1-1-1993

The effects of dimensionality and item selection methods on the validity of criterion-referenced test scores and decisions.

Mohamed Awil Dirir
University of Massachusetts Amherst

THE EFFECTS OF DIMENSIONALITY AND ITEM SELECTION METHODS ON THE VALIDITY OF CRITERION-REFERENCED TEST SCORES AND DECISIONS

A Dissertation Presented

by

MOHAMED AWIL DIRIR

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

DOCTOR OF EDUCATION

May 1993

School of Education

THE EFFECTS OF DIMENSIONALITY AND ITEM SELECTION METHODS
ON THE VALIDITY OF CRITERION-REFERENCED
TEST SCORES AND DECISIONS

A Dissertation Presented

by

MOHAMED AWIL DIRIR

Approved as to style and content by:

H Swaminathan

Hariharan Swaminathan, Chair

Ronald K. Hambleton

Ronald K. Hambleton, Member

Gene Fisher

Gene Fisher, Member

John C Carey

Bailey Jackson, Dean
School of Education

ACKNOWLEDGMENT

I am indebted to many individuals for offering invaluable assistance and support throughout my graduate training. First, I want to thank my wife, Koos, and our three children, for their support and endurance during my long absence. In fact, they forfeited their right to have me at home as husband and father. My other family, my mother, my father, sister and brothers, lost my help during the hardships they suffered throughout the years I was studying in the United States. I would like to thank them for their moral support and patience with me; I hope I will be able to help them in the future.

My dissertation committee deserves many thanks for their assistance, patience, and support. I would like to thank my thesis advisor, Professor Hariharan Swaminathan, who directed and supported me throughout my graduate study, and who offered valuable training, advice, and guidance. I would like to thank Professor Ronald K. Hambleton who encouraged me in researching the topic of my thesis, and helped me throughout the process of this dissertation. I would also like to thank him for involving me in several research projects which provided invaluable experience. I would like to thank Professor Gene Fisher, who offered many long hours of his valuable time, and provided precious consultations, encouragement, and guidance. Finally, many thanks to my friends and classmates in the Research and

Evaluation Methods Program, for their academic and social support during my enrollment in the Program.

ABSTRACT

THE EFFECTS OF DIMENSIONALITY AND ITEM SELECTION METHODS ON THE VALIDITY OF CRITERION-REFERENCED SCORES AND DECISIONS

MAY 1993

MOHAMED AWIL DIRIR, B.Sc., SOMALI NATIONAL UNIVERSITY

M.Ed., UNIVERSITY OF MASSACHUSETTS

Ed.D., UNIVERSITY OF MASSACHUSETTS

Directed by: Professor Hariharan Swaminathan

Many of the measurement models currently used in testing require that the items that make up the test span a unidimensional space. The assumption of unidimensionality is difficult to satisfy in practice since item pools are arguably multidimensional. Among the causes of test multidimensionality are the presence of minor dimensions (such as test motivation, speed of performance and reading ability) beyond the dominant ability the test is supposed to measure. The consequences of violating the assumption of unidimensionality may be serious. Different item selection procedures when used for constructing tests will hav eunkown and differential effects on the reliability and validity of tests.

The purposes of this research were (1) to review research on test dimensionality, (2) to investigate the impact of test dimensionality on the ability estimation and the decision accuracy of criterion-referenced tests, and (3) to examine the effects of interaction of item selection methods with test dimensionality and content categories on ability estimation and decision accuracy of criterion-referenced tests.

The empirical research consisted of two parts: in Part A, three item pools with different dimensionality structures were generated for two different tests. Four item selection methods were used to construct tests from each item pool, and the ability estimates and the decision accuracies of the 12 tests were compared in each test. In Part B, real data were used as an item bank, and four item selection methods were used to construct short tests from the item bank. The measurement precision and the decision accuracies of the resulted tests were compared.

It was found that the strength of minor dimensions affect the precision of the ability estimation and decision accuracy of mastery tests, and that optimal item selection methods perform better than other item selection methods, especially when test data are not unidimensional. The differences in measurement precision and decision accuracy among data with different degrees of multidimensionality and among the different item selection methods were statistically and practically significant.

An important implication of the study results for the practitioners are that the presence of minor dimensions in a test may lead to the misclassification of examinees, and hence limit the usefulness of the test.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Criterion-Referenced Tests

Criterion-referenced tests are used to assess examinee performance on prespecified and well defined content domains, or tasks. These tests are extensively used by school districts, professional organizations, and state departments of education (Hambleton & Rogers, 1989). Hambleton and Rogers (1989) stated that 48 out of the 50 states in the U.S., and more than 900 licensing agencies use criterion-referenced tests. The tests are mainly used for two purposes in schools: to describe students, and to assign students to mastery levels. Subsumed by these two purposes are such goals as evaluating training programs and instruction, diagnosing student weaknesses and progress, and assessing student mastery levels of content domains (Hambleton & Jurgensen, 1990).

The essential components in criterion-referenced test construction, are the definition of the objectives measured by the test, the match between the items and the objectives they measure, and a standard or cut-off score to sort examinees into mastery states. Criterion-referenced tests usually measure more than one objective, and test items are arranged in distinct subsets that reflect the objectives the test measures. In test score reporting, it is desirable to report the scores by objective. In the test construction process, item statistics do not play as important a role as they do in norm-referenced tests. Instead, they are used to detect flawed items which might be revised in the future. Intended uses of the test, objectives to be measured, items for each objective, and item-objective correspondence are central in criterion-referenced test development.

Reliability and validity of test scores are critical in criterion-referenced tests as they are with any test, and are addressed in ways different from the methods used in norm-referenced tests. The reliability indices of norm-referenced tests are not applicable to criterion-referenced tests partly because criterion-referenced test scores are more homogenous than norm-referenced test scores, and mainly because these indices do not provide important information about the scores, namely; the precision of domain score estimates, and the decision accuracy of scores (Hambleton & Jurgensen, 1990). The acquisition of these two pieces of information, which are the basis of test score reliability and validity, is important to criterion-referenced test score uses.

Since many criterion-referenced tests are used to classify examinees into masters and nonmasters, test reliability is often indexed in terms of test-retest or parallel forms decision consistency (see, for example, Hambleton & Novick, 1973; Swaminathan, Hambleton, & Algina, 1974; Huynh, 1976; Subkoviak, 1976;). The decision consistency, which was first introduced by Hambleton and

Novick (1973), is the proportion of examinees consistently classified as masters or nonmasters in repeated measurements of one form or parallel forms of the same tests. Swaminathan, Hambleton, and Algina (1974) recommended the use of coefficient kappa in which the decision consistency is adjusted to account for chance agreement. Decision consistency estimates based on test-retest or parallel forms need two test administrations, a design that is difficult to implement in practice. Decision consistency indices based on single administrations were separately introduced by Huynh (1976) and Subkoviak (1976). Many factors affect the decision consistency of criterion-referenced tests. Among them are the selection of a cut-off score, and the composition of the examinee population.

The validity of criterion-referenced test scores can be partially addressed in terms of content validity, a process in which content specialists assess the item-objective congruence. Lately, it has been legitimately argued that content validity is not enough to represent the validity of test scores (Messick, 1975; Linn, 1980; Hambleton, 1984). Construct validity, criterion-related validity, and content validity are all important in criterion-referenced tests. The investigation and assurance of each type of validity is equally important.

Hambleton (1984) discussed several procedures in which these test score validity investigations could be undertaken. In content validity investigations, the item-objective match is assessed by content specialists. In construct validity, what the test is and is not measuring are examined. In criterion-related validity, the accuracy of the test scores in domain score estimation and in assigning examinees into mastery levels are assessed. It should be stressed that, in all types of validity investigations, the intended use of the test scores is an important factor.

IRT Uses in Criterion-Referenced Testing Practice

Item response theory (IRT) has applications that are useful in addressing many of the practical problems in criterion-referenced tests (for a review, see Hambleton & Rogers, 1989). Test score reporting, test length determination, and item selection are among the areas in which IRT has been found to be valuable in criterion-referenced tests.

The benefits from IRT are realized when its assumptions are met and one of its models fits the data (Hambleton & Swaminathan, 1985). In classical test theory, item statistics are used for test construction purposes. But these item statistics (p's & r's which are obtained by field testing items) are group-dependent, and the examinees scores are dependent on the sample of items administered. These dependencies undermine the equivalence of test forms and their use across groups of examinees. IRT provides item statistics that are independent from the examinee population, and ability scores that are independent from the particular sample of items. These features are potentially valuable in criterion-referenced test construction and uses, particularly in developing item banks, selecting items for a test, and comparing examinees to a common standard though they may have taken different forms of the same test.

In test score reporting, IRT test characteristic curves can be very useful. The ability scale and item parameters are used to obtain the item and test characteristic curves, and each examinee's score can be estimated using these curves. Predictions can also be made about the performance of examinees with certain abilities on any set of items in an item pool. The scores could be reported in any metric, at school levels, at district levels, or at any other desired level. The standard error of measurement for each score can also be added to the reports to enhance accuracy of ability score interpretations. In other words, IRT enables the reporting of the measurement error for each examinee.

In choosing appropriate test lengths for criterion-referenced tests, practitioners often worry about imprecise domain score estimation and incorrect mastery classification of examinees. That often leads to the preference of long tests when classical test theory is utilized. In an IRT framework, the relationship between test length and decision consistency can be formulated (see Hambleton, Mills, & Simon, 1983), and short tests with reasonable decision consistencies and accuracies can be constructed by using suitable item selection procedures. This is accomplished by selecting items that provide most information and hence provide least errors of measurement at any ability of interest-often at the cut-off ability score. Hence, the domain score estimation problem and the decision consistency and accuracy problem are both addressed by using a suitable item selection method.

In test construction, item selection methods based on IRT are generally superior to classical approaches (see Hambleton, Arrasmith, & Smith, 1987; Green, Yen, & Burket, 1989). When constructing tests within a classical measurement framework, items with high biserial correlations and moderate difficulty values (.3 to .8) are selected during the test construction process. The objective that each item measures, and the technical qualities of the items are also considered. Item and test information functions, which stand the place of test reliability in classical test theory, are used in IRT-based methods of item selection. Items are selected on the basis they provide desired information at specified points along the ability scale, and the information function is inversely related to the standard error of measurement at any ability (Hambleton & Swaminathan, 1985). With criterion-referenced tests, usually the number of specified points of interest is one; i.e., the point where the cut-off score is located. A good feature of the item information function is its additive property. The test information function is given by adding up the information functions of the constituent items (Lord, 1977; Hambleton & Swaminathan, 1985).

In constructing tests, one may begin with specifying the standard error of estimation that one can tolerate at a particular ability range or point. Suppose we need to measure a normally distributed ability range of +2 to -2 with standard error of .35. The information at this ability range should be 8.16 or higher. Consequently, items are selected from the pool which contribute to the test information up to the desired level at this ability range. Lord (1977) provides a heuristic procedure for selecting items from an item pool: (1) Decide on the desired test information function; (2) select items that will cover the hard-to-fill areas under the chosen information curve; (3) calculate the resulting test information each time an item is added to the test; and (4) continue until the test information satisfactorily approximates the target information. Content composition, and other psychometric properties of tests that are developed through the use of item information are not neglected but regarded during test construction (see, for example, Ackerman, 1989).

There are several methods of item selection some of which are based on classical test theory and some of which are IRT-based. These methods include the random method in which items are selected randomly from item pools, a classical method in which items with high biserial correlations and moderate difficulties are selected, an optimal method in which items that provide the most information at ability level of interest are selected, and "content-optimal" methods in which items that provide the most information at the ability level of interest are selected while other requirements on the resulting test such as content composition are considered too.

The choice of an item selection method will have an effect on the reliability and validity of the resulted test scores (Hambleton, Arrasmith, & Smith, 1987; Hambleton, Dirir, & Lam, 1992). In criterion-referenced tests, the item selection method has effects on the decision consistency and accuracy. That is especially true when the items in the bank differ in properties that have a notable impact on the results of the test, and when the presence or absence of certain group of items affects the test scores.

Currently, automated item selection methods are receiving more attention among test developers and practitioners. The development of powerful computers has played an important role in the case of automated test development procedures, and many test publishers are using or considering these
approaches to-day (Green, Yen, & Burket, 1989; Stocking, Swanson, & Pearlman, 1990). The automated item selection methods, which use IRT-based item parameters to compute item and test information functions, utilize linear or integer programming procedures and optimization algorithms. Some

literature related to this topic will be reviewed in the next chapter.

The advantages of IRT in solving practical problems in criterion-referenced tests are fairly well established, but the advantages do require that model assumptions are satisfied. One of the critical assumptions of several popular IRT models is that test data should be unidimensional. This assumption is not always met; and hence, it is important to investigate the robustness of the IRT models to the violation of this assumption. In the next three sections an overview of test dimensionality, how it is assessed, and how the unidimensionality assumption is often violated in practice will be discussed.

Background on Dimensionality

The number of traits a test measures is one of its critical aspects. A comprehensive review on the evolution and indices for assessing test dimensionality was done by Hattie (1985). He reported that interest in the issue goes back as far as 1940s, and that more than 80 indices were proposed to assess test dimensionality. These indices vary from those based on answer pattern and test reliability to those based on nonlinear factor analysis and nonparametric approaches. Mislevy (1986) and McDonald (1989) both discussed approaches that are recently being used for dimensionality assessment. These methods include generalized least-squares solutions, and maximum likelihood solutions.

Despite the widespread attention in the topic, several contrasting definitions has been proposed for test dimensionality, and it is sometimes confused with such concepts as homogeneity, reliability, and internal consistency. In common practice, whether a test is unidimensional or not is often assessed, and hence the definition of dimensionality is often based on the unidimensionality of a test. Hattie (1984) distinguished unidimensionality from other terms or methods which do not define it but are used to determine it. He contended that unidimensionality is not defined in terms of unit rank, percent of variance explained by first factor, deviation from perfect scale, type of correlation, or the number of common factors. Dimensionality is the number of abilities that influence the performance of examinees on test items. Hattie (1985) asserted that unidimensionality is "the existence of one latent trait underlying the data" (p.157).

McDonald (1982) insisted that if only one trait influences the distribution of the response patterns of items, then the set of items is unidimensional. His definition is based on the principle of local independence which states that for any fixed ability, the examinee responses to binary items are mutually statistically independent. He claimed that the principle of local independence is basic for the definition of latent traits, and that unidimensionality could not be explicated without the definition of the latent traits. Other methodologists do not agree with this argument. Goldstein (1980), for example, contended that the dimensionality of a test could be specified regardless of the state of the assumption of local independence. He wrote "we can have one-dimensional model such as the logistic either with or without local independence" (p. 239).

The interest in test dimensionality stems from the fact that many measurement models are based on the assumption of test unidimensionality. In other words, measurement practitioners and test users assume that all items in a test measure one trait. Stout (1987) argued that test unidimensionality is important because: (a) it is essential for accurate test interpretation; (b) many measurement models assume it; and (c) we cannot measure individual differences without it (pp. 589-590). However, there are multidimensional test models that are currently in use, but these models have not received nearly as much publicity and usage as compared to the unidimensional models. There is no question about the desirability of unidimensional test models, and the measurement of one ability leads to sound judgement on the performance of examinees. An earlier statement by McNemar (1946), which is related to attitude tests, and which is quoted by many researchers, is as follows:

> Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank can be quantitatively and,

> within limits, qualitatively similar in their attitude towards a given issue. As an example, suppose a test of liberalism consist of two general sorts of items, one concerned with economic and the other with religious issues. Two individuals could thus arrive at the same numerical score by quite different routes. Now it may be true that economic and religious liberalism are correlated but unless highly correlated the meaning of scores based on such a composite is questionable (p. 268).

According to this argument, even two correlated traits cannot be trusted to measure individual differences in a test score.

In a way close to the above assertion, and also related to rank ordering of examinees, Hattie (1985) insisted:

> to make psychological sense when relating variables, ordering persons on some attribute, forming group on the basis of some variable, or making comments about individual differences, the variable must be unidimensional; that is, the various items must measure the same ability, achievement, attitude, or other psychological variable (p. 139).

From this viewpoint, a test must measure just one trait to foster valid conclusions about examinee performance, and optimal rank ordering of examinees might not be attained when the test is not unidimensional. Test unidimensionality is an issue not only for tests intended to measure individual differences and rank order examinees, but also for tests intended to measure whether examinees mastered specific tasks. In other words, the dimensionality of tests affects both criterion-referenced and norm-referenced tests. It seems test specialists have usually been concerned about how the violation of the unidimensionality assumption may affect norm-referenced tests. Less attention has been paid to the impact of the number of dimensions in a test on the psychometric properties of a criterion-referenced test. In short, the effects of test dimensionality on decisions based on mastery tests are no less important than the effects of test dimensionality on rank-ordering examinees.

Assessment of Dimensionality and Related Research

Many methods are currently available for the assessment of dimensionality of a set of test items (see, for example, Hattie, 1984; Hambleton & Rovinelli, 1986; Mislevy, 1986; Knol & Berger, 1991). Some of the widely utilized techniques are linear factor analysis, hierarchical factor analysis, nonlinear factor analysis, and nonparametric approaches. The classical factor analysis and its variations have dominance over the other approaches in use if not in practical value. In this method, a conventional procedure of assessing the dimensionality of binary item pools is to obtain the tetrachoric correlations among the items, get the principal components or common factors, and examine the eigenvalues of the correlation matrix. This examination could entail the inspection of the percent of variance explained by the first factor, the magnitudes of the eigenvalues, and/or the differences of successive eigenvalues.

Full-information IRT models and associated contingency tables and likelihood ratio goodness of fit (Bock, Gibbons, & Muraki, 1988), partial-information factor analysis models (Christofferson, 1975; Muthen, 1978), nonlinear factor analysis (McDonald, 1967a), models that test the hypothesis of unit dimension in nondecreasing monotonic item response functions (Rosenbaum, 1984), and methods that use examinees scores on subset of items to test the dimensionality of the test (Bejar, 1980; Stout, 1987) are other methods which are all currently in use. These models are based either on IRT formulation or common factor analysis formulation. The equivalence of the two formulations has attracted the attention of some researchers, and it was concluded by many that the two approaches are equal (McDonald, 1982, 1985; Takane and De Leeuw, 1987). McDonald (1985) and Takane and De Leeuw (1987) separately proved that the two-parameter normal ogive model and the factor analysis of dichotomized variables as discussed in Christofferson (1975) and Muthen (1978) are in fact equivalent.

Christofferson (1975) introduced a factor analytic approach for dichotomous items using the marginal distributions of single and pairs of items. The loss of information in this procedure compared to full-information maximum likelihood methods is compensated for the less computations it requires. In this model, a set of continuous variables, which are fitted by common factor model, are dichotomized by using threshold values to get binary item responses. The threshold values and the factor parameters are then jointly estimated.

The proportion of examinees passing each item are obtained, and the proportion of examinees passing each pair of items is approximated. Finally, estimates of sample proportions are used to fit the model by generalized least squares (GLS). The GLS estimator is asymptotically distributed as a chi-square with known degrees of freedom, and it could be used to test the number of significant factors in the data. A computer program based on this model was developed, but up to now more than 25 items could not be used on it. Otherwise, the model was rated as promising and adequate (see Hattie, 1984; McDonald, 1985). Muthen modified this model and made it computationally faster by using sample tetrachoric correlations instead of sample proportions passing pairs of items, but the limitation in the number of items has yet to be solved.

McDonald (1967a, 1967b, 1985) developed a nonlinear approach of factor analysis which he suggested would improve upon the usual linear factor analysis that is used in assessing test dimensionality. McDonald (1982) classified common factor models into three categories: 1) those that are linear in both their coefficients and latent traits; 2) those that are linear in their coefficients but non-linear in the latent traits; and 3) those that are non-linear in both coefficients and latent traits (p. 380).

Examples of cases 1 and 3 are Spearman's general factor and IRT logistic model, respectively. He contended that the popular logistic and normal ogive models are nonlinear transformations of the Spearman's general-factor model that are specific for dichotomous items. He further noted that linear approximations could be employed to fit such nonlinear models to datasets. McDonald (1982, 1985) advocated the case in which the functions are linear in the coefficients but not in the traits to be fitted to datasets. He also showed that by harmonic analysis, the normal ogive model can be approximated as closely as desired by polynomial series, and recommended that up to the cubic term would provide good approximations.

Currently research on dimensionality of tests mainly focus on three areas: (1) robustness of IRT unidimensional models to the violation of the unidimensionality assumption; (2) assessment of the performance of various indices of unidimensionality, and (3) the effect of multidimensionality on test uses such as parallel test construction and test scoring. Fewer studies are undertaken in the area of the performance and uses of multidimensional models. In the first category, the original (true) item parameters of the test and the item parameters estimated by the model under investigation are compared. The central question in these studies is: Does the model estimate the item parameters properly when the data is not unidimensional? The relationship between the estimated and true parameters is often examined; mainly by using correlational techniques, and the relationship of the two sets of parameters is used to evaluate the robustness of the model to the violation of the unidimensionality assumption.

In the second approach of dimensionality research, the strengths of various indices to detect multidimensional tests are studied and/or compared. However, these indices differ in their assumptions, uses, and limitations, and each is intended to highlight the dimensions of a test in its own way. Most of the indices are based on judgmental, subjective decision making approaches in which the number of dimensions are determined. Many of them do not have associated statistics, and many of them do not assess how dominant the dimensions in a test data are relative to each other. Each index has its own rules of detecting multidimensionality, and there might not be clear cut criteria for comparing all of the different indices.

The last line of research, the effect of dimensionality on test use, is not as well developed as the other two. Since it is known that there are no strictly unidimensional tests, it is reasonable to probe how dimensionality might influence test score interpretations and subsequent decisions based on it, and address issues like its effects on test construction, results, and uses. Then guidelines can be developed for avoiding or minimizing the effects of multidimensionality. In criterion-referenced tests, it is important to know how test dimensionality affects the reliability and validity of mastery classification decisions, for example.

Some Causes of Test Multidimensionality

Tests are supposed to conform to the unidimensionality assumption required by most of the currently used measurement models (see, for example, Hambleton, Swaminathan, & Rogers, 1991). However, since tests are constructed to meet other criteria such as the presentation of different domains focused on different abilities of examinees in a single test, the presence of different topics of the same subject matter, the satisfaction of targeted test and item statistics, and not to meet specific factor structure, unidimensionality can be violated in different ways and for many sound reasons. Traub (1983) discussed three possible causes of test multidimensionality: Differences in instruction and educational effects among test takers, test speededness, and examinees' tendency to guess. He cautioned against the effects of multidimensionality on test results that might be caused by using IRT models with achievement tests.

Tests could be multidimensional for other reasons too. It has been noted by many researchers that tests could be multidimensional because of presence of minor and unintended traits beyond the major trait the test is purported to measure (Drasgow & Parsons, 1983; Nandakumar, 1991; Stout, 1987; Harrison, 1986). Stout (1987) introduced the concept of essential unidimensionality in which he suggests that tests often have one dominant trait and one or more minor traits. He added that the potency of the minor trait(s) determine the test dimensionality, and that test are essentially unidimensional as long as the minor dimensions are less potent.

For instance, besides the ability the test is intended to measure, a test may be affected by a second trait which might have less influence on the test items, and could affect all or part of the items. For example, reading could be a minor trait in a physics test where the major ability to be measured is physics knowledge. In some situations, there are even more than one minor ability beyond the major trait, and these minor abilities could affect all items of the test or each could be influencing different clusters of items. Mathematics knowledge and reading proficiency could be minor abilities that may affect all items in a physics test. In other instances, different parts of a test may require different strategies of test taking or different abilities for the examinees to answer test items correctly. In yet other situations, different topics of the same subject may require disparate minor abilities in addition to the major trait needed for the mastery of the subject.

The arrangement of tests into different sections, and its composition of different content areas might introduce lack of unidimensionality in the strict sense used in IRT. Each section of a test may require, albeit minor different abilities in responding to the test items. Similarly, each content category may load on a different minor dimension in a testing situation, and the presence, representation, or absence of certain content areas may have an effect on the dimensionality of the test. In all the above mentioned cases, the degree of departure of the test from unidimensionality is related to how "minor" are the minor abilities. The degree of departure could be influenced by such factors as the potency of the major ability, the number of items affected by each minor ability, and the number of minor abilities.

Tests are usually developed by choosing items from large item banks, and it can be argued that the items in these item banks are not strictly unidimensional. In the process of test construction, especially when IRT-based techniques are used, many attributes of items are considered; e.g., their information functions, content, format, the frequency of their use, dependencies among items. Item information functions and item content categories are often used more than the other characteristics in item selection. When both attributes are considered, "content-optimal item selection" results, and items are selected according to the amount of information they provide at the ability levels of interest and according to their content. If item information alone is used in the test design process, it may lead to an unbalanced test in terms of content, and may also lead to a multidimensional test when the item pool is not unidimensional. The reason is that when sampling items in this manner, one is seeking items with desirable properties; that is, items providing most information at the ability level of interest, and these items may load on a specific trait, and hence may affect the dimensionality of the resulting test.

Imagine a test development situation in which items are being selected from an item pool which has many content categories, and which is believed to be a relatively unidimensional item pool. Suppose the content categories represent minor dimensions. If specific content categories are oversampled or selected more than the other categories during the test construction process, the dimensionality of the resulting test might be affected, and this might subsequently reduce the reliability and validity of the test scores.

Purposes

Criterion-referenced tests are being used by many educational and professional organizations for a variety of purposes. Item response theory provides a useful framework and models for the development and use of criterion-referenced tests. The merits gained from using IRT are fully realized when its assumptions are met and the model of choice fits the test data. One of the crucial assumptions that is difficult to meet in practice is the assumption of test unidimensionality. The IRT-based item selection methods used for test construction might even contribute to the problem of multidimensionality because items influencing specific minor traits might be selected through a particular item selection method. In test design processes, the dimensionality of an item pool, the item selection method, and the interaction effects of the two are expected to affect the reliability and validity of criterion-referenced test scores and decisions.

In view of the previous research on test dimensionality, and its central importance to the reliability and validity of criterion-referenced test scores and decisions, this research study has been designed to investigate several important questions:

1. What are the current methods of choice in investigating test dimensionality?
2. How do various amounts of test dimensionality impact on the ability estimation and decision accuracy of criterion-referenced tests?
3. How do item selection methods interact with test content to influence the reliability and validity of criterion-referenced tests?

The first question addresses some essential background information for the study and will be addressed by a comprehensive literature review. The second question will be addressed via a number of carefully designed simulation studies. The third and final question will be addressed using some real data provided by one of the national credentialing organizations.

The study was based on a hypothetical situation where a test was being constructed from an item pool. The examinee responses to items in the pool were assumed to be accounted for by a general ability, and in addition, some minor traits exist which were specific to specific clusters of items. The relative potencies of the general ability and the minor abilities is what constitutes the dimensionality of the test, and that was manipulated in the study.

Both simulated and real data were used in this study. In the simulation part, examinee response data with known degrees of departure from unidimensionality were generated. The computer program used for this purpose was based on the concept of essential unidimensionality developed by Stout (1987). The real data comes from a national credentialing examination administered in December 1988. The exam consists of over 200 items, several content categories, and over 5 item formats. The content categories were treated as minor traits that were tapping different abilities even though the whole test was measuring a general ability.

Significance of the Research

Criterion-referenced tests, which have not received attention equal to that of norm-referenced tests when it comes to the issue of dimensionality, was the focus of this study. Of special interest was how dimensionality affects the decision accuracy of tests; that is, passing masters and failing nonmasters. In criterion-referenced tests, dimensionality might be caused by different objectives or content categories reflected in the test, different cognitive levels, or different item formats. Whichever it might be, these differences might correspond to different abilities, and hence might influence the dimensionality of the test.

One of the main uses of criterion-referenced tests is to assess whether examinees have mastered specific curriculum objectives or tasks. These objectives or tasks covered by the test are often assumed to measure just one ability by those models used for scoring, decision making, and for other testing purposes. What could happen if the objectives or items of the test are measuring several abilities? Are our pass/fail decisions accurate in these cases? Or more generally, are our decisions in passing or failing examinees equally accurate in unidimensional and in multidimensional tests?

If we desire to report the scores of the different dimensions in a multidimensional criterion-referenced test, would it lead to decisions more consistent than aggregating the whole scores and basing our decisions on the average score? Would the relative strengths of the different traits in the test be a factor in our decision making? On the other hand, when a test is constructed by selecting items from an item bank, does the utilized item selection method contribute to the multidimensionality problem? Do the item selection methods have effects on decision accuracy? Answers to these questions were addressed in this research.

In the next chapter, literature related to test dimensionality, applications of IRT to criterion-referenced tests, and optimal test designs will be reviewed. The data simulation and data analysis procedures will be fully discussed in Chapter III. Results and discussion follow in Chapter IV, and the summary and conclusions in Chapter V.

CHAPTER II

LITERATURE REVIEW

In this chapter, studies related to dimensionality, studies on IRT approaches to item selection methods in criterion-referenced tests, and studies on automated item selection methods will be reviewed. Studies on dimensionality will be grouped into four categories: a) studies on the robustness of unidimensional models to the violation of the unidimensionality assumption; b) studies on comparisons of different indices proposed for dimensionality assessment; c) studies that present the item parameters of multidimensional data in polar coordinates and address different issues in testing; and d) studies that use a nonparametric approach in investigating test dimensionality.

Model Robustness Studies

Drasgow and Parsons (1983) studied the robustness of the widely used IRT program LOGIST to the violation of the unidimensionality assumption. They addressed the problem in a classical hierarchical factor analytic approach, and used a model developed by Schmid and Leiman (1957). The model is set in such a way that the test (all items) is influenced by a single general latent trait, while some clusters of items are affected by specific factors. They asserted that the first-order common factors in the examinee responses are correlated, and their correlation is accounted for by a second-order general factor which is the underlying general trait measured by the test. In addition, second-order group factors that are specific to certain clusters of items exist. This approach is attractive and has practical appeal; the second-order general factor could be general mathematical ability, for example, while the second-order group factors are related to specific mathematics topics such as algebra, calculus, and geometry.

Drasgow and Parsons (1983) generated five data sets with varying degrees of dimensionality. These degrees, which were set in terms of the correlations among the first-order common factors were controlled by the parameters of the second-order general and group factors. They formed a matrix of factor loadings of items on the first-order common factors, which was simple structure, loadings of first-order common factors on the general factor, and loadings of the first-order common factors on the second-order group factors. The data sets varied from strictly unidimensional in which factors were perfectly correlated to five-dimensional data in which factors were almost uncorrelated (.02 to .14). The researchers found that as the potency of the general factor decreases the accuracy of the LOGIST estimation decreases. They recommended that LOGIST would provide accurate estimates when the first-order common factors have correlations of .46 to .6. When the correlations among the factors are smaller than .45, they insisted, one may find inaccurate item and ability estimates, and LOGIST will be drawn to one of the common factors instead of the general ability.

One limitation of this study was that a fixed number of traits were used, although a reasonable number was chosen. The number of items per factor which ranged from 5 to 15 was somewhat restricted, and one may wonder what could have happened if wider ranges were used. No replications were made in the study, and the number of items was fixed at 50. The data generation method was chosen to fit a factor analysis model to the data instead of an IRT logistic model, and a relationship between the parameters obtained from the normal ogive model and the parameters obtained from the factor analysis was used in the study. But this relationship is especial for unidimensional data, and whether it holds in multidimensional cases is doubtful. Finally, this study was close to real testing situations in the sense that a test usually measures one general trait and a number of minor abilities. The unidimensionality assumption is often violated through the presence and potencies of the minor abilities.

Harrison (1986) investigated the robustness of IRT parameter estimation in LOGIST to a violation of the unidimensionality assumption. He followed an approach similar to that of Drasgow and Parsons (1983). He further varied test length (30, 50, 70), number of common factors (4, 8), and distribution of items loading on each factor (uniform, or highly skewed). The design he used, in terms of correlations among abilities, was similar to that of Drasgow and Parsons (1983). The estimation of the discrimination parameter was better for the longer tests, the stronger general factor, the uniform distribution of items among factors, and the larger number of factors. The estimation of the difficulty parameter was affected in a way similar to that of the discrimination by all factors. The trait estimates followed the same trend; better estimation was obtained for longer tests, stronger general factor, and uniform item distribution. LOGIST was drawn to the stronger group factor (the one loaded on by most of the items) in the case of the skewed item distribution, and difficulty in estimation was reported in the 30-item, four-factor case. But that is not surprising since the three-parameter logistic model requires around 50 items to provide adequate estimates. The shortcomings of the study were similar to those of the study by Drasgow and Parsons (1983). Furthermore, tests that measure as many as eight traits may not be found in practice.

Drasgow and Lissak (1983) used modified parallel analysis to investigate its effectiveness in assessing dimensionality of binary items. In this procedure item responses were generated using a method similar to that of Drasgow and Parsons (1983). The item and ability parameter estimates of the data were then used to generate artificial data. The plots of the eigenvalues of the corresponding datasets were compared. The ith factor extracted from the original data was considered to be real if its eigenvalue exceeded that of the ith factor of the second set of data. The results of this study matched those reported in the Drasgow and Parsons (1983) research. One expects that corresponding factors have close eigenvalues since the two datasets are so related or the artificial data depends upon the first data. The equivalent item parameters in the two datasets is of concern in comparing the eigenvalues of the extracted factors, and how this will affect the results is not clear.

Reckase (1979) utilized linear factor analysis in investigating the estimation of the 1- and 3-parameter logistic models when used with multidimensional data. The question of interest was how these models perform when used with multidimensional data. Four datasets were generated: (1) One-factor dataset with loadings of 0.9 on each item; (2) two-factor dataset with randomly distributed loadings of 0.9 on items; (3) nine-factor dataset in which there was a dominant factor of 0.7 loadings on all items, and items randomly distributed to other eight factors with 0.6 loadings; and (4) nine-factor dataset with items randomly distributed to the factors with either 0.9 or 0.0 loadings. For the two-factor case, Reckase found that the 3-parameter model was drawn to the second factor, and the 1-parameter model was measuring the sum of the two factors. For the nine-factor simple structure case, he found that the three-parameter model estimates were highly correlated to factor nine, and the one-parameter estimates were highly correlated with the raw scores.

In the data with the one dominant factor and number of specific factors, both models estimated the first dominant factor. Even when used with classroom tests, the first factor was measured in most of the cases. Reckase also addressed the question of how strong should the first factor be in order to get reasonable estimates. He insisted that the first factor should have an eigenvalue of 10 or greater, or account for at least 20 percent of the variance of a 50-item test. He also added that good ability estimates might be found when the variance explained by the first factor are less than 10 percent, but that the item parameter estimates will be unstable. Besides the first unidimensional case which was used as baseline, all of the factor structures in the study have weaknesses. In case two, two orthogonal factors may not be found in real live testing situations, and in case three as many as nine factors in one test data is not common. In the close-to-reality case of a dominant first-factor, it is reasonable that the first factor is loaded on by all items, but the magnitudes of the chosen loadings for the dominant and minor factors were not that different (0.7 and 0.6). In situations like this, one may argue that each item is explained by two factors. Moreover, the data were generated to fit a factor analysis model and not an IRT model.

Ansley and Forsyth (1985) studied the IRT unidimensional estimates derived from two-dimensional data. They generated two abilities with correlations of .0, .3, .6, .9, and .95. Their study was different from those of many others; they used a noncompensatory model while other researchers used compensatory models. In the noncompensatory models, if an item is measuring two dimensions, an examinee with low ability in one dimension and high ability in the other will have low probability in answering the item. The high ability in one dimension will not compensate for the low ability in the other dimension. The reverse is true for compensatory models. They used sample sizes of 1000 and 2000, and test lengths of 30 and 60. With the exception of the 30-item test length, all the other variables and values are suitable for the use of the three parameter logistic model estimation, and so we would not expect these factors to affect the outcomes of the study.

Ansley and Forsyth (1985) chose the item parameters to reflect test data that has two dimensions with one of them slightly more dominant than the other. They reported that the mean of the estimated discrimination values were between the means of the two discrimination values of the two dimensions, and that it approached the value of the first dimension as the correlation between the dimensions increased. The estimated b- values have means and standard deviations that were higher than those of the difficulty values used to generate the data when the correlations between the traits were low. Both mean and standard deviation decreased as the correlation increased. The correlations between the estimated mean b-values and the true b-values (b_1 and b_2) were all high compared to those of the discrimination parameter.

For the ability parameters, the correlations between the estimated and generated abilities increased as the correlation between the traits increased. At low correlations between the two dimensions, the estimated ability was correlated with the first trait, and at the highest correlation between the two dimensions, the estimated ability has equal correlations with the two abilities. In the latter cases, the estimated ability was most highly related to the average of the original abilities, and the design became close to unidimensional. Disparate results, however, were reported for these datasets and unidimensional data. The correlations of estimated and true parameters of the unidimensional data were higher than those found for the two-dimensional data, and their average absolute differences were smaller. One limitation of this approach was that the choice of the noncompensatory model was not justified, and we may question whether the LOGIST program is equally suitable for compensatory and noncompensatory approaches. Another limitation was that no check was made to insure the dimensionality of the data, and the number of abilities was only two in the multidimensional cases.

Studies on Dimensionality Indices

Hattie (1985) classified the indices proposed for unidimensionality assessment as those based on answer pattern, those based on reliability, those based on principal components and factor analysis, and those based on latent traits. These indices were developed with the other developments of the testing field, and have been replaced by subsequent indices after their flaws had been discovered. In addition to these indices, there are nonparametric indices that are currently in use such as Stout's T statistic and Bejar's method of correlation. Yen's Q3 is also used by some researchers to assess the unidimensionality of test items, and indices based on residuals after fitting a model to the data are getting more attention and applications. Many more indices may be developed in the future as well.

Hattie (1984) used the classical factor analysis approach to assess the relative merits of various indices used for testing unidimensionality. Despite the fact that he used small number of items, his approach was beneficial. He simulated 1-factor, 2-factor, and 5-factor datasets in which the factors in the multifactor cases had correlations of either 0.1 or 0.5. The three-parameter logistic, compensatory model was used for the data generation. The study was based on the notion that factor loadings and item discrimination values are related, although the relationship was not mentioned in the study. For the two- and five-factor cases, the simulation was manipulated in such a way that factor intercorrelations of .1 or .5 were produced. First, discrimination values of 1 were formed into a simple structure pattern, and then postmultiplied by a triangular matrix decomposition of factor correlation to provide actual factor loadings to be used in the simulation. Abilities were normally distributed with mean zero and unit variance, and the difficulty values were uniformly distributed between -2 and 2.

Four stage analysis was made to assess the effectiveness of 87 indices in distinguishing between 1-factor and more than 1-factor data. The first criterion was the means of the indices in which it was expected that the mean of each index for one-factor case should be larger or smaller than the mean indices for the multifactor datasets. The second criterion was a three-way MANOVA in which it was evaluated whether the values of the indices calculated from the one-factor cases were significantly different from the values calculated from the multifactor cases. In the third criterion, the number of times the one-factor mean for each index was greater or smaller than the corresponding mean of the two- or five-factor case was inspected. Finally, the number of times the value of an index in one-factor data overlapped the values in the two- and five-factor data was counted. Indices which did not pass each hurdle of the four-stage analysis were excluded from the subsequent analysis, and four indices in two programs (FADIV, and NOHARM) that utilize either the number of residuals greater than .01 or the sum of residuals were reported to be effective in testing unidimensionality. Both programs are based on the two-parameter latent trait model, and both use residual analysis.

The number of items in the Hattie study was small. The correlations among the dimensions in the multifactor cases were restricted; factors or latent traits that have correlations of .1 are almost orthogonal, and it is not unusual to find ability correlations higher than .5 in real tests. How these factors may or may not affect the various indices was not discussed in the study.

In another study, Hambleton and Rovinelli (1986) compared four methods of determining test dimensionality: linear factor analysis, nonlinear factor analysis, residual analysis, and a method developed by Bejar (1980). They used 1500 examinees and 40 items, and two traits with correlations of .1, or .6. They varied the percent of items measuring each trait (50% for each trait, or 75% for the first trait and 25% for the second). One-dimensional data were also used as a baseline, and different criteria were used in assessing the effectiveness of the different methods. For the linear factor analysis, eigenvalue plots, and the matrix of residuals after fitting the factor model to the data were used. For the nonlinear factor analysis, residuals after the model was fitted to the data were used. In residual analysis, the discrepancies between expected and observed probabilities at various ability levels are usually computed. The average absolute-valued residuals, the average absolute-valued standardized residuals, and the distribution of the absolute-valued standardized residuals were examined. Finally, in the Bejar method the correlation coefficients between two difficulty values of a subtest of items; values obtained when the subtest is analyzed separately and values obtained when the subtest is analyzed with the rest of the test, was used.

Hambleton and Rovinelli (1986) found that the linear factor analysis overestimated the number of factors in the datasets, that residual analysis failed to detect test dimensionality, and that the Bejar method was not adequate in illuminating the multidimensionality in the data in most of the cases. For the nonlinear factor analysis, they reported that the number of dimensions in the test data was accurately determined. They mentioned, however, that there were no guidelines to follow in determining the number of factors and polynomial terms to retain. But McDonald (1985) has recommended that the cubic term is sufficient, and there are computer programs such as NOHARM which could be used in fitting nonlinear factor models to binary data. The number of abilities in this research was limited to two, which were moderately correlated or almost orthogonal. In the two-dimensional data, the first trait was used to generate some of the item probabilities while the second trait was used to generate the rest of the data. How and whether this approach of data generation would affect the results of the study is not clear.

Knol and Berger (1991) investigated the relative effectiveness of traditional factor analysis models and IRT multidimensional models to assess the dimensionality of tests. They grouped the IRT models into those which use full-information in the data and those which use partial information contained in the response data. Specifically, they studied the models implemented in TESTFACT, MAXLOG, NOHARM, and those in traditional factor analytic methods such as MINRES and iterative principal factor analysis. Knol and Berger (1991) simulated data of three sample sizes (250, 500, 1000), and three 15-item tests and one 30-item test of varying numbers of dimensions (1, 2, and 3). The number of items was small, but many of the programs they used could not handle large tests.

The criteria they used to compare the programs were in terms of mean squared differences between the true and estimated item parameters, and they divided the criteria into factor analytic and IRT. In the 1-dimensional data, they reported that TESTFACT performed best in both factor analysis and IRT criteria, and NOHARM performed adequate in both criteria. They also reported that the common factor methods performed well with the IRT criteria. In multidimensional data, NOHARM did better than TESTFACT with factor analysis criteria, and factor analysis models did better than TESTFACT with IRT criteria. The researchers concluded that factor analysis methods performed the same or better than IRT full-information models, and that NOHARM did better than its IRT counterparts. Other multidimensional IRT models such as LISCOMP and MIRTE were not included in the study, and the number of items in the study were restricted. The main conclusion of the study was that classical factor analysis are not less effective than the theoretically sound IRT full-information models in detecting the number of dimensions in test data.

Roznowski, Tucker, and Humphreys (1991) compared three indices of unidimensionality: index based on local independence, index based on second-order loadings, and one based on eigenvalues. None of the indices was reported to be satisfactory, but the local independence index was found to be better than the other two, and the eigenvalues index was rated to be the worst. What values would make these indices satisfactory were not mentioned, and recommendations for alternative indices were not made.

Polar Coordinate Studies

Reckase (1985) and others developed another approach of looking at test dimensionality. They introduced multidimensional models in which the item parameters are represented as a vector in the latent space. Three assumptions of these models were: a) probability of answering an item correctly increases monotonically with each dimension being measured; b) assumption of locating each item at a single point in a multidimensional space; and c) the most reasonable point in defining the difficulty of an item in the multidimensional space is where the item is most discriminating, or most informative.

The item difficulty and item discrimination are represented in a polar coordinate format where the direction cosines of the angles of multidimensional difficulty (MDIF) determine the item characteristics as a vector in the latent space. The angle is a measure of the composite of abilities which the item measures, the signed distance from the tail of the vector to the origin is the magnitude of the MDIF, and the length of the vector is the item discrimination. Items with same direction cosines measure the same composite of abilities, a fact that may lead to conclude that items with same direction cosines fulfill IRT unidimensionality requirement although more than one ability is measured. In this modelling, orthogonal abilities are assumed, a fact that will unfortunately limit its use. Add also that more than two abilities were not addressed in the studies that used this model so far, and one may question if the method can handle more than two latent traits, or oblique traits.

To demonstrate the effectiveness of this procedure, Reckase (1985) analyzed a 40-item mathematics test using a program based on a multidimensional two parameter model (M2PL), and the resulted item statistics were compared to results obtained by analyzing the same data with LOGIST and by classical item analysis. An interesting feature in this study was that the first ability mostly measured easy items and the second trait measured the relatively difficult items. Obviously, the item difficulty and dimensionality were confounded. In a correlation analysis of the parameters, the a-parameter from LOGIST was highly related to the second a-value of the two-dimensional analysis, indicating that LOGIST estimated ability of the second dimension of the M2PL. All difficulty parameters of the three analyses were highly related, indicating that the difficulty estimation of the M2PL is adequate. However, the correlations among the discrimination parameters were low, and the high correlations among the b-values could not provide much information about the dimensionality of the test.

Reckase, Ackerman, and Carlson (1988) showed that a two-dimensional test can be robust to the unidimensionality assumption. Both real and simulated data were used to prove this argument. In the simulated part, data with two orthogonal dimensions were generated by using M2PL. The real data consisted of responses of 2738 examinees to 68 multiple-choice items composed of 40 mathematics items and 28 social studies reading items. In the simulated data, the first 20 items measured θ_1, the second 20 items measured θ_2, the third 20 items measured both traits, and the multidimensional difficulty of the last 20 items had directions equally spaced between 0 and 90 degrees with the first trait. Two analyses were made of these data: multidimensional analysis using M2PL and specifying two dimensions, and unidimensional analysis using LOGIST. Yen's Q_3 statistic (Yen, 1984) was used to determine the violation of the unidimensionality assumption. Clusters of items that measured the same composite of abilities (approximately same alpha-vectors) were identified in the real data, and these clusters were reanalyzed as unidimensional subtests, again, computing the Q_3 statistic for each subtest.

In the multidimensional analysis of the simulated data, the four subtests separately analyzed by LOGIST, and the subtest that measured equally both abilities when analyzed with the rest of the test using LOGIST did not violate the unidimensionality assumption as determined by Yen's Q_3. In the real data, the subtest that had almost equal alpha-vectors but measured both mathematics and social studies did not violate the unidimensionality assumption when calibrated with LOGIST either. The rest of the datasets or subtests did violate the unidimensionality assumption. These results led the authors to conclude that items measuring the same composite of abilities could meet the unidimensionality requirement although different traits would be needed for answering the items in the test.

This study was restricted by its use of two orthogonal abilities. In the LOGIST analysis, although the 3-parameter model was used in the real data, test lengths of 16 items were used, and this is not consistent with what is often recommended for the 3-parameter model.

The vector representation modelling has attracted many researchers, and several studies based on this approach have been conducted. Findings of three of these studies are noteworthy. Luecht and Miller (1991) suggested that more accurate parameter estimates can be obtained by clustering multidimensional data and analyzing the clusters by using unidimensional IRT models. They argued that estimates from unidimensional models are more interpretable and stable than estimates obtained from multidimensional models. Ackerman (1991) studied the effect of multidimensionality on parallel forms construction when items are selected by using unidimensionally estimated parameters. He reported that parallel test forms could be constructed by using unidimensional parameter estimates and derived information functions even when the test is multidimensional. Davey and Hirsch (1991) recommended that test scoring by using unidimensionally estimated parameters provide more adequate results than their multidimensional counter parts.

Nonparametric Approach Studies

Stout (1987) introduced a nonparametric approach with an index to assess the dimensionality of test items. The index measures the degree of departure of the test from unidimensionality. The method is based on the notion of essential unidimensionality which Stout contends to be different from the strict unidimensionality used in IRT.

Basic assumptions for this approach are : a) local independence; b) random sampling of examinees from a specific population; c) independence of the response patterns of different examinees; d) a set of fixed items, possibly selected from a large item pool; and e) monotonically increasing item response functions.

The method of computing the index is straight forward. Successive steps of splitting the test into assessment and partitioning subtests, grouping examinees, computing, normalizing, and combining subgroup variance estimates, and other smoothing steps are undertaken. A basic assumption for the statistic is that when there is local independence, and the test is unidimensional, examinees with approximately equal test scores should have approximately equal abilities (Stout, 1987, p.591). The statistic is based on the fact that the theoretical variance of examinee scores on the "assessment subtest" is equal to the unidimensional variance estimate for a fixed, equal ability subgroup. Almost any binary test data could be applied to the model, no matter how large. A minor limitation is the factor analytic or subjective selection of the "assessment test" in which it is required that the subtest be "more homogenous" relative to the rest of the test. What would happen if the assessment subtest is not more homogenous than the rest of the battery, or how effective the procedure would be if there is no homogenous subtest? The selection and nature of the assessment subtest is a source of concern. One may construct a test that measures two traits and when one tries to use this model the assessment subtest could be all the items from one of the traits. The effects of the selection of assessment subtests on the performance of the index need investigation and clarification.

To highlight how the procedure works, Stout (1987) simulated five unidimensional tests that were close in terms of psychometric properties to five widely used real tests, and assessed their dimensionality by using his statistic. Although two of the tests were less than 40 items in length, he used the three parameter logistic model and what he called "three parameter piecewise linear" to generate the item responses. The number of assessment subtest items, the examinee sample size, and the nominal level of significance were all varied (not to many levels though). The three parameter piecewise linear model was included to show that the model works under nonlogistic models as well. In the one-dimensional case, the statistic was powerful in not rejecting the null hypothesis that the data is essentially unidimensional in both types of models.

To assess the power of the statistic with two-dimensional data, two normally distributed and correlated abilities were generated. An additional factor in the two-dimensional case was that each test consisted of n_1 pure items measured by one ability, n_2 items measured by the other ability, and n_3 items measured by both abilities. Five two-dimensional tests that had item parameters similar to those of the five unidimensional tests were simulated. The correlation between the abilities, and the number of items measured by each ability, were varied with cases $n_1=n_2$ and cases $n_1 \gg n_2+n_3$. The value of the guessing parameter was set at either 0.0 or 0.2. In the piecewise linear model the items were either measured by one ability or the other. The statistic exhibited good power in all cases with two-dimensional data, and the power increased as the correlation between the abilities decreased and the number of examinees increased. Under both models, the rejection rates were high.

The design was limited by the fact that only two traits were used in the case of the multidimensional data, and only two moderate correlations were used. One may also wonder why rejection rates as low as 17 percent were obtained in some cases with the multidimensional data. Another concern is why the rejection rates in the two-dimensional test with the two abilities affecting equal numbers of items was not different from the rejection rates when one ability was measuring most of the items. Finally, comparison was not made between the index and other methods used for dimensionality assessment; this would have highlighted how the index is superior or similar to other indices already in use in the testing field.

Nandakumar (1991) did another simulation study that addressed the effectiveness of Stout's index. In the unidimensional model she used, each item was influenced by one dominant ability and one minor ability. In one case there were several minor abilities each influencing small number of items, and in another case there was just one minor ability influencing all items. Due to the fact that the index is designed to be sensitive to the deviation from essential unidimensionality due to the joint variation of discrimination parameters a_1 and a_2, an index of the degree of deviation from essential unidimensionality based on a_1 and a_2 was developed. Test length, number of examinees, and the strength of minor abilities relative to the major ability were varied, and all these parameters and the degree of deviation from unidimensionality influenced the performance of the statistic. As the number of items influenced by the minor abilities increased, the rejection rates went up, and in some cases reached above the nominal level. The rejection rate also increased with the degree of deviation from unidimensionality and sample size, and decreased with test length. However, many of the tests used were less than 50 items, a fact that may prompt questions of model fit since the three parameter model was used.

In the case of one dominant trait and one minor trait, the rejection rate increased with sample size, number of items (25 and 50), and the degree of deviation from unidimensionality. It also increased with the relative strength of the minor ability, and as the value was set at .4, all rejection rates were very high. Nandakumar also assessed the performance of the index in two-dimensional data and showed that the index is sensitive to these cases, and the rejection rates were very high.

In this study, no baseline data were used to highlight the dimensionality of the tests. Other methods of test dimensionality assessment were not compared to this approach, and there is no evidence that this procedure worked better than the other techniques. It is not normal to find a test with more than ten minor abilities each having the same influence on corresponding items relative to the major ability. Also, tests having as many as 26 minor abilities might not be realistic, and if they exist at all, these tests might be expected to be multidimensional. Finally, the degree of departure from dimensionality could be influenced by many factors, not only the variations of the discrimination values. The relationship among the abilities, and the number of items measuring each ability could be factors too.

Summary of Dimensionality Studies

Some findings in the studies in the previous sections are noteworthy, and will be summarized in the following paragraphs. The studies were categorized into those that (1) focused on robustness of unidimensional IRT models to violation of the unidimensionality assumption, (2) presented the item parameters of multidimensional data in polar coordinate form and addressed different issues of testing when the data are not unidimensional, (3) investigated and compared different indices of unidimensionality assessment, and (4) studies which used a nonparametric approach to dimensionality assessment.

In the studies addressed the robustness of unidimensional IRT models to the violation of the unidimensionality assumption, it was found that the robustness mainly depends on the extent to which the test dimensions are correlated. If the correlations among the traits are high the parameter estimation of unidimensional models are adequate. If the correlations are low, on the other hand, the parameters are poorly estimated. If there is one dominant factor in multidimensional data, the model is drawn to that factor. The relative potencies of major and minor abilities were also found to have remarkable effects on the dimensionality of the tests. The discrimination parameter is found to be harder to estimate than the difficulty and ability parameters. One weakness in these studies is that in many cases the data were fit to a factor analysis model and later calibrated in an IRT model. In doing so, a relationship between factor analysis parameters and IRT parameters, which is especial to unidimensional data, is often used.

The models that utilize polar coordinate parametrization have some advantages. They introduce vector representation of item parameters in multidimensional space, and enhance the visualization of multidimensional data by spatial representation of the item parameters. These studies also shed light on a way in which the unidimensional assumption is not violated by multidimensional data; for example when items are equally measuring two abilities. One of the shortcomings of this modelling is that two orthogonal abilities are often assessed in the multidimensional cases. If more than two traits are examined, the graphical presentation could be difficult, and the effectiveness of the models could become questionable.

In the studies comparing the existing models, programs, and indices for dimensionality assessment, it was found that procedures based on residual analysis are the most effective. It was also found that traditional factor analysis methods are not less effective than IRT approaches in assessing the number of dimensions in a test. These studies often used short tests and small number of traits. Nonparametric approach to dimensionality assessment has received attention lately. Stout's procedure (Stout, 1987) is based on sound theoretical background, but has not enjoyed wide applications yet. More research is needed on this procedure, especially studies comparing the procedure with other approaches.

IRT Approaches to Item Selection

Criterion-referenced tests benefit from IRT. In test construction, for example, IRT provides item selection methods that are superior to classical methods (see Hambleton & de Gruijter, 1983). These methods are based on item and test information functions. The relative merits of the item selection procedures in developing criterion-referenced test is well documented.

Hambleton, Mills, and Simon (1983) used simulated data to investigate the effects of item pool heterogeneity, test length, discrimination values, and two methods of item selection on the decision consistency of parallel tests. The two item selection methods they used for constructing the parallel forms were _random_ and _strictly parallel_. In the strictly parallel method, items for the first form were randomly selected from the pool and the items for the second form were selected by matching their statistics to those of the items in the first form. Hambleton et al. found that the strictly parallel method was better in leading to more consistent decisions when the item pool was heterogeneous. They also found that decision consistency increased with test length, item pool homogeneity, and item discrimination values. Their study was limited to short test lengths (2 to 20 items) though short criterion-referenced tests are common in practice.

Hambleton (1983) compared the one-, two-, and three-parameter logistic models in the area of mastery/non-mastery determinations. He investigated the performance of the models in estimating domain scores and making mastery/nonmastery decisions. Hambleton found that the three models were relatively comparable in domain score estimation, and that scores were overestimated at the lower abilities and underestimated at higher abilities. In decision consistency, Hambleton found that the one- and three-parameter models provided the same rates of decision consistency at average and high ability levels, while the one-parameter model provided less decision consistencies at the lower ability levels.

Pozel and Wise (1991) studied the effects of model choice, test length, and sample size on decision consistency and accuracy. They used the content-optimal method to select either 50 items or 100 items from a pool of 142 items. The pool was a national certification examination which was fitted to the one-, two-, and three-parameter logistic models. The decision consistency and accuracy of the 50- and 100-item tests were compared for all models. Reliabilities even higher than that of the full test were obtained for the 50- and 100-item tests in nearly all models. The decision accuracy was the highest for the 3-parameter model for both tests, and moderately low for 1-parameter model and 50-item test (93.6%). These results highlight the benefits that can be gained from using IRT item selection for criterion-referenced tests; a long test can be cut to 30% without compromising the test score reliability and validity. A classical solution is possible and gains would accrue but it would be considerably more difficult to implement.

Hambleton and de Gruijter (1983) examined two item selection methods; random and optimal, for constructing criterion-referenced tests. The goal was to minimize the probabilities of misclassification (passing nonmasters or failing masters), using the smallest possible number of items. For 13 test lengths (8 items to 20 items), the researchers found that optimal item selection gave lower misclassification probabilities in all cases. They also found that substantially less classification errors resulted when both difficulty and discrimination values were used rather than using difficulty values alone.

Haladyna and Roid (1983) studied the effects of random and adaptive item selection methods on domain score estimation. In the adaptive method, the difficulty level of selected items were either close or substantially different from the examinee ability scores. Using either the random or one of three variations of the adaptive method (difficulty of selected items match the examinee ability, selected items are too easy for the examinees, or selected items are too hard for the examinee), tests of varying lengths (10, 20, 30, and 40) were constructed from an item pool. The errors in domain score estimation were compared among the item selection methods and test lengths. Haladyna and Roid found that the on-level adaptive method performed best, and the off-level methods gave the largest errors. They also found that test lengths of 20 to 30 items can provide satisfactory precision.

Hambleton, Arrasmith, and Smith (1987) compared four item selection methods in providing accurate decisions and higher information functions. The four methods were random, classical, optimal, and content-optimal. The researchers used a 249-item credentialing examination as an item pool, and as a criterion test. Using each method, a 20-item test was selected from the pool, and its decision accuracy and information function were compared among the methods. Hambleton et al. found that the optimal method provided the most information, followed by the content-optimal, and that these two methods provided better decision accuracies than the other non-IRT methods. This was true for both the total examinee population and a constrained sample which consisted of those examinees who scored near the cut-off point, and who were the most likely to be misclassified.

There is substantial evidence that optimal methods of item selection are useful for test construction in criterion- referenced testing. These methods lead to the development of short tests that are optimal in domain score estimation and classification of examinees into mastery levels. With the help of computers, the methods could be easily and flexibly implemented, and, in fact, automated.

Automated Test Development Studies

Item and test information functions are among the special features of test construction in using IRT. Computer technology further empowered the test development procedures, and made possible the inception of computer based test construction methods. These methods, which have emerged in the last decade, mainly use mathematical optimization algorithms. Linear and integer programming algorithms, which are famous in operations research, are utilized. In these algorithms, the aspects of the items to be selected for the test are often optimized subject to constraints. These constraints are some properties of the items in the pool or the test, and it could be any of the item parameters or attributes such as content, format, difficulty, discrimination, information function, and so on.

The automated test construction techniques are flexible, and are formulated to optimize some objective function which could be the minimization of test length, maximization of test information, minimization of deviations from the target information, minimization of administration time, or combinations of some of these objectives (van der Linden & Boekkooi-Timminga, 1989). The decision variable is always the selection of an item, and it takes the value of 1 or 0 for selected and not selected items, respectively. Hence, integer programming is the suitable option for the item selection problems. However, an integer programming solution can be very time-consuming (Stocking, Swanson, & Pearlman, 1990; van der Linden & Boekkooi-Timminga, 1989), and some approximations to it are recommended in the literature. These options include the following:

1. Linear solution in which the decision variables are allowed to take noninteger values, and the obtained values are rounded to zero and one.
2. Improved linear rounding in which the decision variables are ordered in descending order, and the

first n of them are rounded to one where n is the desired number of items.

3. Optimal rounding in which a linear solution is sought first, and an integer solution is sought for those variables with fractional values.
4. First 0-1 solution in which the first integer solution is considered although it is not the optimum.
5. Second 0-1 solution in which the second integer solution is considered although it is not the optimum solution.

The linear and improved linear solutions do not always meet the constraints, and the first and second 0-1 solutions need more computer time (van der Linden & Boekkooi-Timminga, 1989; Stocking, Swanson, & Pearlman, 1990). The optimal rounding method is the most favorable in terms of constraint fulfillment and computer time (ibid). The behavior and performance of automated item selection algorithms have been investigated by many researchers.

Theunissen (1985) studied the effects of the size of the item bank, target information function, IRT logistic model, and the addition of content constraints on the automated test development. He particularly investigated the effects of these factors on computer time. Theunissen used an integer solution, and reported that CPU-time increased with the size of the item bank. He also found that the location of the peak and the height of the target information affected the number of items selected. More items were needed for a highly peaked target information function, and more items were needed for targets that were peaked at points away from the mean difficulty of the item bank. As expected, the addition of the content constraint increased the CPU-time. The integer solution, however, was the slowest among the methods used for optimization problems.

Van der Linden and Boekkooi-Timminga (1989) discussed a maximin (maximizing the minimum) model in test development. They introduced a model which can accommodate the selection of items subject to several constraints such as target information, test composition, test administration time, upper and lower limits of certain item parameters or features, inclusion or exclusion of individual items, and inter-item dependencies. They mentioned the difficulty encountered in 0-1 programming in automated test construction which needs excessive CPU-time. They also mentioned the inaccuracy in linear programming which result in items with fractional values, and might lead to lack of satisfaction of some constraints. They recommended a model in which a linear solution is sought first, and the number of items with fractional values are considered as a 0-1 problem. The authors compared four different methods; optimal 0-1 solution, linear solution, optimal rounding, and first 0-1 solution. They showed that the optimal rounding solution is the most effective in terms of time, fulfillment of constraints, and finding the optimal solution.

Adema (1990) studied the effectiveness of integer programming in constructing two-stage tests. He focused the placement of constraints in developing two-stage tests, and compared when constraints are formulated for the two stages at one time and when the stages have separate constraints. Adema constructed a 20-item test from a pool of 300 items using both methods, and reported that imposing constraints on each stage at a time is easier to implement. He argued that imposing constraints for the whole test at one time may raise some difficulties, but these difficulties were not discussed in his paper. The CPU-times needed for the two types of models were not that different; 11.2 seconds for the stage level constraints and 8.274 seconds for the test level constraints.

Baker, Cohen, and Barmish (1988) investigated the characteristics of items selected through linear programming. The variables of their study were (a) IRT model (3 logistic models), (b) target information distribution (uniform & normal), (c) peak of the target information, and (d) the range of the ability of interest. Baker et al. reported that the one-parameter model requires more items to reach the desired target information than required by the more general models. Relatively large discrepancies between obtained and target information curves occurred in the middle range of the ability for the uniformly distributed information functions, and at the ends for peaked information curves. The number of items selected and the discrepancies between realized and target information curves both increased with the range of the ability of interest. Difficulty was reported in the case of the 3-parameter model and uniform target; the items in the pool (500) were not enough to provide the required information at the extremes of the target information.

The difficulty of the selected items were clustered at the extremes for all models when the uniform target was used. When normal targets were used, the b-values were clustered at the center for all models. When the two- and three-parameter models were used, the mean discrimination value of the selected items was higher than the mean discrimination value of the item pool, and the range of the values was small. The researchers observed that the linear programming solution focused on the "worst" areas of the target information; extremes for the uniform target and the peak for the normal target. Baker et al. also compared the linear and optimal rounding methods, and argued that the latter did not significantly contribute above the former although it needed extensive computer time. That finding is not consistent with the findings of other researcher (see Stocking et al., 1990; van der Linden & Boekkooi-Timminga, 1989).

Stocking, Swanson, and Pearlman (1990) reported that the optimal rounding approach did not give them satisfactory solutions when they used it in automated item selection. They introduced a model that enabled them to come "as close as possible to all constraints simultaneously" rather than not fulfilling any one of them (Stocking, Swanson, & Pearlman, 1990, p. 8). They used weights to reflect the relative importance of the constraints, and minimized the weighted sum of deviations from fulfilling all constraints. They named their model the 'successive item replacement algorithm', and it replaced items until the least deviation from satisfying all constraints is attained. Using a 480-item bank, they built 25-item tests by each one of the following item selection methods: (a) crude linear rounding, (b) improved linear rounding, (c) optimal rounding, (d) first 0-1 solution, (e) second 0-1 solution, and (f) their model. The researchers reported that their algorithm performed better than the other methods in terms of CPU-time and/or satisfying the desired constraints.

Green, Yen, and Burket (1989) discussed a computer program they use for test construction. The program uses item and test information functions, and allows the test constructor to manipulate the process in many ways. There is a feature in which content constraints can easily be added to the selection process. There is an option in which all selected items, the objectives they measure, their parameters, and the amount of information they provide at any specified ability could be seen. There is another program that displays the features of the selected items, such as standard error of measurement, the test characteristic curve, and the number of poorly fitting items used. Green et al. concluded "we are impressed with the way [the program] enables us to capitalize on the strengths of an item pool and to build a test rapidly ... we believe that it gives us very good control of the construction because of its basis in IRT" (Green, Yen, & Burket, 1989, p.308).

In automated item selection methods, precalibrated item banks that are fitted to one of the IRT models are always needed. The computer time and the realization of target features mainly depend on the size of the item bank, the number of constraints, and the programming solution. The optimal rounding method is more effective than integer and "strictly" linear solutions. The desirability of automated item selection methods is well understood, and it is hoped that the method will receive wide applications in the near future.

Computer Programs

In this section three computer programs that are suitable for this study will be reviewed.

TESTSIM

This program was developed by Stout and his associates (1991), and builds on the concept of essential unidimensionality introduced in Stout (1987). The program generates examinee binary responses from multidimensional or unidimensional IRT logistic models. It can create data with any of four models:

1. Strictly unidimensional model. Generates strictly unidimensional data. The examinee abilities are

normally distributed with mean zero and unit variance, and the item parameters are sampled from normal distributions, with user specified means and variances.

2. Essential unidimensional model with two abilities. This model generates tests with one dominant and one minor dimension. Both traits influence all items but in different degrees. The influence of the minor trait decreases with the number of items. Abilities are bivariate normal with zero means and unit variances, and they are uncorrelated. The b- and a-values are generated from normal distributions with specified means and variances.
3. Essential unidimensional model with many traits. This model simulates tests with one dominant trait and several minor traits. Each minor dimension influences a subset of items, while the major ability affects the whole test. Two parameters chosen by the user are essential in this model; the number of minor traits and the strength of the major ability relative to the minor abilities. The examinee abilities are generated from N(0,1), and the item parameters are normally distributed with user specified means and variances. If the test is desired to be unidimensional, both the number of minor traits and relative strength of minor traits should be small.
4. Two dimensional model. In this model, tests with two dimensions are simulated. As before, the user

specifies means and variances of the a and b parameters. The user also specifies in this case the correlation between the two traits.

For all models, the guessing parameter is set to a constant. The program simulates situations that are close to real data, especially model 3. It is flexible, and the user is provided many options to generate data. One limitation is that the program generates normally distributed item and ability parameters only.

NOHARM

This program, which is written by Fraser (1983), fits the multidimensional normal ogive IRT model to binary data. It is based on a theory developed by McDonald (1967a, 1982), and approximates the normal ogive model by a polynomial series. The output of the program contains residual covariances obtained after fitting the model to the data. The user would search relatively large residuals which would be seen if the model does not fit the data, but how large the residuals need to be is not known. Originally, there was no fit statistic for the model, however, Gessaroli (personal communication, March 1992) has added a fit index to the program. This program is getting more attention and use, and many researchers who use IRT prefer NOHARM because of its strong theoretical basis. NOHARM can handle large datasets, and is user friendly.

OPTIMAL TEST DESIGN (OTD)

This program, which was written by Verschoor (1991), uses a linear programming algorithm to select items from item banks. The user prepares three input files; item bank file which contains item parameters and other item characteristics, specifications file which contains target information and other constraints, and a third file that contains the names of any item categories (the names are coded as numbers). The target information function is important in the specifications file, and many other constraints, such as number of items from each content or item format, can be imposed on the item selection process. Some of the error messages in the program are not helpful, and there is no option to request the exact number of items needed for test. Improvements can be expected in subsequent releases of the software.

Summary

Studies on test dimensionality, IRT approaches to item selection in criterion-referenced tests, and automated item selection methods, have been reviewed in the preceding sections. It has been seen in the dimensionality studies that IRT unidimensional models are robust to less severe violations of the unidimensionality assumption. But the effects of the mild violations of the assumption on test score validity and reliability were not addressed in any of the dimensionality studies. In the studies on item selection methods, it has been documented that optimal item selection methods provide higher decision consistencies and accuracies than non optimal methods. But these studies mainly used unidimensional tests. What could have happened to decision consistencies and accuracies of the constructed tests if the item banks were not strictly unidimensional has not been studied?

None of the studies addressed the effects of multidimensionality on criterion-referenced tests. The accuracy of mastery/nonmastery decisions based on criterion-referenced tests when the test data are multidimensional and the test model is unidimensional has not been investigated. The performance of optimal item selection methods when the test is multidimensional was not studied. A comprehensive Monte Carlo study in which these situations are examined seems timely. This is the focus of this study, and the methodology will be outlined in Chapter III.

CHAPTER III

METHODOLOGY

Introduction

The procedures followed in this study are based on the assumption that unidimensionality is violated through the presence of minor traits beyond the major trait or ability the test is intended to measure. The situations simulated or investigated reflect cases in which tests are being constructed from item pools. Multidimensionality exists and is being assessed at the item pool level, and its effect on tests developed from the pool will be examined. A common dominant ability underlies the examinee responses on items in the pool, and minor abilities that are specific to particular sections of the test are operating too. In many situations, a test may have a dominant trait and some minor traits. For example, it could be true that reading ability is one minor factor in the examinee performance on a physics test. Another cause of the presence of minor abilities might be the presentation of test items in different formats that require different techniques from the examinees to answer the items. Another possibility is that different sections of a test may require different minor abilities to get correct responses, because the sections usually measure related but different aspects of the same content domain.

When a test is constructed from a multidimensional item pool, the item selection method used may influence or have impact on the dimensionality of the resulting test. If items that tap a specific trait are sampled more than the items tapping other traits, the resulting test may not reflect the item pool in terms of dimensionality. The results might look different if items are selected equally from the different dimensions. Imagine the case where a pool of 100 items has 4 dimensions, each dimension influencing 25 items. If a 20-item test is constructed from this pool by selecting items at random, the dimensionality of the resulting test might be similar to that of the pool, but may not be certainly known. If the 20 items are sampled from the four dimensions proportionally, on the other hand, the resulting test may have dimensionality equivalent to that of the item pool. If all 20 items are chosen from one dimension, the resulting test might be unidimensional. In short, the item selection method may have an impact on the dimensionality of the resulting test when the item pool is not unidimensional, and some item selection methods might work better than others.

This study addressed three issues: (1) Violation of the unidimensionality assumption by the presence of minor traits besides the major ability; (2) test development in situations where items are selected from item pools that are not strictly unidimensional; and (3) the performance of some item selection methods in such situations. The study began with a data simulation in which item pools with different amounts of multidimensionality were simulated. Preliminary analyses were made to assess if the data was being generated as expected. The seeds of the random number generator were changed to see if they have effects on the generated item and ability parameters. The factor structures of the generated data were examined using both linear and nonlinear factor analysis. The generated test data were then calibrated by using the IRT program BILOG (Mislevy & Bock, 1986).

The robustness of the maximum likelihood estimation procedure, as implemented by the widely used computer program LOGIST, to the violation of the unidimensionality assumption has been studied (see, for example, Drasgow & Parsons, 1983; Harrison, 1986; Ansley & Forsyth, 1985; Reckase, Ackerman, & Carlson, 1988). In assessing the robustness of MLE, i.e., LOGIST, researchers often compare the true and estimated item and ability parameters; they assess the estimation accuracy of the program when the data is not strictly unidimensional. They do not, however, examine the model-data fit using residual analysis or some other fit statistics. The goodness-of-fit assessment is an important step for the subsequent analysis of the test data. If an IRT model does not fit the data, the estimation of ability and item parameters might not be accurate, and the conclusions derived from these estimates might be inadequate.

It has been found in several studies that LOGIST is robust to "minor" violations of the unidimensionality assumption, although the model-data fit was not addressed in many of the studies. The robustness of BILOG to the violation of the unidimensionality assumption does not appear to have been studied, and it is hoped that it is not less robust than LOGIST. BILOG provides item and test fit statistics which LOGIST does not provide, and which help in examining the model-data fit. In this study, the model-data fit was insured by examining the fit statistics provided by the program, and by performing residual analysis after fitting IRT models to the data. The estimated parameters were also correlated with their true values to assess how well the parameters in each dataset were estimated.

Short tests were constructed from each generated item pool using each of four methods of item selection. The tests were then analyzed and scored using BILOG (Mislevy & Bock, 1986). The estimated abilities were correlated with the true abilities for each dataset and for each item selection method. The decision accuracies of these tests were compared among item pools, and among item selection methods. Analysis of real data followed. First, the dimensionality of the test data was examined. Second, the data were calibrated with BILOG. Finally, short tests were constructed from the test data by using each of four item selection methods. The measurement precision and decision accuracies of the resulting tests were then compared by item selection method.

PART A: Simulation

Purposes

One purpose of this part of the study was to investigate the effect of item pool dimensionality on ability estimation and decision accuracy. To do so, test data with different degrees of multidimensionality were simulated. Another purpose was to study the influences of item selection methods on decision accuracy, and their interaction with item pool dimensionality.

Data Simulation

A FORTRAN program similar to the IRT program TESTSIM discussed in Chapter II was used for the data simulation. The program is a modified extension of the simulation program DATAGEN (Hambleton & Rovinelli, 1973). It is based on the concept of essential unidimensionality introduced by Stout (1987) and simulates test data with one major dimension and several minor dimensions. It uses a bivariate extension of the two-parameter logistic model which can be written as

$$P_i = \frac{1}{1 + \exp\{-D[a_1(\theta_1-b_1) + a_2(\theta_k-b_2)]\}} \qquad (1)$$

where:

p_i is the probability of answering item i correctly
θ_1 is the dominant ability
θ_k is the kth minor ability
D is an scaling factor equal to 1.7
a_1 is the discrimination of item i in the major dimension
a_2 is the discrimination of item i in the minor dimension
b_1 is the difficulty of item i in the major dimension
b_2 is the difficulty of item i in the minor dimension.

In dimensionality assessment, the guessing parameter usually cause some problems (see, for example, Carroll, 1945; Carroll, 1983; Bock, Gibbons & Muraki, 1988). That might be the reason many dimensionality researchers set the parameter to a constant value, and why many IRT computer programs such as NOHARM, TESTFACT, and TESTSIM constrain it to be constant or treat it differently from the other item parameters. To avoid problems that the c-parameter may have caused in this study, it was set equal to zero.

The data were simulated in a way such that the major trait influenced all items in the pool, and each minor trait affected a cluster of items. Each item was affected by the major trait and one of the minor traits. All minor traits influenced equal numbers of items in the pool, because if any minor trait influenced more items than the other minor traits that minor trait might become more significant than the others. It is not known, however, how many items a minor trait would need to influence in order to become dominant. This issue was not addressed in this study. The number of minor dimensions was set equal to 4; i.e, each item pool was divided into four parts, each part being influenced by the major trait and one of the minor traits.

Strength of minor dimensions. The variation of the potency of each of the minor dimensions could be attained by varying the relative means and variances of the a-parameters of the major and minor traits (Ansley & Forsyth, 1985; Way, Ansley & Forsyth, 1988). In simulating two-dimensional data with one of the traits stronger than the other, Ansley and Forsyth (1985) used a mean of 1.23 and standard deviation of .34 for the discrimination of the dominant ability and a mean of .49 and standard deviation of .11 for the discrimination of the minor ability.

For the same purpose, Stout (1987) introduced an index of lack of unidimensionality which controls the means and variances of the a-parameters in the major and the minor traits. The index, ξ, represents the influence of each minor trait relative to the major trait, and the means and variances of the a-parameter in the major and minor traits could be related as follows:

$$a_1 \sim N((1 - \xi)\mu, \sqrt{1 - \xi}\ \sigma) \tag{2a}$$

$$a_2 \sim N(\xi\mu, \sqrt{\xi}\ \sigma) \tag{2b}$$

$$a_1 + a_2 \sim N(\mu, \sigma) \tag{2c}$$

where a_1 - discrimination parameter for dimension 1 (major)

a_2 - discrimination parameter for dimension 2 (minor)

μ - mean of discrimination parameter for the whole test

σ - standard deviation of the a-parameter for the test

ξ - Strength of minor trait relative to the major trait.

The index ξ varies from 0.0 which means the test is strictly unidimensional to a value of 0.5 which reflects that the minor traits are not less potent than the major trait. A value greater than 0.5 for the index implies that the dominant dimension is not dominant any more; a case that goes beyond the concept of essential unidimensionality. If we choose, for example, a value of .2 for ξ and wish to generate two-dimensional discrimination of mean 1.0 and standard deviation 0.4, we will get a mean of 0.8 for a_1 and a mean of 0.2 for a_2. The standard deviations will be 0.358 and 0.179, respectively. ξ controls the values of the a-parameters for the respective traits, and hence, the potencies of the traits. Nandakumar (1991) studied the effect of the index on the dimensionality of a test and reported that tests might not be essentially unidimensional if the index is set as high as 0.4. In this study, ξ took the same value for all minor traits in each item pool.

In choosing the distributions and descriptive statistics of the ability and item parameters for the simulation process, two strategies were utilized. Real data were analyzed and the resulting descriptive statistics (distributions, means, variances, and ranges) were examined. Secondly, other studies were reviewed and the distributions, means, variances, and ranges of model parameters were examined. The values obtained in the two cases were considered in choosing the means and variances of the ability and item parameters in the data generation process. Two facts were kept in mind: Test scores are more homogeneous in criterion-referenced tests, and most of the reviewed research concerns norm-referenced tests.

Ability. In criterion-referenced tests, the latent trait score distribution is often negatively skewed. In analyzing one national credentialing examination, skewness of -.25, a mean of .094, and a variance of 1.127 were found for the ability distribution. Minimum and maximum values were -4 and +4 respectively because the analyses were made with BILOG which restricts the ability parameters to this range. To simulate ability scores close to these values, a beta distribution with parameters 5 and 3 was used. These parameters will provide a mean of 0.6 and standard deviation of 0.2. The scores were then rescaled to have a mean of zero and variance of 1.

Discrimination. The discrimination parameter is important in dimensionality assessment because it represents the factor loadings in factor analysis. In analyzing test data, Lord (1968) found a range from .4 to 1.7 with a mean of 1.07 and standard deviation of .4. Ree (1979) determined that discrimination usually varies from .5 to 2.5. He used a range between .65 and 1.61 with a mean of .95 and standard deviation of .28. In simulating test data, Swaminathan and Gifford (1983) used a mean of 1.28 for 1000 examinees for an 80 item test. In a simulation study, Hambleton and Cook (1983) used a mean of 1.12. In analyzing a credentialing exam, values lower than the values found in the literature were obtained (mean of .642 and standard deviation of .212).

In this study, two sets of values were used for generating the discrimination parameter: (1) A mean of 1.0 and standard deviation of 0.4 to reflect achievement test data that have a-parameters close to those reported in the above cited research studies, and (2) a mean of 0.6 and standard deviation of 0.2 to reflect a licensure test such as the above mentioned credentialing exam. The intention was not to compare the two types of tests but merely to assess the effect of the presence of minor dimensions on the decision accuracies in both types of tests. The first test (mean of 1.0 and standard deviation of 0.4) will be called Test 1 and the second test (mean of 0.6 and standard deviation of 0.2) will be called Test 2 in the remainder of the study. Each generated a-value will be broken down into two components as will be discussed shortly.

Difficulty. For the difficulty parameter, values obtained in the literature and values obtained in analyzing real data were compared. Lord (1968) reported a range of -1.5 and 2.5 with a mean of .58 and standard deviation of .87. Ree (1979) contended that values typically fall between -3 and +3. Swaminathan and Gifford (1983) used a mean of .15 for 80 items and 1000 examinees. Hambleton and Cook (1983) and Hambleton (1983) used uniformly distributed difficulty in the interval [-2,2]. In one credentialing exam, normally distributed b-values with a mean of -.534 and standard deviation of 1.09 were found. The difficulty parameter is not as critical as the discrimination in dimensionality assessment, and it was deemed that one value for both types of tests would be sufficient. Normally distributed b-values with mean of -.53 and standard deviation of 1 were chosen to be used in simulating both Test 1 and Test 2.

Simulation. When the descriptive statistics for the item and ability parameters were chosen as discussed above for the major trait, the statistics of the parameters for the minor dimensions were calculated by the generating program according to equation 2. Six item pools (three for each test), each consisting of the binary responses of 1000 examinees on 200 items, were generated as follows:

1) Five independent ability scores were generated from a negatively skewed beta distribution for each examinee, corresponding to the major and four minor abilities. The ability scores were rescaled to have zero means and unit variances.
2) Two b-values and one a-value were generated from a normal distribution with the above discussed means and variances for each of the 200 items and for Test 1 and Test 2. The few a-values that turned out to be less than zero were set to zero, and the b's were independent.
3) The value of ξ (see equation 2) was chosen as 0, .3, or .5 for the 3 item pools for each type of test.
4) By equation 2, the magnitudes of the a-parameters in the major and minor traits for each item were controlled. The a-value generated for each item was broken down into

two components, one for the major trait and one for the minor trait. If a is high, both a_1 and a_2 will be relatively high.

5) Each minor trait was affecting 50 items. The first minor trait was affecting the first 50 items, the second minor trait was affecting the next 50 items, and so on.

6) The probability of getting an item correct by an examinee with certain abilities was obtained by equation 1.

7) Uniform random numbers in the interval [0,1] were generated for each item and compared with the probability of each examinee getting each item right. If the probability was less than the random number, the examinee is scored 0 for that item, and 1 otherwise.

The descriptive statistics of the item and ability parameters used to generate the data are highlighted in Table 1. This process resulted in a 1000x200 matrix of binary responses for each of the six datasets.

Table 1

Description of the Parameters
Used to Simulate the Data

Test	Statistics	θ	b's	$a_1 + a_2$
1	Mean	0.0	-.53	1.0
	Std. Dev.	1.0	1.00	0.4
2	Mean	0.0	-.53	0.6
	Std. Dev.	1.0	-.53	0.2

Assessing the program. It was deemed necessary to insure that the program was generating the expected data. To obtain a thorough analysis with reasonable variables, the number of examinees and the number of items were reduced to 500 and 40, respectively. Two extra levels of ξ were also included at this stage; 0.2 and 0.4. The performance of the data generating program was examined in three analyses. The seeds of the random number generator were changed, and the descriptive statistics of the generated item and ability parameters were examined. This analysis was intended to probe whether the starting values of the random numbers had effects on the generated data.

Second, linear factor analyses were performed on the five datasets (with minor dimension strengths of 0, .2, .3, .4 and .5) with one to five factor solutions. The eigenvalues of the matrices consisting of the tetrachoric correlations of the binary data, and the variances explained by each factor were compared among the datasets. This was expected to highlight if the generated datasets had different factor structures. Finally, nonlinear factor analysis, using the program NOHARM (Fraser & McDonald, 1988), was undertaken. A unidimensional solution was fitted to each dataset, and the results provided for the five datasets were examined. The sum of the squares of residuals and the percent of standardized residuals in the variance covariance matrix greater than 1.96 were compared among the datasets. This was expected to highlight whether different results would be obtained when an IRT model is fitted to the five datasets with the different dimensionality structures. These three analyses were used to probe whether the simulation program was working as expected.

IRT analysis. After satisfactory results were obtained from the data generation step, an IRT analysis was performed in each dataset. The datasets were six; two tests (Test 1 & Test 2) with three levels of minor dimension strengths (0.0, 0.3, & 0.5). Each dataset for each type of test had a different dimensionality structure as determined by the relative strengths of the dominant and minor traits. The 2-parameter model of the BILOG program was used to calibrate the item and ability parameters.

It was not possible to calibrate 200 items in one run. So the items in each dataset were grouped into three 90-item sets with overlapping items, and each group was calibrated separately. To justify the calibration of the data in three sections, the equivalence of the parameter estimates of the common items were assessed by a) plotting the two sets of values against each other, and b) using linear regression analysis with the two estimates. It was expected that the values would almost be the same to justify the calibration of each dataset in three parts with BILOG. If they were not, the presence of common items provided a basis for statistical adjustments (i.e. equating). The data-model fit was assessed in two ways: 1) by looking at the item and test fit statistics provided by the program, and 2) by carrying out a residual analysis of the item and ability parameters provided by the program using the RESID computer program (Hambleton & Murray,1983). The data-model fit was necessary for the rest of the data analyses, and if it were not attained for any dataset, another dataset that fit the model would have been generated.

The estimated ability, difficulty, and discrimination parameters were correlated with their true values. The purpose was to examine if the strength of the minor dimensions affected the estimation of the parameters, especially the estimation of the ability scores. The item parameters and examinee true ability scores were kept for further use. The dominant true ability scores (uncontaminated by the minor factors) were used as a criterion. The intended use of the simulated item pools and constructed tests was assumed to be classification of examinees along one ability; the major ability. The item parameters were used to create item banks from which items were later selected, and which of the four minor dimensions influenced each item was also shown in the banks.

Variables

Degree of lack of unidimensionality. This variable which reflects the factor structure of the item pools was varied to three levels in the main analyses of the study, and up to five levels during the evaluation of the data generation process. These levels stand for the influences of the minor traits relative to the dominant trait. In the main analyses, three tests of 200 items each with different factor structures in terms of dimensionality were generated for Test 1 and Test 2. Building on studies by Nandakumar (1991) and Stout (1987), the relative influences of the major and minor abilities for the item pools were 0, 0.3, and 0.5. These values were chosen to vary from data that had no minor dimensions to a test data that had relatively strong minor dimensions (equal values for a_1 and a_2).

Item selection. In developing tests by selecting items from each item pool that corresponded to the six datasets (2 types of tests and 3 levels of ξ) four item selection methods were used:

1. Random method: items were chosen from the item banks at random, and the item statistics were not used. This method is usually used in situations where item statistics are not available, or items are considered to be equally useful.
2. Optimal method: items that provide the most information at the cut-off score were selected. The other item properties; that is, which minor factor influences each item were not considered in this method, and were not considered to be important in the resulting test. This method focuses on the measurement precision near the cut-off score.
3. Optimal-balanced method: items that provide the most information at the cut-off score were selected, and the items influenced by the four minor factors were equally

represented in the resulting test. This method insures the content validity, and measurement precision near the cut-off score of the resulting test. It has been recommended for practical test development (see, for example, Hambleton, Dirir, & Lam, 1992).

4. Optimal-unbalanced method: test items were selected to provide maximum information at the cut-off score, and the number of items from each of the four minor factors that were included in the resulting test was not balanced. Approximately 63 percent of the selected items were from one of the minor traits, and the rest were equally distributed among the other minor traits. This approach reflects cases in which most of the items in an item bank load on one trait, and cases where most of the selected items tap a single dimension. The Optimal Test Design computer software (Verschoor, 1991) was used to select items in methods 2, 3, and 4.

Test length. Using each method of item selection, a test of 40 items was constructed from each of the six item banks (200 items in each bank). This test length is typical of many tests. Also, shorter tests may result in inaccurate parameter estimation, and longer tests might not easily be handled with the available computer facilities. As mentioned earlier, the number of minor traits was four, and each was influencing 10 items for item selection with method 3 (optimal-balanced), 63 percent or 13 percent of the items for method 4 (optimal-unbalanced), and any number of items for methods 1 and 2.

Number of examinees. In all simulated test data and constructed tests, the number of examinees was fixed at 1000. This was also chosen having in mind the accuracy of the item and ability parameter estimation and the available computer capabilities.

Cut-off score. Two arbitrary cut-off scores on the ability scale were used: (1) A point along the ability score where around 75 percent of the examinees passed the tests (i.e, $\theta=-.685$), and (2) the mean of the ability distribution; i.e, 0.0. Figure 1 shows the simulated examinee ability distribution, and the location of the two cut-off scores. The first cut-off score represents tests with high pass rates, and the latter was chosen to reflect tests with passing scores at the middle of the ability distribution, and with comparable numbers of failures and passers.

Evaluation

From each of the six item pools (three for Test 1 and three for Test 2), four tests were constructed using each of the four item selection methods. For the 24 tests constructed, the BILOG program was used to obtain the examinee ability scores. The scores were correlated with the true dominant ability scores of the examinees. The pass/fail decisions for each of the 1000 examinees in each test were compared with the pass/fail decisions based on the

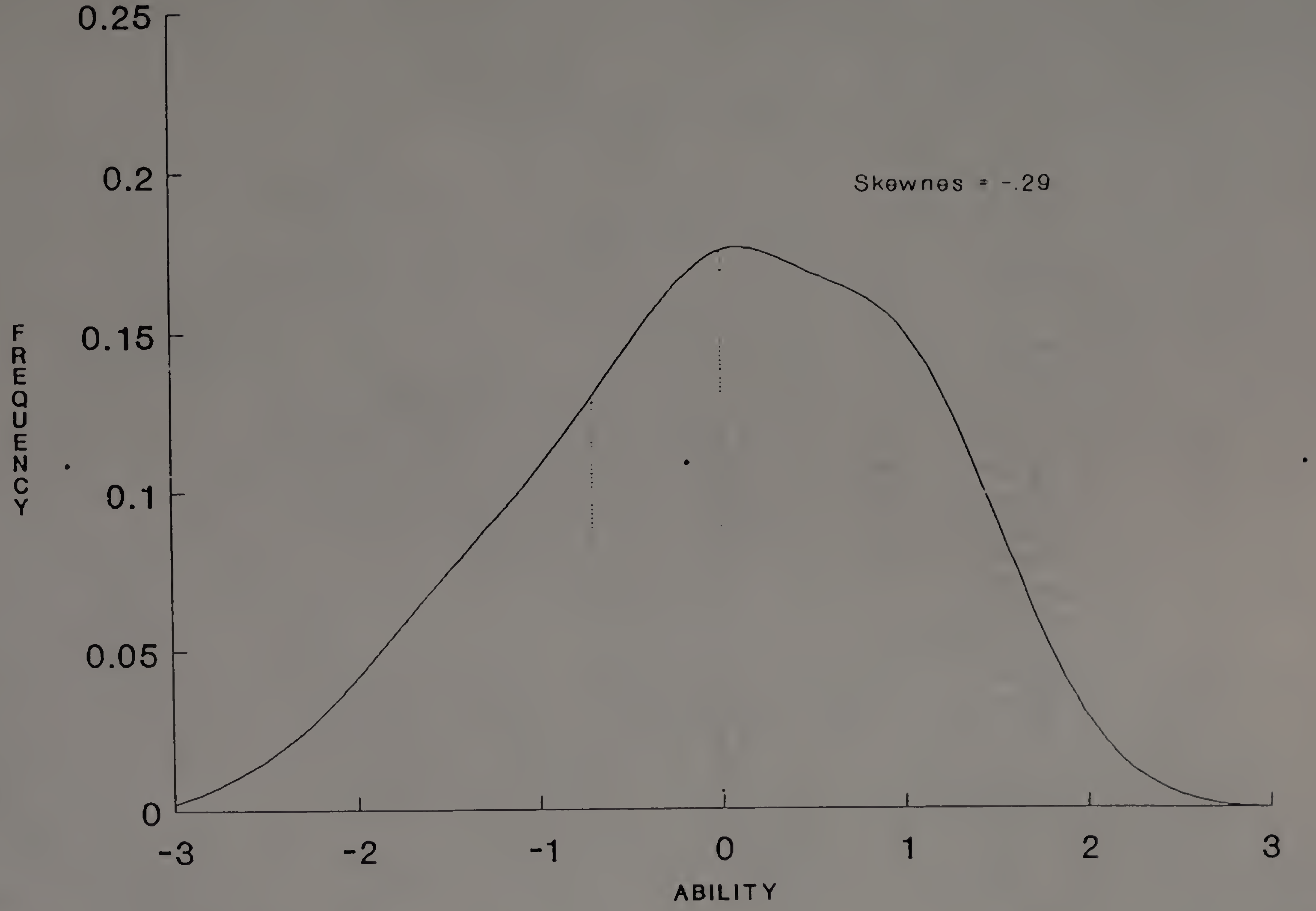


Figure 1. Example of Simulated ability distribution and the Two Cut-off Scores

criterion scores (dominant ability) to obtain the decision accuracies of the tests at both cut-off scores. At a cut-off score of 0.0, five replications were made in generating each of the six datasets, constructing each of the 24 tests, obtaining the examinee scores on each test, computing the correlation coefficients of the abilities, and computing the resulting decision accuracies for each test.

The mean correlation coefficients, and the mean decision accuracies for the 24 cases were then obtained, and compared. Analysis of variances (ANOVA) were conducted, separately for Test 1 and Test 2, to assess if the correlation coefficients of the abilities were different among the item selection methods, whether the coefficients were different among the datasets with the different degrees of lack of unidimensionality, and whether there was interaction between the two effects. Before undertaking the analysis of variance, the correlation coefficients were transformed by using Fisher's r to z transformation which can be written as:

$$z = \tfrac{1}{2}\ln\left(\frac{1 + r}{1 - r}\right) \qquad (3)$$

Another set of analysis of variances were made, again, separately for the two tests, to examine whether the decision accuracies among the dimensionality structures were significantly different from one another, whether the decision accuracies among the item selection methods were significantly different from one another, and if the two effects had interactions in influencing decision accuracy. The proportion agreement was also transformed prior to analysis by using the equation

$$x = \mathrm{Sin}^{-1} \sqrt{p} \tag{4}$$

where x is the transformed decision agreement, and p is the decision agreement of the test and criterion.

Part B: Real Data

Purposes

The general purpose in including real data in the study was to examine if similar results would be found in real and simulated data. The second purpose of this part of the study was to examine if content categories in a particular credentialing exam represented different traits. A third purpose was to examine the performance of item selection methods in affecting the decision accuracies of short tests developed from a credentialing exam.

Data

Candidate item response data from one of the national credentialing examinations were available for use in this part of the research. The exam, which was administered in 1988 to 3965 candidates, consists of 250 items. Twenty items were not included in the analysis because of low biserial correlations (less than 0.2). The test has six content categories, ten item formats, and three categories of cognitive levels. The 230 items were divided into the following content categories:

Content Category	Number of Items
1	27 (11.7%)
2	78 (33.9%)
3	31 (13.5%)
4	30 (13.0%)
5	27 (11.7%)
6	37 (16.1%)

Procedures

The variables in this part were essentially the same as those in Part A. There were differences between the item selection methods compared in this part and in the previous part. In this part, an optimal method, a content-optimal method, a content-random method, and a classical method were compared. The optimal method and content-optimal method were parallel to the previously defined optimal and optimal-balanced methods except that the content categories of the item pool were balanced in developing the 40-item tests. In the classical method, the content specifications were considered while items with high biserial correlations and moderate difficulty (between 0.3 and 0.8) were selected. In the content-random method, the content was also balanced in the resulting test, but the items were selected from each category at random. The optimal-unbalanced method used in the simulation part seemed unimportant since it would not be much different from the other optimal methods used with unidimensional data. The cut-off score of the exam, which was 70%, was used for item selection and decision accuracy computations. Due to computer limitations, the number of examinees was reduced to 2000.

Dimensionality investigation. In dimensionality assessment, the content categories were considered as possible causes of lack of unidimensionality, or to put it in another way, the content categories were treated as minor dimensions in the test while one major trait was being measured by the examination.

In order to get an idea about the factor structure of the data, linear and non-linear factor analysis were undertaken. For the purpose of these analyses, 40 items that represented the 230 items in terms of percent of items from each content category were selected. The tetrachoric correlations of the items were factor analyzed, then the eigenvalues of the correlation matrix were examined. The difference between the first and the second eigenvalues were compared to the difference between the second and the third. If the difference of the differences is large, this implies that the test data are unidimensional. The differences of the successive eigenvalues were also examined, as well as the magnitudes of all eigenvalues.

In the non-linear factor analysis, the NOHARM program (Fraser & McDonald, 1988) was used to fit the normal ogive model to the data. One-, two-, three-, and four-dimensional solutions were investigated in the binary data to provide some additional clues about the dimensionality structure of the test data.

IRT calibration. The next analysis of the data was IRT item and examinee calibration. The BILOG program (Mislevy & Bock, 1986) was used to analyze the data, and provide item and ability parameters. The one-, two-, and three-parameter logistic models were fitted to a representative sample of 65 items. The purpose was to examine if any of the IRT models fits the data, and which model provided the best fit. The item parameter estimates from the IRT analysis were kept in an item bank together with content information for further use.

The next step was dividing the test into two equal parts. One part was used to provide a criterion measure, and the other part served as an item pool from which items were selected later in the test construction process. The odd-numbered items between 1 and 200 and items 201-230 of the data were placed in the item bank, and the even items between 1 and 200 were used as the criterion. The choice of the odd and even items of the test as item bank and criterion, respectively, was arbitrary, and the last 30 items were added to the bank to create a larger pool.

Item selection. The four item selection methods compared in the real data were: (1) Optimal in which items provide most information at the cut-off score were selected; (2) content-optimal in which items provide most information at the cut-off score were selected and the content balance of the resulting test was considered; (3) classical in which test content was balanced while items with high r's and moderate p's were selected; and (4) content-random in which items were selected from the different content categories to reflect the content specifications of the test, but item statistics were not used. Content categories replaced the minor traits in the simulated data, and the OTD program was used to select items from the pool (odd items) in the optimal methods.

Test length. Like the simulated data, test length was set at 40 items. Each method of item selection was used to select 40 items from the item bank.

Evaluation

For the four tests developed, the information functions in the ability range -3 to +3 were computed. The errors of measurement at a selected range near the cut-off score were also computed by using the relationship

$$SE(\theta) = (I(\theta))^{-\frac{1}{2}} \qquad (5)$$

where SE(θ) is the standard error of ability estimates at θ, I(θ) is the test information at ability θ.

The percent of pass/fail decision agreements between each test and criterion was calculated. These percents were compared for the four methods of item selection. The improvement in decision accuracy by the item selection methods over a baseline decision accuracy level was also examined. The content-random item selection method was chosen as the baseline.

CHAPTER IV

RESULTS

Part A: Simulated Data

The Performance of the Computer Software

Before any analyses were made, the accuracy of the computer software was examined. The seeds of the random number generator used to simulate item and ability parameters were changed five times. For this purpose, the abilities and binary scores of 500 examinees on 40 items were simulated. The means and standard deviations of the generated parameters were examined, and compared with their true values (i.e, the means and standard deviations chosen to generate the parameters). In all five runs, the true and simulated means and variances for the ability and item parameters were almost identical. This is an indication that the program was performing as expected, and changing the seeds values of the random number generator had only a small random effect on the performance of the software.

Another investigation on the performance of the software regarding the factor structure of the generated datasets was conducted. For five datasets (5 levels of ξ) generated using the parameters for Test 1 (achievement tests), the binary responses of the 500 examinees on the 40 items were factor analyzed using linear factor analysis. The eigenvalues of the first five factors, and the variance explained by each factor are shown in Table 2. The

Table 2

First Five Eigenvalues and Variances Explained by the Factors for the Five Datasets
(N=500, n=40)

		Level of ξ				
Factor		0.0	0.2	0.3	0.4	0.5
1	λ	10.44	7.67	6.28	4.0	4.2
	%	26.1	19.2	15.7	10.0	10.6
2	λ	1.46	.86	.87	2.3	1.4
	%	3.7	2.1	2.2	5.7	3.4
3	λ	.47	.53	.65	.90	1.0
	%	1.2	1.3	1.6	2.2	2.5
4	λ	.44	.52	.55	.70	.84
	%	1.1	1.3	1.4	1.8	2.1
5	λ	.37	.50	.49	.58	.48
	%	.9	1.3	1.2	1.4	1.2

eigenvalue for the first factor and the variance explained by the first factor decreased as ξ increased. The difference between the first and second factor also decreased with an increase in ξ (one slight exception is at 0.4 and 0.5). The factor loadings were also examined for one, two, three, four, and five factor solutions.

The factor loadings changed as the strength of the minor dimensions (ξ) in the data changed. Almost all items highly loaded on the first factor at ξ of 0.0 in all factor solutions, and items loaded and were divided among the factors as expected at ξ of 0.5. In the latter case, the first 10 items highly loaded on one factor, the next 10 items on another factor, the third 10 items on a different factor, and the last 10 items on a different factor when a four factor solution was requested. These results support that the software was generating datasets with the expected factor structures.

Another step was taken to ensure that the generated data had the expected dimensionality structures. This time, non-linear factor analysis was performed using the IRT program, NOHARM (Fraser & McDonald, 1988). The sum of squared residuals of the variance-covariance matrix after fitting each dataset to a unidimensional solution was examined. The percent of the standardized residuals of the variance-covariance matrix that were greater than 1.96 was also examined for each dataset. Both indices were expected to increase as the potency of the minor dimensions increases. The sum of squared residuals (SSR) and the percent of standardized residuals (PERZ) greater than 1.96 (expected to be not more than 0.05 if the data fits a unidimensional model) for each of the five datasets are shown in Table 3. Multidimensional data were also generated to highlight how large these two indices could be when a

Table 3

Sum of Squared Residuals and Percent of Standardized Residuals Greater Than 1.96 (N = 500, n = 40)

Data	ξ	SSR	PERZ
1	0.0	0.032	0.010
2	0.2	0.041	0.015
3	0.3	0.050	0.031
4	0.4	0.079	0.068
5	0.5	0.121	0.117
6	MD*	0.339	0.295

*Four-dimensional data

unidimensional solution is fitted to a four-dimensional data. It can be seen that both indices increased systematically as the potency of the minor dimensions increased. The intention was not to determine whether each dataset was unidimensional, but merely to show that the misfit statistics are in the expected order and highlight the departure of datasets from unidimensionality as ξ is increased.

IRT Analysis

After satisfactory results were observed in examining the performance of the software, three datasets were generated for each type of test (Test 1 and Test 2). Responses for 1000 examinees on 200 items were simulated as discussed in Chapter III. IRT data calibration followed to obtain item and ability parameter estimates. The main purpose of the data calibration was to develop item banks for the six datasets (three for Test 1 and three for Test 2). A secondary purpose was to examine if the two-parameter logistic model could adequately fit the two-parameter data which were generated. Hence the next step was to analyze the binary datasets using the two-parameter IRT logistic model.

The IRT program BILOG (Mislevy & Bock, 1986) was used for this purpose. A whole dataset (200 items and 1000 examinees) could not be handled in one run or even two runs with the available computer facility. It was found that more than 90 items could not be calibrated in one run because of computer memory limitations. Hence each set of data was divided into three sets of 90 items with overlapping items. The sets were items 1 to 90, items 61 to 150, and items 111 to 200. In that arrangement, sets 1 and 2 had 30-item overlap, and sets 2 and 3 had a 40-item overlap. Three separate analyses of 90 items each were performed for each dataset.

An invariance analysis was undertaken in which the b-values of the common items were plotted against each other. The plot of 40 b-values obtained from calibrating the items separately and calibrating them with 90 items is shown in Figure 2. As can be seen from the figure, the plot is almost a straight line; an indication that the values are almost the same. Second, a regression analysis on the

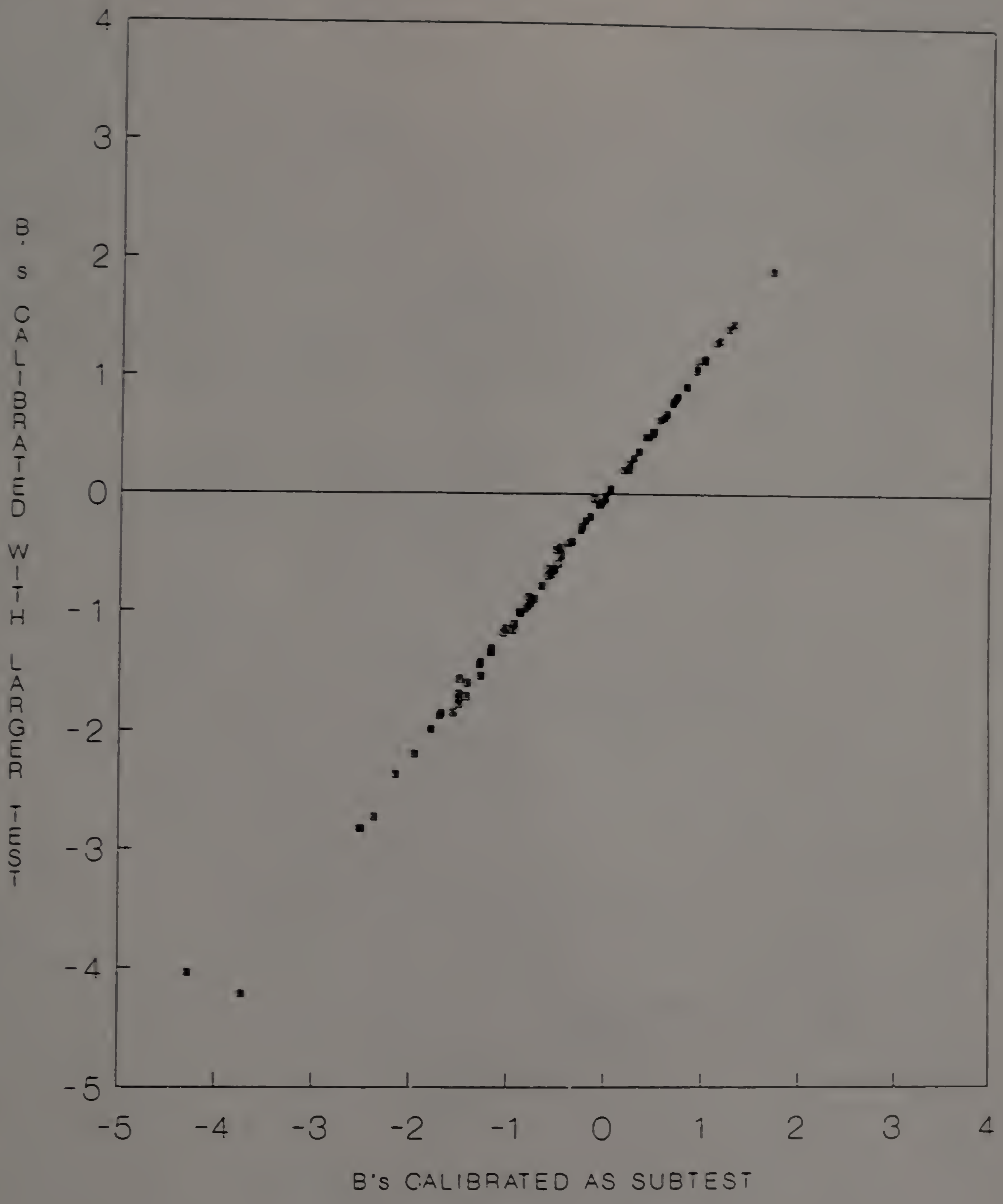


Figure 2. Plot of Difficulty Values for 40 Items Calibrated at Two Runs

discrimination and difficulty parameters was conducted. An intercept of 0.011 and an slope of 0.995 were found for the difficulty parameter, and the correlation between the two b-values (estimated in the two analyses) was .995. For the discrimination parameter, an intercept of -0.006 and an slope of 0.992 were found, while the correlation coefficient of the a-values was 0.958. These results indicate that values obtained for the item parameters in the two analyses were quite close, and hence support that the data could be run in separate sections. Then, the IRT data analyses proceeded, calibrating each dataset three times. Three item pools for Test 1 and three item pools for Test 2 were formed; one for each dataset. For the items with multiple parameter estimates, the average of each parameter was taken and used in the item banking process.

The goodness-of-fit of the data was assessed by computing the residuals using the computer program RESID (Hambleton & Murray, 1983). For each level of ξ, 67 items were sampled from the bank of 200 items. The items were selected so that each set of 50 items that might be affected by a particular factor were equally represented in the selected set. The two-parameter logistic model was fitted to each of the 67-item sets. The resulting standardized residuals provided by the IRT program RESID (Hambleton & Murray, 1983) are shown in Table 4. The last column contains the expected normal distribution of standardized

Table 4

Distribution of Standardized Residuals From Fitting the Two-Parameter Logistic Model to a Sample of 67 Items for the Three Levels of ξ[1]

Standardized Residuals	Level of ξ 0.0	0.3	0.5	Normal
< -3	0.14%	0.27%	0.15%	0.2%
-3 to -2	1.90%	2.04%	2.09%	2.1%
-2 to -1	10.99%	16.28%	17.91%	13.6%
-1 to 0	36.91%	29.58%	30.90%	34.1%
0 to 1	35.82%	36.64%	28.66%	34.1%
1 to 2	12.62%	13.16%	16.87%	13.6%
2 to 3	1.63%	2.04%	3.43%	2.1%
> 3	0.00%	0.00%	0.00%	0.2%
AASR[2]	0.747	0.829	0.919	0.790

[1] The number of residuals was 804.

[2] Average of the Absolute-valued Standardized Residuals.

residuals under the null hypothesis (see, Hambleton, Swaminathan, & Rogers, 1991). The fit was reasonably good at the first two levels of ξ, and was not adequate at the last level. At ξ of 0.5, the fit was relatively poor.

Correlations of Ability and Item Parameters

The parameter estimates for each dataset were correlated with their true values. The goal was to examine how the strength of the minor dimensions influence the IRT parameter estimation, and to probe how close the estimates would be to their true values at each level of ξ. Table 5 shows the correlation coefficients of the true and estimated parameters. The correlation coefficients of all parameters used in unidimensional cases (θ, a, b) with their true

Table 5

Correlations of Estimated and True Parameters

ξ	$r_{\hat{\theta}\theta}$	$r_{\hat{a}a1}$	$r_{\hat{a}a2}$	$r_{\hat{b}b1}$	$r_{\hat{b}b2}$
0.0	.986	.982	----	.988	-----
0.2	.960	.974	.974	.733	-.017
0.3	.927	.974	.974	.616	.075
0.4	.871	.969	.969	.519	.185
0.5	.781	.946	.946	.397	.311

values (θ, a_1, b_1, a_2, b_2) decreased as the strength of the minor dimensions increased. The most substantial decrease was observed for the difficulty parameter (decrement of .591 from dataset 1 to dataset 5). This decrease is very high, and not even close to what is reported in other dimensionality studies. The correlation coefficients of the discrimination parameters decreased but not as much as the other parameters. They decreased from .982 in data 1 to .946 in data 5. The ability parameter, which is more important than the other parameters for the purpose of this study, had decreased significantly as the index ξ increased from 0.0 to 0.5. It had decreased from .986 at ξ=0.0 to .781 at ξ=0.5.

The correlation between the second discrimination parameter a_2 with the estimated a-values was always equal to the correlation between a_1 and estimated a-value (it does not exist at ξ=0.0), and that was expected because a_1 and a_2

were generated to be proportional. The difficulty values were generated randomly and unrelated to a common difficulty, and that is why the correlation between b_2 and b is very different from the correlation between a_2 and a. The correlation coefficients in the table indicated that the estimation procedure, which was based on unidimensional model, became less accurate as the minor dimensions became stronger. It was also apparent that the accuracy of the ability parameter estimation decreased as the multidimensionality of the data increased.

Item Selection and Ability Estimation

The four item selection methods discussed earlier (optimal, optimal-balanced, optimal-unbalanced, and random) were used to select items from the item pools in constructing 40-item tests. The tests were then calibrated with BILOG, and the estimated abilities were correlated with the true dominant abilities. Five replications were made in generating each dataset, developing item pools, constructing tests, calibrating the test with BILOG, and computing the correlation coefficients and decision accuracies. The number of replications were limited because of the high computer costs and limitations of the computer space. The replications were made by using as cut-off point at an ability score of 0.0 (which is the mean of the ability distribution and at which roughly 50 percent of the examinees pass the test) in selecting items and computing decision accuracies.

In Table 6, the mean correlation coefficients of the estimated and true ability parameters, displayed by item selection method, are shown for Test 1 and Test 2. These coefficients are the correlations between true and estimated scores, and hence the average validity indices of the tests developed by each item selection method. The terms correlation coefficient and validity index will be used interchangeably in the rest of the study. As in Table 5, the coefficients decreased as the strength of the minor dimensions increased for each item selection method, and for both types of tests. The decrease was systematic for all item selection methods as the dimensionality increased, but differed among the item selection methods. Test 1 seemed to have higher indices in all cases.

Table 6

Mean Correlations Between
Estimated and True Abilities
(number of replications = 5)

Item Selection Method	Level of ξ					
	Test 1			Test 2		
	0.0	0.3	0.5	0.0	0.3	0.5
Optimal	.970	.935	.830	.962	.916	.812
Balanced	.968	.937	.846	.961	.917	.823
Unbalanced	.969	.921	.770	.957	.902	.763
Random	.969	.925	.806	.941	.878	.763

The coefficients or validity indices dropped more in optimal-unbalanced method (a decrease of .199 in Test 1 and a decrease of .194 in Test 2 from ξ of 0.0 to ξ of 0.5). The second highest decrease was for the random method (a decrease of .163 in Test 1 and a decrease of .178 in Test 2 from ξ of 0.0 to ξ of 0.5). The smallest decreases of the indices were at the optimal-balanced method; 0.122 and .138 in Test 1 and Test 2, respectively. The correlation coefficients were close in all item selection methods when the data was strictly unidimensional; the largest differences were 0.002 in Test 1 and 0.021 in Test 2. However, as the potency of the minor dimensions increased, the differences among the coefficients for the item selection methods increased, and it was highest at ξ of 0.5.

Analysis of variance was conducted to assess whether the correlation coefficients were significantly different from one another among the levels of ξ, whether they were significantly different from one another among the item selection methods, and whether there was an interaction effect between item selection and strength of minor factors. The coefficients were transformed into z-scores using Fisher's z to r transformation as mentioned earlier in Chapter III (see equation 3). The ANOVA tables for the results in Test 1 and Test 2 are shown in Table 7. The main effects and their interactions were all significant, and led to a rejection of the null hypotheses of no differences among levels of ξ and among item selection methods. The result indicates that the strength of the minor dimensions

Table 7

Analysis of Variance of the Validity Coefficients

Test	Source of Variation	SS	df	MS	F	Sign.
1	Strength	8.68	2	4.34	2431.6	<.01
	Item Selection	.12	3	.036	19.9	<.01
	Interaction	.08	6	.014	7.60	<.01
2	Strength	6.75	2	3.38	3289.4	<.01
	Item Selection	.36	3	.119	115.9	<.01
	Interaction	.04	6	.006	5.98	<.01

in an item pool, and the choice of the item selection method in test development have effects on the ability estimation in the resulting tests. It also indicates that some item selection methods work better than others when test data are not strictly unidimensional.

Decision Accuracy

The decision accuracy for the 40-item tests constructed using each of the four item selection methods was computed in each item pool and in each of the five replications. The summary statistics of the decision accuracies for Test 1 are shown in Table 8. Obviously, the decision accuracies decreased as the value of ξ increased, and that was a common trend to all item selection methods. The decrease ranged from 10.7 percent in the optimal-balanced method to 16.1 percent in the optimal-unbalanced method.

Table 8

Summary Statistics of the Decision Accuracies for Test 1

		Item Selection Method			
ξ	Statistics	optimal	balanced	unbalanced	random
	Mean	93.7%	93.7%	94.3%	93.1%
0.0	Std. Dev.	0.90	1.38	0.66	0.39
	Range	2.4	3.5	1.6	0.9
	Mean	89.8%	90.2%	88.5%	88.5%
0.3	Std. Dev.	1.34	1.09	1.36	1.21
	Range	3.4	2.7	3.8	3.0
	Mean	81.9%	83.0%	78.2%	80.5%
0.5	Std. Dev.	1.40	1.02	1.40	1.32
	Range	3.8	2.5	3.7	2.9

The differences in decision accuracies among the item selection methods were very small when the item pool was strictly unidimensional. The optimal-unbalanced method provided a decision accuracy 0.6 percent higher than the other optimal methods and the random method provided a decision accuracy 1.2 percent less than the optimal-unbalanced method. The differences were largest when the minor dimensions were as strong as the major dimension; that is, when ξ was 0.5. At that level, the optimal-balanced method provided the highest decision accuracy (83 percent), the optimal-unbalanced method provided the lowest (78.2 percent), and the random method provided the second lowest (80.5 percent). At ξ of 0.5 the differences in decision accuracies among the item selection methods was larger than when ξ is 0.0. The decision accuracy in the optimal-balanced method is 4.8 percent higher than that of the optimal-unbalanced.

Similarly, the summary statistics of the decision accuracies in Test 2 are shown in Table 9. The decision accuracies was all lower than those obtained in Test 1, but the same trend of decrements were seen as ξ increased. The largest drops in decision accuracy were seen in the optimal-unbalanced and the optimal methods; 14.8 percent and 13.6 percent, respectively. In the random method, the drop was 12.9 percent, and the smallest drop was seen in the optimal-balanced (12.3 percent). There were slight differences between the results reported in Tables 7 and 8. At ξ of 0.0, for example, the optimal-unbalanced method had the highest decision accuracy in Table 8 while the other optimal methods had higher decision accuracies in Table 9. Also, the random method had higher decision accuracy than the optimal-unbalanced method at ξ of 0.5 in Table 8, while the decision accuracies of the two methods are comparable in Table 9.

The results of the analysis of variance undertaken to test the effects of ξ and item selection method on decision accuracy for Test 1 and Test 2 are reported in Table 10. The computed proportion agreement statistics (i.e, decision accuracy) were transformed as discussed in Chapter III; taking the arcsin of the square root of the proportion. Clearly, both variables had significant effects on decision

Table 9

Summary Statistics of the
Decision Accuracies for Test 2

		Item Selection Method			
ξ	Statistics	optimal	balanced	unbalanced	random
	Mean	93.0%	92.8%	92.4%	90.3%
0.0	Std. Dev.	1.12	0.62	0.61	0.74
	Range	2.6	1.5	1.5	1.8
	Mean	87.2%	87.3%	86.9%	84.3%
0.3	Std. Dev.	1.18	0.96	0.83	1.15
	Range	3.1	2.2	2.0	2.8
	Mean	79.4%	80.5%	77.6%	77.4%
0.5	Std. Dev.	1.30	1.00	1.30	1.38
	Range	3.2	2.6	3.3	3.0

Table 10

Analysis of Variances for the Decision Accuracy

Test	Source of Variation	SS	df	MS	F	Sign.
	Strength	16.05	2	8.02	457.3	<.01
1	Item Selection	.24	3	.079	4.49	<.01
	Interaction	.25	6	.041	2.36	<.05
	Strength	13.49	2	6.75	794.2	<.01
2	Item Selection	.61	3	.202	22.4	<.01
	Interaction	.12	6	.019	2.15	.065

accuracy. For the interaction of the effects, it was significant for Test 1 at 0.05 level but came short in Test 2 (0.065). However, since the number of replications were small, one may argue that the latter interaction could have been significant (0.05) had the sample size (number of replications) been increased.

Effects at a Lower Cut-off Score

The cut-off score used to construct tests and compute decision accuracies was lowered from 0.0 to -0.685 along the ability scale where approximately 75 percent of the examinees passed the test. That is typical of many mastery tests where high percent of the examinees pass the test, and where the middle of the ability distribution is higher than the cut-off score. The goal was to examine the effects of minor factor strength and item selection method on decision accuracy and ability estimation in a such situation. No replications were made at this time, and the decision accuracies and correlation coefficients for Test 1 are reported in Table 11. The indices are all higher than the corresponding indices for Test 1 in Tables 5 and 8 in all item selection methods and at all levels of ξ. But that is not unexpected since more classification errors are prone to be made at an ability level located at the middle of the ability distribution than at ability level where fewer examinees are located.

In Table 11, at the lowest level of ξ, the optimal methods provided almost the same decision accuracies, and the random method provided a decision accuracy less than those of the optimal methods by more than 1 percent. For the correlation coefficients, the optimal and the optimal-unbalanced methods provided indices higher than the other

Table 11

Accuracy and Correlation at a Lower Cut-off Score for Test 1 (θ_c=-.685)

ξ		Item Selection Method			
		Optimal	Balanced	Unbalanced	Random
0.0	accuracy	95.8%	95.6%	95.4%	94.3%
	correlation	.974	.969	.974	.969
0.3	accuracy	92.3%	92.0%	91.5%	91.1%
	correlation	.938	.941	.931	.929
0.5	accuracy	88.5%	88.7%	84.9%	85.2%
	correlation	.838	.849	.780	.797

two methods. At the highest level of ξ, the decision accuracies and the correlation coefficients were ranked among the item selection methods in a descending order as: optimal-balanced, optimal, random, and optimal-unbalanced. This trend was seen in Tables 5 and 7 when the cut-off score was 0.0 and the five replications were made. The differences in decision accuracy among the item selection methods at the smallest ξ was lower than when the minor dimensions were stronger. At ξ of 0.0, the largest difference was 1.5 percent, and at ξ of 0.5 the largest difference was 3.5 percent.

As ξ went from the lowest to the highest levels, the decision accuracy and the correlation coefficients dropped for each item selection method. The item selection methods, however, differed in the amount of drop of these indices. The optimal-unbalanced method resulted in the largest decrease of 10.5, while the optimal-balanced method resulted in the smallest decrease of 6.9. This trend was also similar to that reported in Table 8.

The same analyses were repeated for Test 2 (credentialing type) using an ability level of -0.685 as cut-off score. The results of these analyses are in Table 12, and are quite close to those found in Tables 5 and 8. A little difference between the two results was that the decision accuracy of the optimal method was not as high as those of the other optimal methods at the lowest level of ξ. Another difference was that the decision accuracy provided by the optimal method at the highest ξ was 0.7 percent higher than that provided by the optimal-balanced method which was providing the best decision accuracies in all other analyses. Apart from these two cases, the results in Table 12 are equivalent to those in Table 9.

Discussion of Part A

In this section, the results found in analyzing the simulated data will be discussed. First, the results in examining the dimensionality structure of the generated data, and the IRT analysis of the data will be reviewed. Second, the findings for Test 1 will be discussed, followed by the findings for Test 2. The results obtained when the cut-off score was lowered will be discussed at the end.

Table 12

Accuracy and Correlation at a Lower Cut-off Score for Test 2 (θ_c=-.685)

		Item Selection Method			
ξ		optimal	balanced	unbalanced	random
0.0	accuracy	93.0%	93.9%	94.0%	92.5%
	correlation	.963	.962	.961	.946
0.3	accuracy	92.2%	91.3%	91.0%	90.6%
	correlation	.921	.914	.906	.891
0.5	accuracy	87.5%	86.8%	84.7%	84.0%
	correlation	.810	.830	.768	.755

Dimensionality and IRT analysis. The performance of the program in generating the data was adequate. In addition to the assessment made by changing the seeds of the random number generator and comparing the resulting descriptive statistics of the parameters, the linear and non-linear factor analysis have showed that the program was generating the data as expected. That can be seen by examining the results in Tables 1 and 3. In Table 2 the decrease of the eigenvalues for the first factor shows that the data was departing from unidimensionality as ξ increased. The same interpretation could also be given by the decrease in the variance explained by the first factor. The ratios of the eigenvalues of the first and second factors also revealed the same results. The ratio was large for the first three levels of ξ while it was small at the last two levels. The last two levels also showed unexpected results; the first factor at level 0.4 was supposed to have λ and σ values higher than those at level 0.5.

Linear factor analysis was not a good method in assessing the dimensionality of binary test data, but it provided a crude estimation of the test dimensionality. In this study, it actually provided an idea of how the factor structure of the generated data would look. The results were consistent with what other researchers found. Nandakumar (1991), for example, recommended that tests depart from essential unidimensionality as ξ reaches 0.4.

In the non-linear factor analysis, similar results were found (see Table 3). For one thing, the trend clearly showed how the dimensionality of the data changed with the change of the values of ξ. The values of the percent of standardized residuals greater than 1.96 also showed that the data could qualify as unidimensional up to ξ of 0.3.

The two-parameter logistic model, which was used to generate the data, provided adequate fit of the data at the two lower levels of ξ (0.0 and 0.3), but not when the strength of the minor dimensions was set at 0.5. That was not unexpected given the results found in the factor analysis step. Since it was possible to obtain real tests that fit the model as poorly as was found for the last set of data (ξ of 0.5), it was decided to accept the poor fit and proceed with the rest of the analyses.

Test 1. The correlation coefficients of the ability scores and criterion (true ability scores) is often used as test validity, and the decision accuracy is often used as a validity index with criterion-referenced tests. The indices shown in Tables 5 and 8 were obviously high in all cases, especially when the data were strictly unidimensional. But that is not surprising since a good criterion (without errors) was used in the study. Apparently, both indices decreased as the strength of the minor dimensions increased. The results in Table 6 also highlighted that the optimal method and the optimal-balanced method are superior in selecting more valid tests than the random and the optimal-unbalanced methods as the test data departed from strict unidimensionality. Same claim could be made by looking at the decrease in decision accuracies in Table 8.

One may wonder whether a small decrease in validity (correlation coefficient) or decision accuracy is important or practically significant. Lord (1963) showed that with a test of moderate validity (0.6), a decrease of 0.03 in validity could be obtained by reducing the test length by half. Let us take as an example the case of the optimal-balanced and optimal-unbalanced methods when ξ is 0.3 in Test 1 (Table 6). The difference in validity mean is 0.016. Since the validity indices are all high, let us assume they are at their limits which are the square roots of the corresponding reliability indices; that is, the mean reliability for each item selection method and level of ξ will be the square of the corresponding validity index. Using the relationship between test validity and test length (see, for details, Gullikson, 1950), an increase in validity of a test constructed through the optimal-unbalanced method by 0.016 requires an increase of the test length by 30 percent; that is, to add 12 more items to the test.

In the case of the decision accuracy, let us take one of the replications, as an example, in which the optimal-balanced and optimal-unbalanced methods differ by 1.7 percent when ξ is 0.3. The test needed to be increased by 50 percent to increase the decision accuracy of the optimal-unbalanced method by 1.7 percent. That requires adding 20 more items to the test; i.e, making the test and testing time longer, and increasing the test expenses.

In short, validity and decision accuracy gains of the order seen in Table 6 and Table 8 are significant. For example, the average decision accuracy improved 1.7 percentage points (from 88.5% to 90.2%) in switching from optimal-unbalanced to optimal-balanced at ξ of 0.3. This improvement is about 15 percent of the maximum improvement possible in decision accuracy of the optimal-unbalanced method. On the other hand, a decrease in decision accuracy of 1.7 percent will misclassify 170 examinees if the test was taken by 10000 examinees, and it is common for many tests to be taken by as many as 50000 examinees per year.

Also important is the fact that these gains in decision accuracy and validity could be attained by using desirable item selection methods such as optimal and optimal-balanced methods instead of substantially increasing the test length and testing time.

The statistical significance of the differences in correlation coefficients among the item selection methods and levels of ξ is clear in the analysis of variance results in Table 7. It is also clear that there is an interaction effect between item selection and strength of minor dimensions, which means some item selection methods reduce the decrease in validity more than others when the test data departs from unidimensionality. Similarly, the significance of the differences in decision accuracy among item selection methods, among strengths of the minor dimensions, and their interactions can be seen in Table 10. The significance of the interaction reveal that some item selection methods perform better than others in developing tests with high decision accuracies when the test is not strictly unidimensional; these are the optimal and optimal-balanced methods.

The cut-off score was lowered from 0.0 to -0.685, a point where 75 percent of the examinees passed Test 1. It is not uncommon in many licensure tests to have similar cut-off scores where 70 percent or more examinees pass them. The decision accuracies and most of the correlation coefficients at this level of cut-off score were higher than at the other cut-off score (ability score of 0.0). The validity increased for the optimal methods because the true item difficulties of the item pools were -0.53 and the estimates were even some times lower. Since many more items were available in the region of the cut-off score, the estimation of ability scores could have been better. As the cut-off score was moved away from that region where the item pool was concentrated, it was likely that the errors in the ability estimation would be increased near the new cut-off score. For the increments in decision accuracy, the effect could be attributed to the fact that the examinee population was concentrated at the other cut-off score (θ of 0.0), and hence more decision errors could result than in using this lower cut-off score where fewer examinees fail. At the lower cut-off score the optimal and optimal-balanced methods performed better than the other methods, and the difference was more profound as the data departs from unidimensionality.

<u>Test 2</u>. Test 2 was generated to represent credentialing exams that have lower discrimination values. The decision accuracies and correlation coefficients in all cases and cut-off scores were lower than the corresponding values in Test 1. The effect could be attributed to the fact that lower a-values usually result in less accuracy in ability estimation, and hence will result in less decision accuracies and validity indices. The trends seen in Test 1 were also seen in Test 2; that is validity and decision

accuracy decreased as data departed from unidimensionality, and optimal and optimal-balanced methods perform better than the other methods especially when the strength of the minor factors increased. The differences in validity and decision accuracy were also significant. For example, the mean validity of the tests developed by using the optimal-unbalanced at ξ of 0.3 (0.902) was lower than the mean validity of the tests developed by using the optimal-balanced method (0.917) by 0.015. To obtain equal validity indices for the two tests, the former needs to be increased by 20 percent or lengthened to 48 items. As another example, the decision accuracies of the two tests developed by using optimal-balanced and random methods at ξ of 0.3 differed in one of the replications by 3.0 percentage points. The test developed by the random method needed to be increased by 100% in order to attain equal decision accuracies for the two tests.

The significance of the differences in correlation coefficients and decision accuracies is also supported in the analysis of variance results in Table 7 and Table 10. The interaction effect was also significant for the correlation coefficient, and close but not in the case of the decision accuracy. The latter finding was difficult to explain, but one may argue that this result could be a type 2 error since all other results showed significance. Another argument could be that this interaction effect might become significant if the sample size was increased.

When the cut-off score was lowered to -0.685 in Test 2, the optimal method performed differently than how it performed in the other cases; i.e, in Test 1 and at higher cut-off score in Test 2. At ξ of 0.0, it provided decision accuracy about 1 percent lower than the optimal-balanced and the optimal-unbalanced methods. At ξ of 0.5, it provided decision accuracy 0.7 percent higher than the accuracy of the optimal-balanced method. Apart from these minor changes in the decision accuracies provided by the optimal method, all other results were similar to previously found results in Test 1 and Test 2. Those little changes might be caused by different, some times opposing effects; the lowered cut-off score, the low discrimination values, and/or the fact that the item pools were concentrated near the lower cut-off score.

Conclusion. The linear and non-linear factor analyses of the datasets both provided results showing how the factor structure of the generated data changed when the strength of the minor dimensions was changed. Both dimensionality investigations showed that the datasets could be ranked as unidimensional up to ξ values of 0.3. The goodness-of-fit analyses showed that the two parameter logistic model satisfactorily fit the datasets at lower two levels of ξ, and less adequately but acceptably fit the datasets at the highest level of ξ.

The validities and decision accuracies of the constructed tests decreased as the strength of the minor dimensions were increased for all cut-off scores and both simulated achievement tests and credentialing exams. Some item selection methods performed better than others, and the differences were more noticeable when the test data departed from unidimensionality. The optimal methods provided better tests in terms of decision accuracy and ability estimation when the item pool was strictly unidimensional, and the optimal and optimal-balanced methods performed better than the random and optimal-unbalanced methods when the test was not strictly unidimensional.

It was shown that the choice of an item selection method matters in test construction, and that the choice is more important when the item pool is not strictly unidimensional. Small differences in validity and decision accuracy among the item selection methods appear to be practically significant. One might need to substantially increase the length of a test constructed with a random or optimal-unbalanced method to match its validity or decision accuracy to a test constructed with optimal or optimal-balanced method. In other words, the optimal and optimal-balanced item selection methods might cut the test length or the testing time in half without any loss of test validity and decision accuracy.

Part B: Real Data

Goodness-of-Fit Analysis

As in the simulated data, linear and non-linear factor analyses were performed on the real data. In doing so, the responses of 1000 examinees on 40 items selected to represent the content categories of the actual test were analyzed. Six items from category 6, 14 items from category 2, and 5 items from each of the other four categories were selected. The first five eigenvalues of the tetrachoric correlation matrix of the binary data were 4.18, 0.55, 0.49, 0.41, and 0.39. These values suggested that the test was unidimensional. In non-linear factor analysis, the same 40 items were fitted to one-, two-, and three-factor solutions using the IRT program NOHARM (Fraser & McDonald, 1988). The percent of the standardized residuals greater than 1.96 were, respectively, 0.033, 0.026, and 0.017. The sum of the squared residuals of the variance-covariance matrix were, respectively, 0.025, 0.022, and 0.022. These results provide additional evidence that the test data was unidimensional.

A sample of 65 items were selected from the 230 test items, and the one-, two-, and three-parameter logistic models were fitted to the sample test. Table 13 contains the standardized residuals after fitting the three models to the data. The results showed that the three-, and two-parameter models fit the data adequately, while the fit of the one-parameter model was not adequate. For the one-parameter model, for example, 25.32 percent of the residuals were greater than 1. Since the two-parameter logistic model was used in the first part of the study, it was decided to use it in this part of the study too.

Table 13

Distribution of Standardized Residuals From Fitting the Three Logistic Models to a Sample of 65 Items[1]

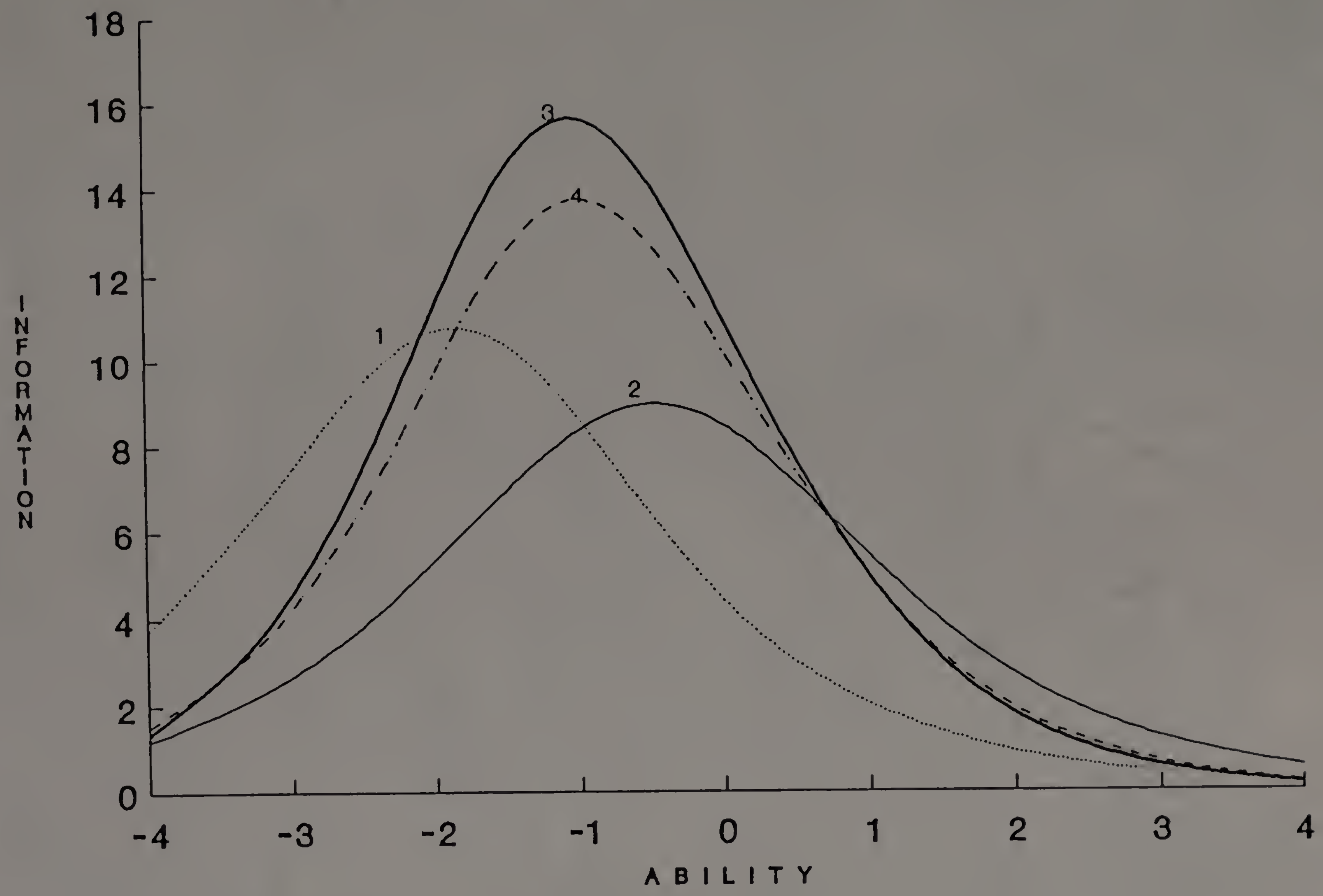
Standardized Residuals	Logistic Model 1	2	3	Normal
< -3	1.82%	0.28%	0.46%	0.2%
-3 to -2	5.59%	2.24%	3.69%	2.1%
-2 to -1	13.71%	10.49%	14.00%	13.6%
-1 to 0	24.20%	31.61%	26.62%	34.1%
0 to 1	27.37%	36.22%	41.08%	34.1%
1 to 2	20.14%	15.66%	12.77%	13.6%
2 to 3	4.48%	3.08%	1.38%	2.1%
> 3	0.70%	0.42%	0.00%	0.2%
AASR[2]	1.092	0.849	0.847	0.790

[1] The number of residuals were 780

[2] Average of the Absolute-valued Standardized Residuals.

Measurement Precision of the Constructed Tests

The four item selection methods discussed earlier in Chapter III (optimal, content-optimal, classical, and content-random) were used to select items from a pool of 130 items and to construct 40-item tests. Items were selected in the optimal test development using the cut-off score of the test, which was 70 percent and equivalent to -0.215 in the ability metric. The information functions provided by the four tests are shown in Figure 3. The optimal method provided the highest information, the content-optimal method provided the second highest information, and the random method provided the lowest information. At high ability levels (greater than 1), the classical method provided more information than the optimal methods.



1. Random, 2. Classical, 3. Optimal, 4. Content-Optimal.

Figure 3. Test Information Functions for the Four 40-Item Tests

The measurement errors at selected ability levels in the range [-1.215,.785] were also computed for each test, and the results are shown in Table 14. These results were obtained by using equation 5 in Chapter III, and are similar to the results provided by the information functions. The table shows that the two optimal methods provided the least errors at all ability levels except at 0.785 where the classical method provided the least errors. Optimal methods are usually focused at the cut-off points, and because of the location of the cut-off ability score (-0.215), they did not provide smaller errors at the higher ability levels. In this case, the optimal methods did not include many difficult items.

Table 14

Measurement Errors at Selected Ability Levels
Near the Cut-off Score

Item Selection Method	Ability				
	-1.215	-0.715	-0.215*	0.285	0.785
Content-Random	0.32	0.37	0.44	0.54	0.66
Classical	0.35	0.33	0.33	0.36	0.40
Optimal	0.25	0.25	0.28	0.33	0.41
Content-Optimal	0.27	0.27	0.29	0.34	0.41

* Cut-off score.

Decision Accuracies of the Constructed Tests

The decision accuracies could not be compared among dimensionality structures since the data had only one; and it was unidimensional. Hence, the decision accuracies were compared among the item selection methods alone. After the four tests were constructed as described in the preceding section, the decision accuracy was computed for each of them using the other 100 items (even items) of the test as criterion and the cut-off score of the test (-0.215 in the ability metric). Since the dimensionality assessment had shown that the test was unidimensional, the relative performances of the item selection methods were expected to be comparable to those obtained when ξ was 0.0 in the simulated data. The decision accuracies of the tests constructed with the four item selection methods are reported in Table 15. The content-optimal method provided the highest decision accuracy, the content-random method provided the lowest decision accuracy, and the optimal method produced the second highest decision accuracy.

Table 15

Decision Accuracies for the Four Item Selection Methods

Item Selection Method	Decision Accuracy	Improvement Factor
Content-Random	81.0%	--
Classical	83.9%	15.3%
Optimal	84.2%	16.8%
Content-Optimal	85.0%	21.1%

The method with lowest decision accuracy (content-random) was used as a baseline, and the percent improvement in decision accuracy over the maximum improvement possible in the baseline decision accuracy (19%) was computed for each of the other methods. The improvement factors of the optimal methods was quite substantial as can be seen from the table.

Discussion of Part B

The linear and non-linear factor analyses provided results showing that the data was unidimensional, and that the content categories or the other characteristics of the test do not constitute multidimensionality. That is not a general hypothesis for any test that may consist of different content categories but a particular aspect of this test. It could be true that the content categories of this particular test were measuring just one trait, while the content categories of another test could be measuring different but related traits. The IRT analysis showed that the data fit the two and the three parameter models, but did not adequately fit the one parameter model. The residual analysis revealed that the two-parameter logistic model provided a reasonable fit to the test data, and hence it was used in the rest of the analyses.

The optimal and content-optimal methods provided tests with ability measurement precisions higher than those provided by the random and classical methods. In terms of producing tests with the least measurement errors at the cut-off score, the item selection methods would be ranked (in a descending order) as follows: optimal, content-optimal, classical, and random. The content-optimal method, which is more desirable in terms of protecting the content validity of the resulting test, provided measurement errors reasonably close to those of the optimal method. The small differences in measurement precision among the item selection methods were still practically significant. At the cut-off score, for example, the test produced by the content random method would need to be increased by 153 percent or lengthened to 101 items to provide information as high as the content-optimal test. The classical test also would need to be increased by 32 percent to provide same information as the content-optimal test.

The decision accuracies were much lower than the previously reported values in the simulation study. The reason is merely that the two criteria are different; the criterion used for the real data analyses was part of the larger test while the criterion used in the simulation was the true abilities of the examinees. The latter criterion had fewer errors and closely matched the estimated ability scores. The decision accuracies of the tests constructed with optimal methods were higher than the decision accuracy of the random and classical methods. Between the optimal methods, the content-optimal method performed better than the optimal method. One explanation could be that in the real data, the representation of the content categories in the criterion matched the representation of the content categories in the pool, and hence the representation of the content categories of the test constructed with the content-optimal method.

The importance of the small differences among decision accuracies was discussed earlier in this chapter, and it is enough to restate that these differences are practically significant. The improvement factor of the item selection methods is another indication of the significance of the differences among decision accuracies. Even if the baseline was changed to the classical methods (since some testing agencies still use this procedure), the improvement factor of the optimal methods would be significant. The content-optimal method will have an improvement of 6.8 percent over the maximum improvement possible in decision accuracy of the classical method (16.1 percent). This improvement would be gained without a loss of content validity and from the same available item pool. In short, an item selection method in which the content validity of the resulting test was considered led the item selection methods in providing the highest decision accuracy.

CHAPTER V

SUMMARY AND CONCLUSIONS

In this last chapter of the dissertation, the summary of the findings in the research will be outlined. Second, the conclusions that follow from the study will be highlighted. Third, some implications of the results for the practitioner will be pointed out. Finally, some limitations of the study, and some suggestions for further research will be introduced.

Summary

Item response theory is used in the testing field for a variety of applications, and it is hoped that it will enjoy more extensive usage in the future. It provides excellent models and a useful framework for many practical psychometric problems such as equating, item bias studies, adaptive testing, item banking, and test development. Perhaps the most valuable property of IRT is the invariance property of the ability estimates and item statistics. This property, however, may not be attained unless a satisfactory fit between one of the IRT models and the data is obtained, and the strong assumptions of the theory are fulfilled. One of the hard-to-realize assumptions of IRT is the assumption of test unidimensionality which requires that the test data measure one common trait. There is abundant literature on the issue of unidimensionality, and ample evidence that in practice this assumption is often violated.

The violation of the unidimensionality assumption is often unintentional; that is, tests are developed to be unidimensional in most situations. However, a multitude of factors may cause the departure of a test from being unidimensional. Among these factors are the test administration process, the mode of presentation of test items, and measurement of different aspects of one subject in one test. It has been noted by many researchers that ability and achievement tests often mildly violate the unidimensionality assumption, and it has been proposed that the root of test multidimensionality is often the presence of minor traits beyond the major trait the test is intended to measure (Drasgow & Parson, 1983; Traub, 1983; Stout, 1987). Stout (1987) introduced, accordingly, the concept of essential unidimensionality, and many studies have been carried out along these lines (see, for example, Nandakumar, 1991; Sykes, Ito, & Potter, 1992). The issue of dimensionality became as Nandakumar (1991) puts it "how effective the minor dimensions should be" to label a test as multidimensional or essentially unidimensional. Another related question is how minor the minor factors should be to affect the quality of the test. This leads to the more fundamental question of whether the presence of the minor dimensions affect the reliability and validity of ability and achievement tests.

One purpose of this research was to examine the effect of the presence of minor dimensions on ability estimation and decision accuracy of mastery tests. A second purpose was to examine the performance of different item selection methods at different levels of test dimensionality. The results of the simulation study show that the decision accuracy decreases as the strength of the minor dimensions increases, and that the accuracy of ability estimates also decreases as the minor dimensions get stronger. In two types of simulated tests; one intended to simulate an achievement test and the other to simulate a credentialing exam, the effect of departure from unidimensionality on decision accuracy and validity was significant. This was true when the cut-off scores were at the center of the ability distribution and at a point where 75 percent of the examinees pass the test.

The optimal and content-optimal (optimal-balanced in the simulation) item selection methods did perform better than others in almost all situations. The differences in performance among the item selection methods, however, was more notable as the test departed from unidimensionality. Optimal item selection methods performed better than the random method of item selection in unidimensional item pools, and two optimal methods performed reasonably better when the strength of the minor dimensions was increased. One was the optimal method in which the items were selected according to the information they provide at the cut-off score regardless of which minor factor affects them. The other was the optimal-balanced method in which items were selected on the basis of the information they provide at the cut-off score, and the representation of the minor factors in the resulting test was balanced. There was an interaction effect between item selection methods and strength of minor dimensions which means that some item selection methods were more suitable than others as the test became multidimensional. The optimal-balanced and optimal methods are preferable when the test data are not unidimensional.

The differences in validity indices and decision accuracies among item selection methods and levels of lack of unidimensionality may appear small in magnitude but are significant in practice. Equalizing the decision accuracies of tests developed through two methods of item selection could mean increasing the test constructed with one method as much as 100 percent. Hence, from a practical point of view, the small differences in validity indices and decision accuracies among item selection methods and levels of minor factor strength are significant.

One purpose in the second part of the research was to examine the effect of item selection methods on decision accuracy and measurement precision. A secondary purpose was to assess whether this particular data was unidimensional. The real data was found to be unidimensional. The significant finding in the analysis of the real data was that optimal item selection methods provide better tests in terms of decision accuracy and measurement precision.

Without loss of content validity, and without requiring additional information in the item pool, the IRT-based optimal methods provided tests with high measurement precision, and the content-optimal method provided the test with the highest decision accuracy.

Conclusions

Criterion-referenced tests are being used by many state departments of education, credentialing agencies, armed services, and many other institutions to assess the competence and achievement levels of examinees. Item response theory offers models that overcome the shortcomings of the classical test models in the applications of criterion-referenced tests. IRT, however, has assumptions that are sometimes hard to meet in real life testing situations. One of the most difficult to meet is the assumption of unidimensionality. Several studies have been carried out on the robustness of IRT estimation programs and models to the violation of the unidimensionality assumption, and it has been found that the models are robust to the violation of the assumption to some extent. What has been missing, however, is research on the effects of the violation of the unidimensionality assumption on the validity and the reliability of mastery tests. This study examined one aspect of that issue; namely the effect of the presence and the strength of minor, unintended factors on validity and decision accuracy. Whether some item selection methods perform better than others in the presence of minor dimensions of different strengths was also investigated. In real data, the performance of several item selection methods in developing tests with high measurement precision and decision accuracy was also examined. Several conclusions can be derived from the results obtained in this study.

First, the strength of minor dimensions in a test do affect the validity and decision accuracy of criterion-referenced tests. This could happen due to (1) less adequacy of model-data fit, and/or (2) the fact that one ability is being measured while the examinees need to use more than one ability to answer the test items correctly.

Second, optimal item selection methods perform better in test development than the classical and random methods, and the optimal method and optimal-balanced method perform better than other methods especially when the test is not strictly unidimensional. Hence, the choice of item selection method will have an effect on validity, measurement precision, and decision accuracy of mastery tests. This effect is not unexpected since optimal methods select items that discriminate, and hence provide least errors of estimation, at the cut-off score of interest. What is not optimal, and will eventually lead to less decision accuracy, is to over sample one part of the test which is mainly affected by one minor factor when the test has several minor factors.

Third, the differences among item selection methods become more notable as the minor dimensions become stronger.

In other words, there is an interaction between the choice of item selection method and strength of minor dimensions. The optimal method which selects items based on the information they provide at the cut-off score, and the optimal-balanced method in which items are selected on the basis of their information functions in addition to the balance of the minor dimensions in the resulting tests, provide higher decision accuracies and better ability estimation, especially as the test departs from unidimensionality.

Finally, the methods that provide better tests in terms of decision accuracy, validity, and measurement precision do so without any additional cost or expenses. The other characteristics of the test such as content validity can be protected, and the methods use the same item pools that are available to all item selection methods. In other words, tests with higher validities, reliabilities, and decision accuracies can be developed easily without compromising the qualities of the required test. The optimal item selection process is made easier and simpler by the computer technology, and there are computer programs already available for these purposes such as OTD (Verschoor, 1991).

Implications of the Research for the Practitioner

Test data are not unidimensional in most practical situations, and the assumption is violated in a multitude of ways such as those discussed in Traub (1983), Drasgow and Parsons (1983), and Stout (1987). It is not uncommon to violate the unidimensionality assumption through the presence of minor dimensions beyond the major trait the test is intended to measure. The presence and the strength of the minor traits affect the validity of the mastery/ nonmastery decisions of the tests, and the test development process. The findings of this research have some implications for the testing practitioner which could be summarized as follows:

1. Assessment of test dimensionality is important for the intended use of the test. Unintentional minor factors such as reading in a mathematics test may affect the reliability, validity, and decision accuracy of the scores.
2. Goodness-of-fit investigations may not be sensitive to the presence of minor factors. A test with significant minor traits might well fit an IRT model as was the case in this study when ξ was set at 0.3 and 0.5 (see Table 4). Linear and non-linear factor analyses appear to be more effective ways of detecting the presence and the strengths of minor traits.
3. One way to detect the presence and the strengths of the minor dimensions could be to analyze the data with a multidimensional IRT program such as NOHARM (Fraser & McDonald, 1988) or MIRTE (Carlson, 1987), and examine the item discrimination indices for the different dimensions. Knowing the relative potencies of the minor dimensions will help the test developer decide on

whether to use the test as if it were unidimensional. If the mean discrimination of the minor dimension is as high as 30 percent of the total discrimination, unidimensional analysis and interpretation of the data is not a good choice.

4. The presence and the strength of the minor traits do affect the ability estimation, validity, and decision accuracy of mastery tests. This may lead to misclassification of substantial number of examinees, and may undermine the usefulness of the test. The problem might be avoided by fitting a multidimensional model to any data with potent minor dimensions.
5. Optimal item selection methods, which use IRT-based formulations, provide tests with relatively high validity indices and decision accuracies even when minor dimensions are operating in addition to the major trait the test is supposed to measure. These methods also provide relatively high levels of measurement precision. One exception is when there are several minor traits, and a high percentage of the items are selected from one of the minor traits. In that case, the ability estimation and decision accuracy might be lower than even the random method of item selection. The optimal item selection methods are not the best solution for the validity and decision accuracy of a test with strong minor dimensions, but are merely better than other methods of test construction.

Limitations of the Research

This research has highlighted a practical problem which is common to many tests that are developed for examinee classification into mastery levels. However, several limitations should be noted. First, the number of replications performed for each type of test was small compared to what is often seen in many Monte Carlo studies.

Second, the situations investigated might not always be found in real tests. The study design was built on other empirical research and real data in choosing its variables, the number of minor dimensions, the test length, the pool size, the choice of item parameter statistics for the simulation, and the potencies of the minor dimensions. These variables and parameters were chosen to fit many common situations found in practice but obviously could not match all situations.

Third, the criterion used for the classification decisions in the simulation study was more valid than any criterion that might be used in testing practice. This has resulted in the high decision accuracies and ability correlations reported in this study. This could have been avoided, and these numbers could have been smaller, for example, if another variable which is correlated to the examinee dominant abilities was used as criterion. In the real data, the criterion (even items) may not be desirable since it is part of the original test.

Fourth, the guessing parameter was not included in either the simulation or the real data analysis. Chance success is common in testing practice, and it is not clear how this parameter could have affected the results of the study.

Fifth, the ability score is only one of test scores reported in applied testing situations. Some practitioners prefer the number correct-score or transformation of it. Finally, the real data did not have minor dimensions to facilitate thorough analyses comparable to those performed in the simulated data. It represented the best case where the test is strictly unidimensional, and hence limited the value of real data analysis.

Suggestions for Further Research

In light of the results and limitations of the research, a number of suggestions for further research can be offered. The case where items are not equally divided among the minor factors is not addressed in this study. This is related to the number of items a minor factor needs to affect in order to be effective. Hence, an investigation is needed to assess the effect of number of items per minor dimension on decision accuracy, and whether this factor causes or interacts with the strength of minor dimensions.

Item pools in some testing agencies are quite large, and pool size might have an effect on the performance of item selection methods. Further research in which the item pool size is varied, its effect on decision accuracy is assessed, and its interaction with item selection methods and potency of minor dimensions is examined would also be useful. It is not uncommon for some items to be affected by more than two dimensions. It is possible for one item to measure one major ability and two minor abilities. A study investigating the effect of dimensionality on decision accuracy of test data in which items are affected by more than one minor dimension beyond the major ability could also be carried out. Finally, reliability of mastery tests is also important in practice. The effects of dimensionality on test reliability is an area where further investigation is needed.

REFERENCES

Ackerman, T. A. (1989, April). An alternative methodology of creating parallel test forms using IRT information function. Paper presented at the meeting of the NCME, San Francisco.

Ackerman, T. A. (1991, April). An examination of the effect of multidimensionality on parallel forms construction. Paper presented at the meeting of the NCME, Chicago.

Adema, J. J. (1990). The construction of customized two-stage tests. Journal of Educational Measurement, 27, 241-253.

Ansley, T. M., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Baker, F., Cohen, A., & Barmish, B.R. (1988). Item characteristics of tests constructed by linear programming. Applied Psychological Measurement, 12, 189-199.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program. Iowa City: American College Testing Program.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 10, 1-19.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In H. Wainer & S. Messick (Eds.), Principles of modern psychological measurement. Hillsdale, NJ: Erlbaum.

Christofferson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-22.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Davey, T., & Hirsch, T. (1991, April). Examinee discrimination and the measurement properties of multidimensional tests. Paper presented at the meeting of the NCME, Chicago.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Green, D. R., Yen, W., & Burket, G. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2, 297-312.

Gullikson, H. (1950). Theory of mental tests. New York: John Wiley & Sons.

Haladyna, T., & Roid, G. (1983). A comparison of two approaches to criterion-referenced test construction. Journal of Educational Measurement, 20, 271-281.

Hambleton, R. K. (1983). Application of item response models to criterion-referenced assessment. Applied Psychological Measurement, 7, 33-44.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 199-230). Baltimore: Johns Hopkins University Press.

Hambleton, R. K., Arrasmith, D., & Smith, I. L. (1987). Optimal item selection with credentialing examinations (Laboratory of Psychometric and Evaluative Research Report No. 157). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., & de Gruijter, D. N. (1983). Applications of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 20, 355-367.

Hambleton, R. K., Dirir, M. A., & Lam, P. (1992, April). Effects of optimal test designs on measurement precision and decision accuracy. Paper presented at the meeting of the AERA, San Francisco.

Hambleton, R. K., & Jurgensen, C. (1990). Criterion-referenced assessment of school achievement. In C. R. Reynolds & R. W. Kamphaus (Eds.), Handbook of psychological and educational assessment (pp. 456-476). New York: The Guilford Press.

Hambleton, R. K., Mills, L. N., & Simon, R. (1983). Determining the lengths for criterion-referenced tests. Journal of Educational Measurement, 20, 27-38.

Hambleton, R. K., & Murray, L. (1983). RESID: A computer program (Laboratory of Psychometric and Evaluative Research Report). Amherst, MA: University of Massachusetts, School of Education.

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

Hambleton, R. K., & Rogers, H. J. (1989). Solving criterion-referenced measurement problems with item response models. International Journal of Educational Research, 13, 145-160.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. Journal of Educational Statistics, 11, 91-115.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Huynh, H. (1976). On the reliability of decisions in domain- referenced testing. Journal of Educational Measurement, 13, 253-264.

Knol, D., & Berger, M. (1991). Empirical comparison between factor analysis and multidimensional item response models. Multivariate Behavioral Research, 26, 457-477.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. Applied Psychological Measurement, 4, 547-561.

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

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

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

Luecht, R., & Miller, T. (1991, April). Unidimensional calibration and interpretations of multidimensional tests. Paper presented at the meeting of the NCME, Chicago.

McDonald, R. P. (1967a). Nonlinear factor analysis. Psychometric Monographs (No. 15).

McDonald, R. P. (1967b). Numerical methods for polynomial models in nonlinear factor analysis. Psychometrika, 32, 77-112.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379-396.

McDonald, R. P. (1985). Unidimensional and multidimensional models for item response theory. In D. J. Weiss (Ed.), Proceedings of the 1982 Computerized Adaptive Testing Conference (pp. 127-148). Minnesota: University of Minnesota.

McDonald, R. P. (1989). Future directions for item response theory. International Journal of Educational Research, 13, 205-219.

McNemar, Q. (1946). Opinion-attitude methodology. Psychological Bulletin, 43, 289-374.

Messick, S. (1975). The standard problem: Measuring and values in measurement and evaluation. American Psychologist, 30, 955-966.

Mislevy, R. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Mislevy, R., & Bock, R. D. (1986). BILOG: Maximum likelihood item analysis and test scoring with logistic models. Mooresville, IN: Scientific Software.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28, 99-117.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, 25, 193-203.

Ree, M. (1979). Estimating item characteristic curves. Applied Psychological Measurement, 3, 371-389.

Roznowski, M., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary items. Applied Psychological Measurement, 15, 109-127.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53-61.

Stocking, M., Swanson, L., & Pearlman, M. (1990, April). Automated item selection using item response theory. Paper presented at the meeting of the NCME, Boston.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. Psychometrika, 52, 589-618.

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-276.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 11, 263-267.

Sykes, R., Ito, K., & Potter, R. (1992, April). Assessing the impact of multidimensionality on the classification decisions of an IRT-based licensure examination. Paper presented at the meeting of the NCME, San Francisco.

Taguchi, G. (1987). System of experimental design. White Plains, New York: Kraus International Publications.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52, 383-408.

Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Traub, R. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 57-70). Vancouver, British Columbia: Educational Research Institute of British Columbia.

van der Linden, W., & Boekkooi-Timminga, W. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.

Verschoor, A. (1991). Optimal test design (a software package). Arnhem, The Netherlands: CITO.