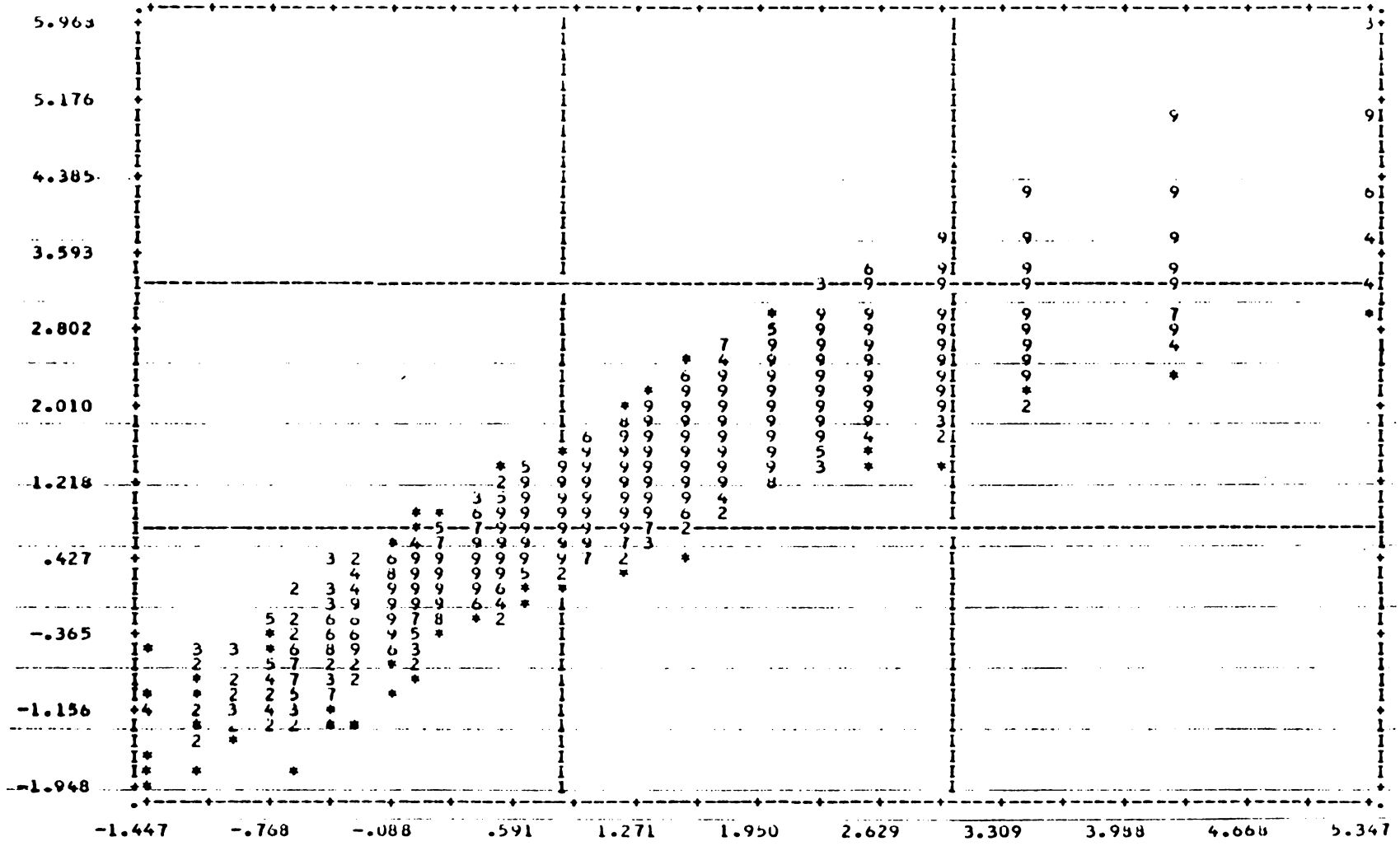
APPENDIX N

Scattergram 20 Item Test x 75 Item Test Sample C



APPENDIX O

Scattergram 30 Item Test x 75 Item Test Sample C



APPENDIX P

Proportion of Students Misclassified with Scores Within the Error Band of The Cut-Score on Long and Short Forms of the MFRT

	Total N Misclass.	Long-Form # in Error Band	Long-Form % in Error Band	Short-Form # in Error Band	Short-Form % in Error Band
10 Item Test					
Sample A	888	474	53%	641	72%
Sample B	903	485	54	661	73
Sample C	856	467	55	614	72
20 Item Test					
Sample A	645	424	66%	432	67%
Sample B	700	463	66	477	68
Sample C	617	389	63	424	69
30 Item Test					
Sample A	496	395	80%	487	98%
Sample B	456	366	80	441	97
Sample C	474	381	80	467	99

The three page vita has been removed from the scanned document. Page 1 of 3

The three page vita has been removed from the scanned document. Page 2 of 3

The three page vita has been removed from the scanned document. Page 3 of 3

AN INVESTIGATION OF THE EFFECTS OF TEST LENGTH ON SHORT-FORM BASIC SKILLS COMPETENCY TESTS DEVELOPED BY USING THE ONE-PARAMETER ITEM RESPONSE MODEL

By

Leroy J. Tompkins

(ABSTRACT)

The purpose of this study was to investigate the effects of test length on the estimation of functional reading ability levels and mastery/nonmastery classifications of ninth grade students in the state of Maryland with short-form tests developed by using the one-parameter item response model.

Using the item responses of approximately 36,000 students to 75 items on a functional reading test, item parameters were estimated with the one-parameter item response model. Three nonoverlapping short-form tests of 10, 20, and 30 items were developed using items targeted at the cut-score of the test. The study investigated the extent to which estimates of pupil functional reading ability levels and mastery/nonmastery classifications obtained from three short-form tests were the same as or related to those obtained on the original 75-item test. Three nonoverlapping samples of 5,000 students each were used to make the comparisons. The extent to which estimates of pupil performance on the short-form measures were the same as that on the original tests was analyzed using a multi-variate analysis of variance design. The results showed that the ability estimates obtained on each of the short-form tests differed significantly ($p<.000$) from that obtained on the original test. The differences

were, however, trivial, measuring less than .06 of the standard deviation of the shorter test.

Pearson's product moment correlation coefficients obtained in this study were, on average, .80, .89, and .94 between the original test (75 items) and the 10-, 20-, and 30-item tests, respectively.

Analysis of the mastery/nonmastery classifications resulted in observed indices of agreement between the short- and long-form tests ranging from .82 for the 10-item test to .91 for the 30-item test. Kappa indices of agreement between the the short- and long-form measures ranged from .64 for the 10-item test to .81 for the 30-item test.

The study concluded that there is a relationship between test length and estimation of pupil functional reading ability and student mastery/nonmastery classifications. It is proposed, however, that a substantial reduction in testing time and student testing burden can be realized by using short-form tests developed and administered in a manner described in the study.