A COMPARISON OF THREE ITEM SELECTION METHODS
IN CRITERION-REFERENCED TESTS

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Hui-Fen Lin, B.A., M.Ed.
Denton, Texas
August, 1988

This study compared three methods of selecting the best discriminating test items and the resultant test reliability of mastery/nonmastery classifications. These three methods were (a) the agreement approach, (b) the phi coefficient approach, and (c) the random selection approach.

Test responses from 1,836 students on a 50-item physical science test were used, from which 90 distinct data sets were generated for analysis. These 90 data sets contained 10 replications of the combination of three different sample sizes (75, 150, and 300) and three different numbers of test items (15, 25, and 35).

The results of this study indicated that the agreement approach was an appropriate method to be used for selecting criterion-referenced test items at the classroom level, while the phi coefficient approach was an appropriate method to be used at the district and/or state levels. The random selection method did not have similar characteristics in selecting test items and produced the lowest reliabilities, when compared with the agreement and the phi coefficient approaches.
Copyright by
Hui-Fen Lin
1988
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td></td>
</tr>
<tr>
<td>Purposes of the Study</td>
<td></td>
</tr>
<tr>
<td>Null Hypotheses</td>
<td></td>
</tr>
<tr>
<td>Background of the Study</td>
<td></td>
</tr>
<tr>
<td>Significance of the Study</td>
<td></td>
</tr>
<tr>
<td>Basic Assumptions</td>
<td></td>
</tr>
<tr>
<td>Definition of Terms</td>
<td></td>
</tr>
<tr>
<td>Chapter References</td>
<td></td>
</tr>
<tr>
<td>II. REVIEW OF LITERATURE</td>
<td>25</td>
</tr>
<tr>
<td>Methods Using Preinstruction-Postinstruction Group Differences</td>
<td></td>
</tr>
<tr>
<td>The Pretest-Posttest Approach</td>
<td></td>
</tr>
<tr>
<td>The Individual Gain Approach</td>
<td></td>
</tr>
<tr>
<td>The Net Gain Approach</td>
<td></td>
</tr>
<tr>
<td>The Internal Sensitivity Approach</td>
<td></td>
</tr>
<tr>
<td>Method Using Uninstructed-Instructed Group</td>
<td></td>
</tr>
<tr>
<td>Methods Using Only A Postinstruction Group</td>
<td></td>
</tr>
<tr>
<td>The B Index Approach</td>
<td></td>
</tr>
<tr>
<td>The Latent Trait Approach</td>
<td></td>
</tr>
<tr>
<td>The Phi Coefficient Approach</td>
<td></td>
</tr>
<tr>
<td>The Agreement Approach</td>
<td></td>
</tr>
<tr>
<td>Comparative Studies Focusing Upon the Magnitude of Agreement and Correlation in Selecting the Best Discriminating Items</td>
<td></td>
</tr>
<tr>
<td>Comparative Studies Focusing Upon the Reliability of Mastery/Nonmastery Classification</td>
<td></td>
</tr>
<tr>
<td>Chapter References</td>
<td></td>
</tr>
<tr>
<td>III. METHODOLOGY</td>
<td>59</td>
</tr>
<tr>
<td>Overview</td>
<td></td>
</tr>
<tr>
<td>Instrumentation</td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td></td>
</tr>
<tr>
<td>Selection of Samples</td>
<td></td>
</tr>
<tr>
<td>Determination of Number of Test Items</td>
<td></td>
</tr>
<tr>
<td>Procedures for Analysis of Data</td>
<td></td>
</tr>
<tr>
<td>Data Preparation</td>
<td></td>
</tr>
<tr>
<td>Analysis of Magnitude of Agreement</td>
<td></td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1.</td>
<td>Possible Outcomes of Preinstruction-Postinstruction Test</td>
</tr>
<tr>
<td>2.</td>
<td>Distribution of Students Answering Correctly Item i in Terms of Pre- and Posttest Total Test Scores</td>
</tr>
<tr>
<td>3.</td>
<td>Distribution of Students Answering the Item Correctly and Incorrectly in Terms of Classification of Mastery and Nonmastery</td>
</tr>
<tr>
<td>4.</td>
<td>Block Diagram for ANOVA Design</td>
</tr>
<tr>
<td>5.</td>
<td>Percentage of Overlap of Selected Test Items in Samples of 75 Subjects</td>
</tr>
<tr>
<td>6.</td>
<td>Percentage of Overlap of Selected Test Items in Samples of 150 Subjects</td>
</tr>
<tr>
<td>7.</td>
<td>Percentage of Overlap of Selected Test Items in Samples of 300 Subjects</td>
</tr>
<tr>
<td>8.</td>
<td>Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for $P_0$ in the 15-Item Tests</td>
</tr>
<tr>
<td>9.</td>
<td>Scheffe Post Hoc Comparisons for $P_0$ Mean Differences Among Item Selection Methods in the 15-Item Tests</td>
</tr>
<tr>
<td>10.</td>
<td>Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for $P_0$ in the 25-Item Tests</td>
</tr>
<tr>
<td>11.</td>
<td>Scheffe Post Hoc Comparisons for $P_0$ Mean Differences Among Item Selection Methods in the 25-Item Tests</td>
</tr>
<tr>
<td>12.</td>
<td>Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for $P_0$ in the 35-Item Tests</td>
</tr>
<tr>
<td>13.</td>
<td>Scheffe Post Hoc Comparisons for $P_0$ Mean Differences Among Item Selection Methods in the 35-Item Tests</td>
</tr>
</tbody>
</table>
14. Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for K (Kappa) in the 15-Item Tests ..........................84

15. Analysis of Variance Table With Tests of Simple Main Effects due to Item Selection Methods.....85

16. Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods at the 75-Sample Size.................................86

17. Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods at 150-Sample Size........................................86

18. Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods at 300-Sample Size.................................87

19. Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for K (Kappa) in the 25-Item Tests.................................89

20. Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods in the 25-Item Tests.................................89

21. Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for K (Kappa) in the 35-Item Tests.................................90

22. Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods in the 35-Item Tests.................................91

23. Relationship Between the Mean of K (Kappa) Values and Test Length.................................91

24. Analysis of Variance Table With Tests of Simple Main Effects due to Sample Sizes.................................92

25. Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Sample Sizes at the Phi Coefficient Approach.................................93

26. Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Sample Sizes at the Agreement Approach.................................93

vii
## LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Relationship between the agreement and the latent trait approaches in terms of the item characteristic curve</td>
<td>41</td>
</tr>
<tr>
<td>2.</td>
<td>Example of joint distribution of ( f(A,B) )</td>
<td>68</td>
</tr>
<tr>
<td>3.</td>
<td>Illustration of the relationship between item selection methods and sample sizes in the 15-item tests</td>
<td>88</td>
</tr>
<tr>
<td>4.</td>
<td>Illustration of the relationship between sample sizes and item selection methods in the 15-item tests</td>
<td>88</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

In most applications of criterion-referenced testing procedures, test scores are used for making mastery decisions, that is, to determine whether students have mastered an instructional objective and therefore, are able to proceed to the next objective or have not mastered and need further instruction (Hambleton, 1974; Swaminathan, Hambleton, & Algina, 1974). If a student's performance exceeds the predetermined criterion level of performance, he is designated as having satisfactorily mastered the objective; otherwise, a nonmastery designation is declared, indicating that he has not satisfactorily mastered the objective. Therefore, in constructing a criterion-referenced test, the primary purpose is to select items that can be maximally sensitive to the differences between test scores corresponding with "masters" and "nonmasters" (Harris, 1983).

There are statistical methods of item analysis which assist in selecting appropriate criterion-referenced test items. This sort of item analysis is used to select a subtest of items from a large item pool so that the resulting test will have certain psychometric characteristics. The goal of item analysis in criterion-
referenced tests is to select items having the property of "differential sensitivity"; that is, masters answer the item correctly and nonmasters answer it incorrectly (Mellenbergh & van de Linden, 1982).

A number of item selection methods for criterion-referenced measurements have been proposed in the literature. Among these are (a) the pretest-posttest approach (Cox & Vargas, 1966), (b) the individual gain approach (Roudabush, 1973), (c) the net gain approach (Kosecoff & Klein, 1974), (d) the internal sensitivity approach (Kosecoff & Klein, 1974), (e) the uninstructed-instructed approach (Klein & Kosecoff, 1975), (f) the \( B \) index approach (Brennan, 1972), (g) the latent trait approach (van de Linden, 1981), (h) the phi coefficient approach (Hsu, 1971), and (i) the agreement approach (Harris, 1983).

Berk (1980a) suggested that in choosing an appropriate item selection method to analyze test items, the "law of parsimony" should be considered; that is, the item selection method should be "conceptually and computationally simple yet statistically sound" (p. 59). Among the proposed item selection methods, the agreement approach requires only one test administration and has similar theoretical characteristics to the latent trait theory (Harris, 1983); the phi coefficient approach, a special case of the product moment correlation coefficient, requires only one testing
procedure, and is best used with items scored as pass/fail (Hsu, 1971); and the random selection approach is a common and straightforward method, requiring no computation (Popham, 1978, Hambleton, 1982). In this study, the phi coefficient approach, the agreement approach, and the random selection method were used because they were practical and meaningful.

Test length can affect the reliability of mastery/nonmastery classification (Hambleton, Mills & Simon, 1983). As noted by Hambleton (1984), "Mastery classification based on scores obtained from short tests are often unreliable and invalid... Low probabilities of misclassification can usually be assured when tests are very long" (p. 144). With reference to how many test items should be included for the final version of the test, several researchers have offered different statistical strategies. Millman (1973) proposed the use of a binomial model in which the approximate true score for each examinee was required. Novick & Lewis (1974) suggested the use of a Bayesian method in which the probability that an examinee's domain score exceeded a given cut-off score should be set. Based on the item response theory, Hambleton et al. (1983) proposed the use of computer simulation methods to investigate the relationship between test length, test reliability, and validity. However, it seems no single method provides an appropriate determination of an acceptable test length. For instance, "There is no
magic number or even magic formula for determining test length" (Wilcoxon, 1980, p. 44); "None of the test length determination methods ... is without shortcomings" (Hambleton, 1984, p. 167); and "As you can see, reaching a decision about an acceptable minimum test length is a long way from being simple" (Popham, 1978, p. 102).

In spite of these limitations, some guidelines for determining the test length are available. Wilcoxon (1979) said, "Typically mastery tests consist of 25 or less items" (p. 20). The rule of thumb suggested by Popham (1978) for determining test length is a test consisting of somewhere between 10 and 20 items per behavior. Berk (1980b) recommended between 5 and 10 items per objective be used for classroom decisions, and between 10 and 20 items be used for school, system, and state level decisions. Moreover, Swezey (1981) stated, "In selecting final items for inclusion in a test, the test developer should have available a pool of about twice as many items as are required for the final version of the test" (p. 97). Based on these recommendations, 15 items (between 10 and 20 items), and 25 items (half of the 50-item pool used in this study), respectively were used for the final forms of test in this study. Since Hambleton (1984) suggested that longer tests assured low probability of misclassification, a 35-item instrument was also included.

In conducting item analysis, Swezey (1981) suggested that at least 50 percent more subjects were needed in the
tryout sample than there were items in the item pool, while Nunnally and Wilson (1975) suggested that at least 5 times as many subjects as items were necessary in order to avoid the complication arising from chance. The prediction that sample size will influence the effect of item selection methods needs to be tested. Since the instrument used in this study constructed from a pool of 50 items, samples of 75, 150 and 350 subjects were used to investigate the relationship between different sample sizes and item selection methods.

In assessing the consistency of mastery/nonmastery classifications, two coefficients are frequently used: $P$ and $K$ (kappa) (Berk, 1984). $P$ is defined as the proportion of individuals consistently classified as masters and nonmasters on two parallel tests. $K$ is the proportion of consistent classifications beyond that expected by chance. In this study, both $P$ and $K$ were used as reliability indices for measuring the consistency of mastery/nonmastery classifications for the tests constructed by the agreement, the phi coefficient, and the random selection methods.

Statement of the Problem

This study was proposed for the following reasons:

1. The "best" method in selecting criterion-referenced test items is still unknown.
2. There is insufficient information to determine if a one-test item method is better than the pure random selection method.

3. There are insufficient data to determine the effect of different sample sizes and test lengths on methods in selecting criterion-referenced test items.

Purposes of the Study

The purposes of this study were to:

1. Determine to what extent the three methods agree with one another in their selection of items having desired discrimination effects in samples of 75 subjects.

2. Determine to what extent the three methods agree with one another in their selection of items having desired discrimination effects in samples of 150 subjects.

3. Determine to what extent the three methods agree with one another in their selection of items having desired discrimination effects in samples of 300 subjects.

4. Determine whether the three item selection methods yield tests with similar $P_\omega$ values when the consistency of mastery/nonmastery classification is measured.

5. Determine whether the resultant tests have similar $P_\omega$ values when three different sample sizes are considered.

6. Determine whether the three item selection methods yield tests with similar $K$ values when the consistency of mastery/nonmastery classification is measured.
7. Determine whether the resultant tests have similar $K$ values when three different sample sizes are considered.

Null Hypotheses

This study assumed the following null hypotheses:

1. The three methods will select the same test items that best discriminate between masters and nonmasters in samples of 75 subjects.

2. The three methods will select the same test items that best discriminate between masters and nonmasters in samples of 150 subjects.

3. The three methods will select the same test items that best discriminate between masters and nonmasters in samples of 300 subjects.

4. The three methods will yield tests with similar $P$ values when the consistency of mastery/nonmastery classification is measured.

5. The resultant tests will have similar $P$ values when three different sample sizes are considered.

6. The three methods will yield tests with similar $K$ values when the consistency of mastery/nonmastery classification is measured.

7. The resultant tests will have similar $K$ values when three different sample sizes are considered.
Background of the Study

Traditionally, educators have used norm-referenced tests to evaluate and interpret student achievement. Such measures can provide information about the capability of a student in relation to the capability of other students (Popham & Husek, 1969; Popham, 1978). For instance, a classroom teacher can tell, based on a norm-referenced test of computation speed, whether student A solves single-digit addition problems quicker than student B. On the other hand, student A’s achievement is interpreted in terms of a comparison between his performance and the performance of other students. Such testing results are dependent on a particular student population rather than on a predetermined domain of behaviors.

While norm-referenced tests provide information regarding the ordering of students relative to a proficiency standard, they provide little or no information concerning mastery or nonmastery of criterion tasks (Davis, 1974). Thus, as Glaser (1963) indicated, “They [norm-referenced measures] tell that one student is more or less proficient than another, but do not tell how proficient either of them [students] is with respect to the subject matter tasks involved” (p. 520).

Criterion-referenced tests differ from norm-referenced tests by indicating what a student with a given score can or
cannot do, without any reference to scores of other students (Glaser, 1963; Popham, 1981; van de Linden, 1982).

The term "criterion-referenced test" was first introduced by Glaser in 1963, and then extended by Popham and Musek in 1969. Glaser (1971) defined a criterion-referenced test:

A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual. Measurements are taken on representative samples of tasks drawn from this domain and such measurements are referenced directly to this domain for each individual measured. (p. 41)

In 1978, Popham defined the purpose of a criterion-referenced test as "... to ascertain an individual's status with respect to a well-defined behavior" (p. 63). These two definitions clearly indicate some important points. First, student performance is interpreted and evaluated in terms of predetermined performance standards rather than in comparison with the performance of other students (Mathews, 1985). Second, the objectives measured in a criterion-referenced test need to be well-defined. That is, the
content or behaviors defining each objective should be clearly described. Third, the word "criterion" refers to a domain of content or behavior to which test scores will be referenced rather than to a standard or cut-off score (Hambleton, 1980).

Millman (1974) emphasized the importance of a well-defined domain in criterion-referenced tests, since "test scores can be interpreted most directly in terms of performance task, and also, the percent of the population of tasks the students would answer correctly or in a given direction can be estimated" (p. 314). Nitko (1980) made a similar suggestion. He said that well-defined domains were necessary for criterion-referenced tests because the basic idea of criterion-referenced testing was to generalize how well an examinee could perform in a broad class of behaviors.

Many educators have stated that criterion-referenced testing has multiple applications for educational practice and decision-making. Some have claimed that criterion-referenced testing is applicable to curriculum development (Klein & Kosecoff, 1975; Popham, 1978), assessment of students' progress (Millman & Popham, 1974; Klein & Kosecoff, 1975; Hambleton, 1978; Hambleton & Eignor, 1978; Popham, 1978; Hambleton, 1980), and diagnostic assessment (Hambleton, 1978; Hambleton & Eignor, 1978; Popham, 1978; Hambleton, 1980; Swezey, 1981; Mills, 1983). Additional

In viewing the wide applicability of criterion-referenced tests, researchers have been concerned about how to select appropriate items that function consistently with the intended purpose. Judgmental evaluation is one of the methods used to select items. In this approach, the content experts examine each item for degree of item-objective congruence. Each item is evaluated in relation to the intended instructional and behavioral objectives from the defined domain of tasks (Popham, 1978; Berk, 1980a, 1984). An item is retained if it meets the predetermined item-objective congruent criterion; otherwise, it is eliminated from the item pool. Once criterion-referenced test items have been developed to be consistent with a carefully defined domain of tasks, the test is constructed by random selection since all items in an item pool should be homogeneous and interchangeable (Popham, 1978; Hambleton, 1982; Hambleton & de Gruijter, 1983). When the items are
randomly selected from item pool, the construction of a test is finished and no further empirical item analysis is needed. Consequently, Popham (1978) and Hambleton (1982) said that the random selection approach was a common and straightforward strategy in constructing criterion-referenced tests.

The use of empirical data to analyze items is another method. This approach involves utilizing indices of item difficulty and item discrimination. Item difficulty refers to the proportion of correct answers to an item; item discrimination refers to the extent to which an item effectively discriminates between those students possessing more of a trait and those possessing less (Popham & Husek, 1969). Researchers like Millman (1974) and Hively (1974) conclude that in criterion-referenced tests, once an item was included in the domain, no further empirical analysis should lead to reconstructing the item or eliminating it from the domain. However, other researchers draw different conclusions. They argue that item writing often involves subjectivity, which might lead to the use of items with unwanted properties; such items can only be detected through the use of item analysis. Therefore, these researchers have suggested that empirical item analysis is absolutely necessary (Saupe, 1966; Lange, Lehmann & Mehrens, 1967; Hsu, 1971; Smith, 1978; Haladyna & Roid, 1981; van de Linden, 1981; Mellenbergh & van de Linden, 1982).
Popham and Husek (1969) stated that traditional item selection methods were not appropriate when empirically analyzing items of criterion-referenced tests. The traditional procedures are dependent on correlational analyses, and correlational analyses rely on a reasonable degree of variability in scores. If scores are not distributed normally, correlational analyses may be invalid. With criterion-referenced tests, the variability of scores is not so important. Although any criterion-referenced test will get variant scores, the variability is not a necessary condition for a good criterion-referenced test (Popham & Husek, 1969). A "good" criterion-referenced test administered after adequate instruction may result in very low variance within scores. In this sense, if an item accurately measures an essential behavior and the examinees respond either all correctly or all incorrectly, the item cannot be eliminated. Furthermore, traditional item selection methods are devised in accordance with the norm-referenced tests' purpose of permitting relative comparison (Popham, 1978). In criterion-referenced testing, however, the primary interest is to compare student performance with some predetermined standards rather than with those of other students (Glaser & Nitko, 1971). As a result, new methods of item analysis for criterion-referenced testing purpose are necessary (Cox, 1971).

Several methods have been proposed for selecting items for a criterion-referenced test. Each approach has its own
theoretical perspective and statistical basis. Methods using preinstruction-postinstruction group differences involve testing one group of students twice, before instruction (preinstruction) and after instruction (postinstruction). These methods include: (a) the pretest-posttest approach, which reflects the difference between the proportion of students answering the item correctly on the posttest and the proportion of students answering the item correctly on the pretest (Cox & Vargas, 1966); (b) the individual gain approach, which measures the proportion of students answering the item incorrectly on the pretest and correctly on the posttest, after a correction for guessing has been applied (Roudabush, 1973); (c) the net gain approach, which measures the difference between the proportion of students answering the item incorrectly on both the pre- and posttests and the proportion of students answering the item correctly on the pretest but incorrectly on the posttest (Kosecoff & Klein, 1974); and (d) the internal sensitivity index approach, which reflects the difference between the proportion of students who are masters on the posttest but nonmasters on the pretest and the proportion of students who are nonmasters on both the pre- and posttests (Kosecoff & Klein, 1974).

A method using two groups, one with and another without instruction, is called the uninstructed-instructed method. This approach was proposed by Klein and Kosecoff in 1975.
Methods using only a postinstruction group involve testing only one group of students after instruction. They are: (a) the B index approach, which represents the difference between the proportion of students in the upper group who answer the item correct and the proportion of students in the lower group who answer the item correctly (Brennan, 1972); (b) the latent trait approach, which measures the power of an item to discriminate between masters and nonmasters at a given cut-off score (van de Linden, 1981); (c) the phi coefficient approach, which "is a special case of Pearson product moment coefficient, for use with dichotomous variables" (Swezey, 1981, p. 108); and (d) the agreement approach, which represents the proportion of masters answering the item correctly and the nonmasters answering the item incorrectly (Harris, 1983).

Comparative studies of the differing item selection methods have been conducted by many researchers. These studies are reviewed in Chapter II of this study.

Significance of the Study

Most research has focused on the magnitude of agreement of item selection methods in selecting the same best-discriminating items. Only three studies, Smith (1978), Harwell (1983), and Silva (1985), could be found that examined the effect of item selection method on the accuracy and consistency of mastery/nonmastery classification decisions.
The purpose of this study was to evaluate three item selection methods on the magnitude of their agreement in selecting the best discriminating items, and the resultant test reliability. The phi coefficient is simple to compute and requires only one test administration. The agreement approach is recently proposed. This method is also easy to compute and needs only one testing procedure. Of most importance, the agreement approach has theoretical properties similar to the latent trait theory. The random selection method is a common and straightforward approach. It requires no mathematical computation.

The importance of this study was fourfold. First, this study further examined the usefulness of the agreement approach. Harris and Subkoviak (1986) suggested that "the agreement statistic...be tentatively recommended for further study and for possible classroom use" (p. 506).

Second, the effect of item selection methods on the resultant test reliability was examined. Berk (1984) stated, "It is important to appraise how the magnitude of the discrimination indices affects the accuracy and consistency of mastery-nonmastery classification decisions" (p. 133).

Third, the relationship between the agreement approach and the phi coefficient approach was evaluated. Kosecoff and Klein (1974) stressed, "the most important approach for evaluating item quality is an examination of the item in
context with total test performance" (p. 25). Both the agreement and phi coefficient approaches select items on the basis of total test score. It is important, therefore, to understand and compare the characteristics of these two methods.

Fourth, the random selection method was used for comparative study. If the results indicate that the random selection method has similar or equal efficiency in selecting items, compared with the agreement and phi coefficient methods, the random selection approach may be recommended for classroom use because such a method requires no computation.

Basic Assumptions

It was assumed that test items used in this study were drawn from a well-defined and carefully constructed item pool.

Definition of Terms

The following terms have restricted meaning and are defined as follows for this study:

1. A cut-off score is the score that classifies examinees into either the mastery or the nonmastery category.

2. Masters are those examinees whose total test scores are above the predetermined cut-off score.
3. **Nonmasters** are those examinees whose total test scores are below the predetermined cut-off score.

4. "**Best**" **items** are those items that masters answer correctly and nonmasters answer incorrectly.

5. **Reliability** refers to the consistency of mastery/nonmastery classifications.
Chapter References


CHAPTER II

REVIEW OF LITERATURE

As noted in Chapter I, several methods have been proposed for selecting items for a criterion-referenced test. Each approach has its own theoretical perspective and statistical basis. Methods using preinstruction-postinstruction group differences are (a) the pretest-posttest approach (Cox & Vargas, 1966), (b) the individual gain approach (Roudabush, 1973), (c) the net gain approach (Kosecoff & Klein, 1974), and (d) the internal sensitivity index approach (Kosecoff & Klein, 1974). A method using two groups, one with and another without instruction, is called the uninstructed-instructed method. This approach was proposed by Klein and Kosecoff in 1975. Methods using only a postinstruction group are (a) the B index approach (Brennan, 1972), (b) the latent trait approach (van de Linden, 1981), (c) the phi coefficient approach (Hsu, 1971), and (d) the agreement approach (Harris, 1983).

Methods Using Preinstruction-Postinstruction Group Differences

In group difference methods, the same group of students is tested twice, with the same or a parallel test, before instruction (preinstruction) and after instruction
(postinstruction). Item statistics are based on differences in the pretest-posttest item difficulties, with the assumption that high values reflect a movement from nonmastery to mastery.

The Pretest-Posttest Approach

The pretest-posttest ($D_{PP}$) approach was proposed by Cox and Vargas in 1966. This method has been used in evaluation of quality of instruction and mastery learning (Carroll, 1963, 1970). The $D_{PP}$ index indicates the difference between the proportion of students who answer the item correctly on the posttest and the proportion of students who answer it correctly on the pretest. Specifically, the $D_{PP}$ is defined by Cox and Vargas (1966) as:

$$D_{PP} = P_{\text{pp}} - P_{\text{pl}}.$$  

where $P_{\text{pp}}$ = proportion of students who answer correctly on the posttest.

$P_{\text{pl}}$ = proportion of students who answer correctly on the pretest.

The range of $D_{PP}$ values is from $-1.00$ to $+1.00$. A positive value of $D_{PP}$ reflects the movement of students from nonmastery to mastery over the instructional period. A negative $D_{PP}$ value is undesirable because it indicates that
more students answered the item correctly on the pretest than on the posttest. Furthermore, such a condition is considered aberrant, since it is unlikely that instruction would cause students to have a poorer performance. One advantage of the pretest-posttest approach is ease of computation and interpretation. This approach is, however, not without its own problems. The index does not indicate which students or how many students actually improved in performance. It is also insensitive to individual performance change and has problems related to carry-over effects (Berk, 1980, 1984). In addition, the pretest-posttest approach depends on the effect of instruction as the basis for discrimination and requires two test administrations (Harris, 1983; Harris & Subkoviak, 1986).

The Individual Gain Approach

Roudabush proposed the individual gain (\(D_{ig}\)) approach in 1973. This approach measures the proportion of the students who answer the item incorrectly on the pretest and correctly on the posttest, after a correction for guessing has been applied. Roudabush (1973) defined the \(D_{ig}\) as (see Table 1):

\[
D_{ig} = \frac{n_2'}{(n_1' + n_2')} \quad (2)
\]

where \(n_2'\) = the number of students correctly responding to the item on the posttest but incorrectly
on the pretest, after correction for
guessing has been applied.

\[ n_1' = \text{the number of students incorrectly}
\]
responding to the item on both the pretest
and the posttest, after the correction for
guessing has been applied.

Kosecoff and Klein (1974) labeled this index as an
external sensitivity index because the measure of an item
discrimination ability in this approach is independent of
the total test score. This approach has similar advantages
and disadvantages to the pretest-posttest approach. In
addition, Kosecoff and Klein (1974) claimed that the \( D_{11} \)
index did not really measure "gain", although it was called
an "individual gain" approach. Meanwhile, Berk (1980, 1984)
questioned the assumption regarding the correction for
guessing; that is, students passing the item on the pretest
and failing it on the posttest answered the pretest item
correctly by guessing.

The Net Gain Approach

With the net gain \( (D_{n_1}) \) approach, the index is computed
by subtracting the proportion of students who answer the
item incorrectly on both the pretest and the posttest from
the proportion of students who answer the item correctly on
the posttest but incorrectly on the pretest (Kosecoff & Klein, 1974). Kosecoff and Klein (1974) defined the $D_{nj}$ as (see Table 1):

$$D_{nj} = \frac{n_2 - n_1}{N} \quad (3)$$

Like the individual gain approach, the index of net gain is an external sensitivity index because it is independent of the total test score. The range of $D_{nj}$ values is also from -1.00 to +1.00. A value of +1.00 indicates maximum change in the direction of learning. A +1.00 is the ideal, because

Table 1
Possible Outcomes of Preinstruction-Postinstruction Test

<table>
<thead>
<tr>
<th></th>
<th>Posttest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pretest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$n_1^*$</td>
</tr>
<tr>
<td>1</td>
<td>$n_2$</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$n_3$</td>
</tr>
<tr>
<td>1</td>
<td>$n_4$</td>
</tr>
</tbody>
</table>

Note. 1's = correct response.
0's = incorrect response.

$n_1^*$ = number of students answering the item incorrectly on both the pretest and posttest.

$$N = n_1 + n_2 + n_3 + n_4.$$

(Source: Kosecoff & Klein, 1974).
it indicates a lack of knowledge before and successful knowledge following instruction. A value of -1.00, on the other hand, indicates that all students fail the item on the pretest and the posttest, and therefore, no instructional effect occurs. Such a result suggests that either the instruction is of no benefit to students or that the item does not discriminate among students.

Like the individual gain approach, the advantages and disadvantages of the net gain approach are similar to the pretest-posttest approach. However, compared with the individual gain approach, the net gain approach is "statistically more conservative because it considers the individual's gains and losses resulting from the instruction rather than just the total group gain or loss" (Berk, 1980, p. 61).

**The Internal Sensitivity Index Approach**

The internal sensitivity index (D) approach described by Kosecoff and Klein (1974) is used to measure "an item's performance, within the context of the total test, comparing how a given item and the entire test discriminate among learners" (p. 8). This index reflects the proportion of students correctly answering the item who fail the pretest but pass the posttest minus the proportion of students correctly answering the item who fail both the pre- and posttests. In other words, this index discriminates among
students (correctly answering the item) who were nonmasters before instruction and masters after instruction. Since the measure of an item discrimination in this approach is dependent on the total test score, Kosecoff and Klein (1974) called it an "internal sensitivity index". Kosecoff and Klein (1974) defined the $D_{ij}$ as (see Table 2):

$$D_{ij} = \frac{n_x - n_{ij}}{N} \quad (4)$$

The values of $D_{ij}$ range from -1.00 to +1.00. A value of -1.00 occurs when all students are nonmasters in both the pre- and posttests, but correctly answer the item. Such an item is not desirable because it is not sensitive to instruction and does not discriminate between masters and nonmasters. A value of +1.00 means that all students answering the item correctly are nonmasters on the pretest but masters on the posttest. Such an item is an ideal one, because it can discriminate between students who are nonmasters prior to instruction and masters after instruction.

According to Kosecoff and Klein (1974), the internal sensitivity approach has some satisfactory properties. It is easy to compute and understand; it is similar in form to other approaches; and it has positive correlations with other methods. However, because the identification of mastery or nonmastery for both the pre- and posttests needs to be made, there is a tenuous dependence on a criterion
Table 2

Distribution of Students Answering Correctly Item i in Terms of Pre- and Posttest Total Test Scores

<table>
<thead>
<tr>
<th>Fail posttest</th>
<th>Pass posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail pretest</td>
<td>$n_1$</td>
</tr>
<tr>
<td>Pass pretest</td>
<td>$n_3$</td>
</tr>
</tbody>
</table>

Note. $n_1$ = number of students failing both the pretest and posttest among those who correctly answer item i.

$n_2$ = number of students failing the pretest but passing the posttest among those who correctly answer item i.

$n_3$ = number of students passing the pretest but failing the posttest among those who correctly answer item i.

$n_4$ = number of students passing both the pre- and posttests among those who correctly answer item i.

$N = n_1 + n_2 + n_3 + n_4$.

(Source: Kosecoff & Klein, 1974).

(Berk, 1980). Under these conditions, two kinds of errors may occur: (a) a false positive error—declaring a learner as master when, in fact, he is not; and (b) a false negative
error—declaring a learner as nonmaster when, in fact, the the reverse is true (Hambleton & Novick, 1973). In addition, like the pretest-posttest approach, the D_{i|s} also requires two test administrations, depends on instruction as the basis for discrimination, and is not independent of the group upon whose responses it is computed (Harris, 1983).

Method Using Uninstructed-Instructed Group

The statistic computed with the uninstructed-instructed (D_{ui}) approach represents the proportion of students in the instructed group who answer the item correctly minus the proportion of students in the uninstructed group who answer it correctly (Crehan, 1974). The D_{ui} is defined by Kosecoff and Klein (1975) as:

\[ D_{ui} = \frac{n_1}{N_1} - \frac{n_2}{N_2} \]  

where \( n_1 \) = number of students in the instructed group who answer the item correctly.

\( n_2 \) = number of students in the uninstructed group who answer the item correctly.

\( N_1 \) = total number of students in the instructed group.

\( N_2 \) = total number of students in the uninstructed group.
In this approach, when the uninstructed and instructed groups are available, the item analysis can be conducted at one time (Berk, 1980). In addition, from an experimental design perspective, this approach is often recommended because the uninstructed group acts as a control group against which to measure the effectiveness of the instruction (Campbell & Stanley, 1963). Nevertheless, using such an approach, there are difficulties in obtaining an equivalent group and in defining appropriate criteria with which to identify the groups (Berk, 1980, 1984).

Methods Using Only A Postinstruction Group

These methods include: (a) the B index approach (Brennan, 1972); (b) the latent trait approach (van de Linden, 1981); (c) the phi coefficient approach (Hsu, 1971); and (d) the agreement approach (Harris, 1983).

The B Index Approach

The B index approach uses a traditional upper-lower discrimination index $D$, which is defined by Johnson (1951) as:

$$D = \frac{U}{n} - \frac{L}{n} \quad (6)$$

where $U$ = the number of students in the upper group who get the item correct.

$L$ = the number of students in the lower group
who get the item correct.

\[ n = \text{the total number of students in each group.} \]

In this approach, the total number of subjects is required to be equal in both upper and lower groups. Derived from index D, the index B is defined by Brennan (1972) as:

\[ B = \left( \frac{U}{n_1} \right) - \left( \frac{L}{n_2} \right) \quad (7) \]

where \( n_1 \) can be equal to or different from \( n_2 \). On the other hand, the total number of subjects in both the upper and lower groups are not necessarily equal.

Because there is no requirement to have equal number of subjects in the upper and lower groups, the B index approach gives researchers freedom to choose appropriate cutoff scores between these groups. In addition to this advantage, Brennan (1972) also indicated that the B index was more appropriate than the D index in a situation where total test scores were not normally distributed.

**The Latent Trait Approach**

Based on the latent trait theory, van de Linden (1981) proposed the latent trait approach \( \lambda_1 (\Theta) \) as a replacement for the pretest-posttest approach as an item selection technique for criterion-referenced tests. A latent trait theory is a theory that specifies a relationship between an
examinee's observable test performance and the unobservable abilities assumed to underlie performance on the test. This theory supposes that students' performance on a test can be predicted or explained by defining their characteristics and can be estimated on the basis of his/her latent traits. Furthermore, latent trait theory assumes that a student's response to an item is independent from his response to any other items in the test; and the items in a test measure a single ability or latent trait (Hambleton & Cook, 1977, Hambleton & Swaminathan, 1985).

At a given cut-off score, the latent trait approach can measure the power of an item to discriminate between masters and nonmasters. According to van de Linden, $I_\perp(\theta)_{\perp}$ is defined as:

$$I_\perp(\theta)_{\perp} = P_\perp(+|\theta) \ast [1 - P_\perp(+|\theta)]$$

where the $P_\perp(+|\theta)$ is the probability of responding correctly to an item given ability level $\theta$. Since the $I_\perp(\theta)_{\perp}$ statistic is derived from latent trait theory, it can select more reliable items than the pretest-posttest approach (Harwell, 1983). In addition, it requires only one test administration. However, the $I_\perp(\theta)_{\perp}$ statistic has some serious disadvantages. These limitations include: the need for a relatively large sample size; the use of a computer and appropriate software for computation; and the background in latent trait theory to understand and interpret the
results (van de Linden 1981; Harris, 1983; Harwell, 1983). Therefore, Harris and Subkoviak (1986) concluded, "Thus, while it [the latent trait approach] is to be preferred on theoretical and empirical grounds, it is impractical for classroom use in certain respects" (p. 498).

The Phi Coefficient Approach

The phi coefficient ($\phi$) approach is "a special case of Pearson product moment correlation coefficient, for use with dichotomous variables" (Swezey, 1981, p. 108). It was proposed by Hsu (1971). In this approach, the examinees are divided into masters and nonmasters by their total test score and a predetermined cut-off score. According to Hsu (1971), the Phi ($\phi$) (see Table 3) is computed as:

$$
\phi = \frac{(n_1 \times n_4) - (n_2 \times n_3)}{\sqrt{(n_1 + n_2) \times (n_4 + n_4) \times (n_1 + n_3) \times (n_2 + n_4)}}
$$

(9)

The values of phi ($\phi$) may range from -1.00 to +1.00. An item is considered as acceptable if phi ($\phi$) is greater or equal to +0.30 (Swezey, 1981). The phi ($\phi$) is easy to compute and interpret, requires only one test administration, and does not depend on instruction as the basis of discrimination (Harris, 1983). In addition to these advantages, phi ($\phi$) tends to favor the items of moderate difficulties (Englehart, 1965), is best used with
items scored as pass/fail (Ferguson, 1981), and when there are approximately the same number of masters and nonmasters in the subject sample (Swezey, 1981). However, Swezey (1981) found that when a very small sample size was used, for instance, less than eight subjects, phi ($\phi$) was not appropriate. In addition, phi ($\phi$) is not suitable to be used in the situations when every student is declared a master or nonmaster or when the item is answered correctly or incorrectly by every student (Hsu, 1971).

**The Agreement Approach**

In viewing the problems existing in the pretest-posttest approach and the latent trait approach, Harris (1983) proposed the agreement approach as a compromise between these two methods. The $P(X_c)$ is computed as the number of masters answering the item correctly plus the number of nonmasters answering the item incorrectly, divided by the total number of subjects. Harris (1983) defined the $P(X_c)$ as (see Table 3):

$$P(X_c) = \frac{n_1 + n_4}{N} \quad (10)$$

The rationale for this index is that an item reflects the "agreement between outcomes on a particular item and outcomes on the total test" (Harris & Subkoviak, 1986, p. 499).
Table 3

Distribution of Students Answering the Item Correctly and Incorrectly in Terms of Classification of Mastery and Nonmastery

<table>
<thead>
<tr>
<th>Master</th>
<th>Nonmaster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>$n_1$</td>
</tr>
<tr>
<td>Incorrect</td>
<td>$n_3$</td>
</tr>
</tbody>
</table>

Note. $n_1$ = the number of masters answering the item correctly.

$n_2$ = the number of nonmasters answering the item correctly.

$n_3$ = the number of masters answering the item incorrectly.

$n_4$ = the number of nonmasters answering the item incorrectly.

$N = n_1 + n_2 + n_3 + n_4$.

(Source: Harris & Subkoviak, 1986).

Harris (1983) claimed that the agreement approach had several satisfactory properties. It is easy to compute, requires only one test administration, does not depend on instruction as the basis of classification, and most
important, has similar theoretical characteristics to the latent trait theory. Harris (1983) explained the basis of a close relationship between the agreement approach and the latent trait approach in Figure 1. The area below the item characteristic curve in Figure 1 represents the students who correctly answer the item. So, area I corresponds to $n_1$ (who are masters and correctly respond to the item), area II to $n_2$, area III to $n_3$, and area IV to $n_4$. In selecting items, the latent trait approach is based on the magnitude of the information function at $\theta = \lambda'$, and the agreement approach is based on the magnitude of $(n_1 + n_4) / N$. From the item characteristic curve in Figure 1, Harris (1983) concluded that the agreement approach and the latent trait approach should "select much the same items" (p. 93). While the agreement approach exhibits these satisfactory properties, the agreement approach has, like the internal sensitivity index approach, the problem of "criterion dependence" because it needs to make use of a cut-off score in its computation.

Comparative Studies Focusing Upon the Magnitude of Agreement and Correlation in Selecting the Best Discriminating Items

Cox and Vargas (1966) compared the pretest-posttest method to the traditional upper-lower discrimination method. Fifty students from grade one through four were administered
Figure 1. Relationship between the agreement and the latent trait approaches in terms of the item characteristic curve (Source: Harris & Subkoviak, 1986).

Note. $\theta_\zeta$ = cutoff score.

$\theta$ = the probability of correct response.
a 31-item addition test, and 25 students from grade four to six were administered a 40-item multiplication test. Two indices for each item were computed and then ranked. The ranked order correlation was found to be positive but small, that is, .37 on the addition test, and .40 on the multiplication test. While these values were significant and indicated a tendency for two methods to rank items similarly, some discrepancies were found. For instance, the item ranked first by the pretest-posttest method was ranked at 23.5 by the upper-lower discrimination method.

Cox and Vargas also examined the magnitude of agreement in these two methods when the highest ranking one-third items were selected from the item pool. The results revealed that approximately 60% and 28% on the addition test and multiplication test, respectively, were the same for the two methods. Cox and Vargas concluded that these two methods produced sufficiently different results in selecting test items and advocated the use of the pretest-posttest method for criterion-referenced tests.

Hsu (1971) compared the phi coefficient, the pretest-posttest and the point-biserial approaches under three different situations: heterogeneous samples with symmetrical distributions; homogeneous samples with skewed score distributions; and items with a variety of item difficulties. The mastery level was set at 80% of items correct. The intercorrelations among these three item
selection indices were analyzed. When samples were heterogeneous and test scores were symmetrically distributed, the pretest-posttest, the phi coefficient, and the point biserial methods were found to be highly correlated to each other, for instance, .77 between the pretest-posttest and the point biserial, and .94 between the pretest-posttest and the phi coefficient. However, when samples consisted of homogeneous subjects and test scores skewed to either left or right, the correlations between pretest-posttest and point biserial, and between phi coefficient and point biserial were considerably lower, that is, .50 for the former and .54 for the latter. With respect to the effect of item difficulty on the correlations among these three methods, higher correlations were found only on items with middle difficulty (.40 to .70), in which correlations between .68 and .93 were found between the point biserial and the phi coefficient; between .81 and .93 were found between the phi coefficient and the pretest-posttest; and between .79 and .91 were found between the pretest-posttest and the point biserial. Hsu concluded that the pretest-posttest and the phi coefficient were consistent with the point biserial in most cases in selecting the criterion-referenced tests.

Gorth and Hambleton (1972) replicated the 1966 Cox and Vargas study, using different items and examinees. The methods considered in this study were: the point biserial;
the pretest-posttest; and the pretest-delayed-posttest. The cut-off score was not indicated in this study. The ranked correlations were found to range from -.31 to .38 between the point biserial and two pretest-posttest methods. Gorth and Hambleton also compared the magnitude of item overlap when the top ranking 25%, 50%, and 75% of item pool were selected. Overlap was found to range from 5.2% (between the point biserial and the pretest-delayed-postest) to 90% (between the pretest-posttest and the pretest-delayed-posttest). Gorth and Hambleton concluded that the choice of statistics had a significant effect on the final selection of test items.

Crehan (1974) compared six item selection methods. The procedures were the pretest-posttest, the modified B index ($n_1$ and $n_2$ were redefined as independent groups of instructed and uninstructed students, respectively), the posttest agreement, teacher rating, the point biserial and the random selection. Eighteen junior high and senior high teachers from various content areas wrote instructional objectives and constructed multiple choice test items to accomplish the objectives. Item pools were then compiled containing approximately 32 items. Within each item pool, the items were ranked by each of the six methods. Tests consisting of the best 1/3, 1/2, and 2/3 of item pools were assembled by each method. The results indicated that none of the methods maintained a consistently high reliability;
however, in all cases the modified B index and the pretest-posttest methods resulted in higher observed classification validity across the various subjects and teachers. Crehan concluded that the modified B index and the pretest-posttest approaches were, of the six, the ones to be favored.

Kosecoff and Klein (1974) proposed the internal sensitivity index and the net gain approach, and compared them with two other item selection methods: the phi coefficient; and the point biserial. Two sets of test data were employed: one had 7 items; the other, 71 items. The test mean was used as the cut-off score. The correlations between the indices in the 70-item test showed that the internal sensitivity approach was significantly correlated with both the phi coefficient (.83) and point biserial (.82), compared to .23 between the net gain with phi coefficient and .21 between the net gain with point biserial. Kosecoff and Klein concluded that the internal sensitivity index was an appropriate method in distinguishing between masters and nonmasters.

Haladyna (1974) studied three item selection methods: the point biserial, the pretest-posttest and the combined groups (pre- and postinstruction) point biserial. One hundred and eighty nine undergraduate students participated in this study and three test forms were employed. The mastery level for each test form was set at 70% of items correct. In the study of ranked correlation, the pretest-
posttest method was found to yield information similar to that of the combined groups point biserial (ranging from .64 to .86), but quite different information to that of the point biserial (ranging from .05 to .49). With respect to the study of magnitude of difference between item discrimination indices, the combined groups point biserial was found to produce higher discrimination indices than the pretest-posttest method did, in which the median of item discrimination value was from 0.18 to 0.54 in the combined groups point biserial, while from 0.13 to 0.40 for the pretest-posttest method. Haladyna concluded that the pretest-posttest method was more conceptually satisfying in communicating the meaning of criterion-referenced item discrimination because of its computing item difficulties for pre- and postinstruction sample; whereas the combined groups point biserial was more efficient in obtaining information about the adequacy of the criterion-referenced items because of its producing higher discrimination indices.

Van de Linden (1981) studied the relationship between the latent trait method and the pretest-posttest approach. A physics unit was taught to 156 tenth grade students, and a 25-item multiple choice test was administered as a pretest and as a posttest. The $D_p$ and $I_1(\Theta_c)$ statistics were computed for each item, and the correlation between these two statistics was computed across items. The correlation
was found to be -.19 when the 80% cut-off score was
employed. Using a much lower cut-off score (58%), a
correlation of .23 was found. Van de Linden concluded that
the pretest-posttest and the latent trait methods tended to
select very different subsets of items from the item pool.

Helmstadter (1972) compared the pretest-posttest
method, the traditional upper-lower discrimination method,
and the method of observing the shift in item difficulty
between pre- and postinstruction. A 59-item test was
administered to 28 students before and after instruction.
The cut-off score was set at 80% of items correct.

The ranked correlation between the pretest-posttest
approach and the method of observing the shift in item
difficulty between pre- and postinstruction was found to be
.78; that between the traditional upper-lower discrimination
approach and the method of observing the shift in item
difficulty between pre- and postinstruction was .29; and
that between the pretest-posttest and the traditional upper-
lower discrimination methods was .47.

Helmstadter concluded that the results confirmed those
which cautioned the use of the traditional upper-lower
discrimination method in the criterion-referenced
situations. Meanwhile, he suggested the pretest-posttest
method be used in constructing the criterion-referenced
tests.

investigated the interrelationship among seven methods used
in criterion-referenced tests: the pretest-posttest; the uninstructed-instructed; the net gain; the individual gain; the point biserial; the combined groups point biserial; and the phi coefficient. Sixty-six seventh grade students were randomly assigned to one of three groups: a preinstruction-postinstruction; an uninstructed group; and an instructed group. A 25-item test was then administered to these students.

Several results were revealed in this study. The pretest-posttest, the individual gain, the net gain, and the combined groups point biserial methods measured item discrimination similarly, but the uninstructed-instructed method tended to select different items. The phi coefficient and the point biserial tended to select less items proportionally than the other five methods, especially when the remedial instruction was effective. One interesting finding was that when the instruction was moderately effective, all the methods selected more items than when the instruction was highly effective.

Haladyna and Roid (1983) compared the latent trait approach and random selection method in terms of precision of measurement in criterion-referenced test construction. Four sets of items were employed, representing a variety of criterion-referenced domain-based tests. Test length varied from 10 to 40 items. Tests constructed by the latent trait approach were found to yield greater precision in domain
score estimation than tests created by random selection method. The result also indicated that test forms composed of 20 to 30 items were found to provide satisfactory domain score estimation, but not for test lengths of less than 20 items.

Harris (1983) proposed the agreement approach and compared its magnitude of agreement in selecting the best items with the pretest-posttest and the latent trait approaches. A number of distinct data sets were simulated to examine a variety of test construction. The mastery level was set at 75% and 85% of items correct, respectively. A ranked correlation of .91 was found between the agreement and the latent trait methods, while -.17 was found between the pretest-posttest and the latent trait methods. Furthermore, an average of 94 items out of 100 selected by the agreement approach was the same as those selected by the latent trait method. Harris concluded that the agreement method selected items similarly to the latent trait method and recommended it to be used in classroom settings because of its ease in computation and understanding.

Sarvela (1986) compared four discrimination indices: the B index; the phi coefficient; the traditional upper-lower D index; and the point biserial. One hundred and ten computer simulated subjects were created and three 20-item tests were administered to these simulated subjects. The scores were systematically distributed—normally, bimodally,
and negatively skewed—to represent test scores frequently occurring in military testing situations. The cutting score on each test was 10 items correct (50%). Analysis of variance and t-test were performed to compare the differences between discrimination index values.

In normal data sets, the traditional upper-lower $D$ statistic produced the largest average discrimination value (0.53) and point biserial (0.42) next, while the phi coefficient and $B$ index approaches produced least but identical discriminating values (0.35). Within the bimodal data sets, similar results were obtained. The results in skewed data, however, were completely different. The $B$ index approach produced the largest discrimination value (0.46), followed by point biserial (0.44), the traditional upper-lower $D$ statistic (0.41) and the phi coefficient (0.34). Sarvela concluded that the distributions of test scores could influence the values obtained from various indices. However, he suggested that in terms of cost efficiency, the $B$ index approach was the ideal method to be employed.

Comparative Studies Focusing Upon the Reliability of Mastery/Nonmastery Classification

Silva (1985) compared five item selection methods in the consistency of mastery classification for criterion-referenced tests. These methods included: the pretest-
posttest; the combined groups point biserial; the phi coefficient; the one-parameter and two-parameter latent trait methods. Nine hundred forty-five first and second grade students together with 1,796 forth, fifth, and sixth grade students participated in this study. A set of items, evaluated by content specialists, was employed as the initial data set. Some of the less relevant items were then added to this initial data set in order to simulating invalid items.

Items were ranked by all statistics at four cut-off scores: (a) the predetermined mastery level on the basis of each objective; (b) 50% of items correct; (c) 62.5% of items correct; and (d) 75% of items correct. For each item analysis procedure, a long (80 items) and short (15 items) test was constructed with the items ranked as best by each procedure. Students were designated as master or nonmasters on each of the resulting tests, and coefficients $K$ (kappa) were computed for the purpose of comparing the efficiency of classification consistency in these methods. The results revealed that the mean coefficient $K$ (kappa) for the two latent trait methods was 0.85, while a mean of 0.80 for the pretest-posttest, the phi coefficient, and the combined groups point biserial methods. Silva concluded that the latent trait approaches performed better than the three non-latent-trait methods.
Smith (1978) compared five item selection methods in the efficiencies of classification accuracy and classification consistency. These procedures included: the point biserial; the pretest-posttest; the \( E \) index; the phi coefficient; and the random selection methods. Three sets of 70-item response data for 1,000 subjects were simulated respectively under the situations of: ineffective instruction; effective instruction; and differentially effective instruction.

Based on each item selection method, 40 items were drawn from the original item pool, then 20 items were selected from the groups of 40 items to form the final test for assessing the classification accuracy and classification consistency. The mastery level was set at 80\% of items correct.

The results of analyzing the classification accuracy revealed that no single method was consistently superior to others in correctly classifying examinees when instructional effectiveness was varied. With respect to the measure of classification consistency, a similar finding was obtained, that is, no particular method was consistently superior to others under three situations of instructional effectiveness. However, Smith suggested that (a) when instruction was effective, the pretest-posttest approach was recommended; (b) when instruction was relatively ineffective, the \( E \) index and the phi coefficient methods
were the best instruments; and (c) when the instruction was differentially effective, the point biserial was recommended.

Using simulated data, Harwell (1983) compared the pretest-posttest and the latent trait methods on the reliability of mastery classification decisions. Different sample sizes (30, 60, and 300), test lengths (20 and 80) and cut-off scores (75% and 85% correct items) were examined, respectively. The average coefficient $P_0$ of .80 was found in tests with items selected by the pretest-posttest method, while .85 was found for those tests constructed by the latent trait approach. The average coefficient $K$ (kappa) was 0.51 for the pretest-posttest method and 0.35 for the latent trait approach. Harwell concluded that in terms of simplicity in computation and conceptualization, the pretest-posttest method was a desirable approach of item analysis in classroom settings.

In summary, many comparative studies among the various methods of selecting items for criterion-referenced tests have been conducted. Two conclusions can be made based on these studies. First, as suggested by Popham and Husek (1969), the traditional item selection methods are indeed not appropriate for analyzing items of criterion-referenced tests. This was emphasized in the studies conducted by Hsu (1971) and Sarvela (1986). Both Hsu and Sarvela indicated that a low correlation was found between the traditional
item selection methods and the methods proposed for criterion-referenced measurements when the distribution of test scores was either positively or negatively skewed. Second, almost all of the researchers selected the pretest-posttest approach to compare with other item selection methods, and some of these researchers recommended it to be used for analyzing items of criterion-referenced tests (for instance, Cox & Vargas, Hsu, Hambleton, Crehan, Haladyna, Helmstadter, Smith, and Harwell). The reason that the pretest-posttest approach is used and recommended so often by some researchers is, perhaps, that the pretest-posttest approach is the prototype of all criterion-referenced item selection methods (Harwell, 1983), and is the easiest as well as the simplest one, in terms of computation and conceptualization.

However, the need for two test administrations, problems related to carry-over effects and instructional dependency as the basis for discrimination in the pretest-posttest approach have also caused some other researchers to advocate other approaches. Although these methods are not as simple as the pretest-posttest approach in computing and understanding, they are more theoretically and statistically sound (for example, Harris, Silva, and van de Linden). Among the advocated alternatives, the agreement approach which correlates very well with the latent trait approach is the one to be recommended.
Chapter References


CHAPTER III

METHODOLOGY

Overview

This study was designed to compare the agreement approach, the phi coefficient approach, and the random selection method on their (a) extent of agreement in selecting the best discriminating test items and (b) resultant test reliability of mastery/nonmastery classification decisions over multiple data sets (subjects-by-item responses data matrices). In this study, the number of subjects and the number of test items were also examined for their effect on item selection. Specifically, three different sample sizes and three different item numbers were examined, for a total of $3 \times 3 = 9$ combinations. Each combination was replicated 10 times, resulting in a total of $9 \times 10 = 90$ distinct data sets.

In the agreement study, the best 15 discriminating items selected from the item pool were used as the best test items, then, the average percentage of item overlap over 10 replications between any two of these three methods in each sample size was examined on the basis of a predetermined 67% criterion.
In the reliability study, the reliability coefficients $P_0$ and $K$ (kappa) were computed separately over the 90 distinct data sets. For each test length, a split plot SPF-3.3 factorial analysis of variance (Kirk, 1982) was performed in which the sample size was the between factor, the item selection method was the within factor, and coefficients of $P_0$ and $K$ were the dependent variables.

Instrumentation

The data that were used in this study were collected through the use of a district-developed criterion-referenced test. The 50-item instrument was a measure of skills in physical science at the high school level. The items were evaluated by curriculum personnel with respect to curricula/content validity and the objectives were in close agreement with those objectives that were measured by the Texas Assessment of Basic Skills (TABS) tests.

Population

The test was not actually administered to the student in this study; instead, existing test results were used as data for statistical analysis. A total of 1,836 students' item responses were used in this study. The students were from a large urban school district in the Southwestern United States. Since the "population" employed in this study, in fact, referred to the students' test responses
rather than the students themselves, the "samples" and/or "subjects" used in the following sections were to represent "students' item responses".

Selection of Samples

As noted in Chapter I, Swezey (1981) said that at least 50 percent more subjects were needed in the tryout sample than items in the item pool. Since there were 50 items in the tests, 75 subjects were used in this study. In addition, Nunnally and Wilson (1975) said that at least 5 times as many subjects as items were needed so as to avoid taking advantage of chance. Therefore, for the purpose of studying the effects of sample size on the item selection, 150 and 300 subjects were also considered. All of these subjects were drawn at random and with replacement from the population.

Determination of Number of Test Items

As mentioned in Chapter I, several researchers provided some guidelines for the "acceptable" test length. Popham (1978), for example, recommended between 10 and 20 items per objective. Berk (1980b) suggested between 5 and 10 items per objective for classroom settings, and between 10 and 20 items for school, system, and state level decisions. Wilcox (1979) stated that a typical mastery test consisted of 20 or less items. Moreover, Swezey (1981) advocated an initial
item pool containing about twice as many items as the final test. According to these researchers' suggestions, 15 items (between 10 and 20 items), and 25 items (half of the initial 50-item pool), respectively, were used as the final test form lengths in this study. Also, because longer tests assured low probabilities of misclassification (Hambleton, 1984), 35 items were included in this study. All of these items were selected at random but without replacement from the initial item pool in the random selection method.

Procedures for Analysis of Data

Data Preparation

To perform statistical analysis for this study, a set of data was selected randomly and with replacement from the initial data matrix. A total test score for each individual student and a total number who passed each item were obtained first. The second step was to rearrange students based on their total test scores from the highest to the lowest values. Since the agreement and phi coefficient methods required a mastery score as part of their computation, the cut-off score was set at 80% of the test items correct. Hsu (1971), Helmstadter (1972), Smith (1974) and van de Linden (1981) also used this criterion as cut-off score in their studies. The $\phi$ and $P(X \geq x)$ indices were then computed by using equations (9) and (10), respectively.
Once the item discrimination indices were computed, the items were sorted in descending order on the basis of the size of each index. For the $P(X)$ and $P$ statistics, the highest 15, 25 and 35 items were used separately as final test forms. With reference to the random selection method (RM), 35 items were selected at random and without replacement from the initial 50-item pool. The first 15, 25, and 35 items were used separately as final test forms.

The above procedure was repeated 10 times, resulting in a total of 90 distinct data sets (as noted in Table 4). The replication was generated for the main purpose of computing a within cell error term for the split plot factorial analysis of variance design. According to Cohen (1977), the 10 replications "per cell" for each method-by-sample size combination with a medium-effect size had a statistical power of .80 at alpha = .05 significance level (p. 273-402).

Analysis of Magnitude of Agreement in Selecting the Same Best Discriminating Items

The highest ranking 15 items selected from the item pool were used as the best test items. The average amount of item overlap between the agreement approach and the phi coefficient approach, between the agreement approach and the random selection method, and between the phi coefficient approach and the random selection method over 10 replications in each sample size were examined. The
Table 4

Block Diagram for ANOVA Design

<table>
<thead>
<tr>
<th>Items</th>
<th>subjects</th>
<th>Replications</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>75</td>
<td>10</td>
<td>P(X&lt;sub&gt;c&lt;/sub&gt;)</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>75</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>75</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

criterion of 67% (2 out of 3 items selected being common to any two methods) was set a priori. In any two methods, if 67% or above of selected items were the same, then these two methods were declared as selecting the "best" discriminating items in common. Harris (1983) used this criterion to determine the amount of item overlap between items selected by various methods and suggested (in personal communication, see Appendix A) that the 67% was a common rule-of-thumb and adequate for reducing "computation intensity" which would result from item selection methods.
Analysis of Reliability of Mastery/Nonmastery Classification Decision

To compare the effect of these three selection methods on reliability of mastery/nonmastery classification decisions, coefficients $P_o$ and $K$ (kappa) were calculated over the 90 distinct data sets. $P_o$ represents the proportion of consistent decisions (mastery/mastery and nonmastery/nonmastery) based on two parallel tests, which can be estimated from a single test administration. Coefficient $K$ represents the proportion of consistent classifications corrected for chance, which can also be calculated from a single test administration.

Huynh's Beta-Binomial Method

To estimate $P_o$ and $K$ coefficients, Huynh's beta-binomial method (1976) were employed. Huynh's method has several advantages. It requires only one test administration (Berk, 1980, 1984), is based on an elegant mathematical model (Peng & Subkoviak, 1980), and provides relatively accurate estimates of $P_o$ and $K$ (Subkoviak, 1978; Gross & Shulman, 1980; van de Brink, 1982).

Huynh's beta-binomial method assumes that all items in the item pool are exchangeable; that is, all items are homogeneous in difficulty and content (Huynh & Saunders, 1980). From a single test administration (test A), it is
possible to estimate the probabilities associated with scores on a second unadministered test (test B). Within the beta-binomial framework, the mean ($\mu$), variance ($\sigma^2$), Kuder-Richardson coefficient 21 ($\alpha_{21}$) of test A, and parameters $\alpha$ and $\beta$ are computed, according to Huynh (1976), as follows in order to simulate test B:

$$\alpha_{21} = \frac{n}{n-1} \times [1 - \frac{\mu \times (n - \mu)}{n^* \sigma^2}] \quad (11)$$

$$\alpha = (-1 + \frac{1}{\alpha_{21}}) \times \mu \quad (12)$$

$$\beta = -\alpha + \frac{n}{\alpha_{21}} - n \quad (13)$$

where $n$ is the total number of items on the test. According to Keats and Lord (1962), the parameters $\alpha$, $\beta$ and $n$ can determine the joint distribution of mastery/nonmastery on tests A and B for every possible set of scores. This distribution, symbolized $f(A,B)$, represents the probability of each pair of $A$ and $B$ values. To generate the joint distribution $f(A,B)$, every possible value of $A$ is paired with every possible value of $B$, and the probability of the joint occurrence of these values is computed. For given values of $A$ and $B$ equal to 0, Huynh showed that the $f(A,B)$ is computed as:
\[
\frac{2n + \beta - n}{2n + \alpha + \beta - n} = \frac{f(n)}{n=1} \frac{2n + \beta - n}{2n + \alpha + \beta - n} \quad (14)
\]

and

\[
f(A+1, B) = f(A, B) \times \frac{(n - A)(\alpha + A + B)}{(A + 1)(2n + \beta - A - B - 1)} \quad (15)
\]

For given values of \(A\) and \(B\) close to \(n\), Huynh provided more manageable formulas:

\[
f(n, n) = \frac{2n + \alpha - n}{2n + \alpha + \beta - n} = \frac{f(n)}{n=1} \frac{2n + \alpha - n}{2n + \alpha + \beta - n} \quad (16)
\]

and

\[
f(A-1, B) = f(A, B) \times \frac{A \times (2n + \beta - A - B)}{(n - A + 1)(\alpha + A + B - 1)} \quad (17)
\]

This process is continued, using expressions (14) and (16), until the entire \(f(A, B)\) distribution has been generated.

However, since \(f(A, B)\) is symmetric in the sense that the \(f(A, B)\) probability is the same as the \(f(B, A)\) probability, only half the distribution actually has to be generated.

Once the \(f(A, B)\) distribution has been generated for a particular subjects-items difference combination, the \(f(A, B)\) values can be tabled in the format shown in Figure 2.
Figure 2. Example of joint distribution of $f(A, B)$
(Source: Huynh, 1976).
The region labeled master/master, for instance, represents the proportion of the \( f(A,B) \) distribution in which students score above cut-off level on test A and also above the cut-off level on test B. The sum of the \( f(A,B) \) probabilities in the master/master portion, therefore, represents the proportion of students consistently classified as masters on both tests A and B. \( P_Q \) is defined as the proportion of consistent classification (master/master and nonmaster/nonmaster) across two tests. Huynh (1976) defined \( P_Q \) as (see Figure 2):

\[
P_Q = P_{mm} + P_{nn}
\]  

(18)

Coefficient \( K \), the proportion of consistent decisions corrected for chance, is defined by Huynh (1976) as:

\[
K = \frac{P_O - P_C}{(1 - P_C)}
\]  

(19)

where \( P \) represents the proportion of consistent decision due to chance, and is also a function of the marginal proportions of masters and nonmasters in Figure 2. The \( P_C \) is defined by Huynh (1976) as (see Figure 2):

\[
P_C = P_{ma} \cdot P_{mb} + P_{ma} \cdot P_{nb}
\]  

(20)

where \( ma, mb \) represent the marginal totals of the proportion of students consistently classified as masters on tests A and B while \( na, nb \), represent the proportion of students consistently classified as nonmasters on tests A and B.
computer program was written to perform the preceding procedures. The source program was presented in Appendix B.

Split Plot Factorial Analysis of Variance

To determine whether the agreement approach, the phi coefficient approach, and the random selection method differed in their effect on resultant test reliability, the following procedure was employed. Once the coefficients, either $P_Q$ or $K$, over 90 distinct data sets had been computed, they were entered as observations into a split plot SPF-3.3 factorial analysis of variance design for each test length in which the item selection method was the within factor and the sample size was the between factor (Kirk, 1982). The split plot factorial analysis of variance was then computed by using PROC ANOVA in Statistical Analysis System (SAS) (SAS Institute Inc. 1985). The alpha level was set at .05 a priori. The interaction effect due to method-by-sample size was tested first. If significant interaction effect occurred, the test of simple main effect followed. When there were significant simple main effects, Scheffé post hoc comparisons were conducted.

If there was no significant interaction effect, main effects were tested. Following any significance of main effect, Scheffé post hoc comparisons were conducted.
The factor of test length was, unfortunately, unable to be considered in the split plot factorial analysis of variance design due to the fact that there was dependency among the selected items. That is, in the agreement and the phi coefficient approaches, the items were sorted and selected on the basis of their corresponding discrimination values, therefore, the selected 15 items overlapped in the selected 25 items and in the 35-item case. However, the mean coefficient, either $P_0$ or $K$, for each test length was presented to investigate whether longer tests generated higher reliability coefficients.
Chapter References


The analysis involved examining the magnitude of agreement in selecting the best discriminating items and the resultant test reliability of mastery/nonmastery classifications for the agreement, the phi coefficient, and the random selection methods. As noted earlier, the agreement study was conducted by selecting the best discriminating 15 items from an initial item pool and the a priori criterion of 67% was set. In any two methods, if 67% or above of the selected items were the same, these two methods were declared as selecting "best" discriminating items in common. In the reliability study, the reliability coefficients, $P_0$ and $K$, were tested separately, using the split plot SPF-3.3 factorial analysis of variance design. The item selection method was the within factor and the sample size was the between factor.

Results of the Agreement Study in Selecting the Same Best Discriminating Items

**Hypothesis One**

Hypothesis one states that these three item selection methods will select the same test items that best
Table 5

**Percentage of Overlap of Selected Test Items in Samples of 75 Subjects**

<table>
<thead>
<tr>
<th>Replication</th>
<th>$P(X_{qa}) &amp; \phi$</th>
<th>$P(X_{qa}) &amp; RM$</th>
<th>$\phi &amp; RM$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80%</td>
<td>27%</td>
<td>33%</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>3</td>
<td>73%</td>
<td>40%</td>
<td>27%</td>
</tr>
<tr>
<td>4</td>
<td>73%</td>
<td>33%</td>
<td>20%</td>
</tr>
<tr>
<td>5</td>
<td>67%</td>
<td>20%</td>
<td>27%</td>
</tr>
<tr>
<td>6</td>
<td>87%</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>7</td>
<td>60%</td>
<td>47%</td>
<td>33%</td>
</tr>
<tr>
<td>8</td>
<td>80%</td>
<td>13%</td>
<td>27%</td>
</tr>
<tr>
<td>9</td>
<td>53%</td>
<td>20%</td>
<td>33%</td>
</tr>
<tr>
<td>10</td>
<td>80%</td>
<td>33%</td>
<td>33%</td>
</tr>
</tbody>
</table>

Mean  
73.3%          28.6%          28.6%

**Note.** 67% is used as the criterion.

discriminate between masters and nonmasters in samples of 75 subjects. The data in Table 5 show that the average percentage of item overlap between the agreement and the phi coefficient approaches exceeds the a priori criterion of 67%; however, only 28.6% overlap exists between the agreement and the random selection methods, and between the phi coefficient and the random selection methods.

**Hypothesis Two**

Hypothesis two states that the three item selection methods will select the same test items that best discriminate between the masters and nonmasters in samples
of 150 subjects. The data in Table 6 show that the average percentage of item overlap between the agreement and the phi coefficient approaches exceeds the a priori criterion of 67%; however, only 28.8% and 25.9% exist between the agreement approach and the random selection method, and between the phi coefficient approach and the random selection method, respectively.

Table 6

Percentage of Overlap of Selected Test Items in Samples of 150 Subjects

<table>
<thead>
<tr>
<th>Replication</th>
<th>$P(X_{c}) \ &amp; \ \phi$</th>
<th>$P(X_{c}) \ &amp; \ \text{RM}$</th>
<th>$\phi \ &amp; \ \text{RM}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60%</td>
<td>27%</td>
<td>20%</td>
</tr>
<tr>
<td>2</td>
<td>87%</td>
<td>7%</td>
<td>7%</td>
</tr>
<tr>
<td>3</td>
<td>80%</td>
<td>27%</td>
<td>33%</td>
</tr>
<tr>
<td>4</td>
<td>87%</td>
<td>7%</td>
<td>13%</td>
</tr>
<tr>
<td>5</td>
<td>60%</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>6</td>
<td>73%</td>
<td>20%</td>
<td>13%</td>
</tr>
<tr>
<td>7</td>
<td>67%</td>
<td>47%</td>
<td>40%</td>
</tr>
<tr>
<td>8</td>
<td>60%</td>
<td>27%</td>
<td>40%</td>
</tr>
<tr>
<td>9</td>
<td>80%</td>
<td>53%</td>
<td>40%</td>
</tr>
<tr>
<td>10</td>
<td>87%</td>
<td>13%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Mean 74.1% 28.8% 25.9%

Note. 67% is used as the criterion.

**Hypothesis Three**

Hypothesis three states that the three item selection methods will select the same test items that best discriminate between masters and nonmasters in sample of 300
subjects. The data in Table 7 show that the average of percentage of item overlap between the agreement approach and the phi coefficient approach exceeds the a priori criterion of 67%; however, only 30% and 30.6% overlap exist between the agreement approach and the random selection method, and between the phi coefficient approach and the random selection method, respectively.

Table 7

Percentage of Overlap of Selected Test Items in Samples of 300 Subjects

<table>
<thead>
<tr>
<th>Replication</th>
<th>( P(X_e) &amp; \phi )</th>
<th>( P(X_e) &amp; RM )</th>
<th>( \phi &amp; RM )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80%</td>
<td>20%</td>
<td>33%</td>
</tr>
<tr>
<td>2</td>
<td>67%</td>
<td>33%</td>
<td>33%</td>
</tr>
<tr>
<td>3</td>
<td>80%</td>
<td>27%</td>
<td>33%</td>
</tr>
<tr>
<td>4</td>
<td>67%</td>
<td>40%</td>
<td>27%</td>
</tr>
<tr>
<td>5</td>
<td>87%</td>
<td>27%</td>
<td>33%</td>
</tr>
<tr>
<td>6</td>
<td>73%</td>
<td>33%</td>
<td>27%</td>
</tr>
<tr>
<td>7</td>
<td>60%</td>
<td>27%</td>
<td>33%</td>
</tr>
<tr>
<td>8</td>
<td>73%</td>
<td>33%</td>
<td>27%</td>
</tr>
<tr>
<td>9</td>
<td>73%</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>10</td>
<td>73%</td>
<td>20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Mean 72.6%  30%  30.6%

Note. 67% is used as the criterion.

Results of \( P_o \) analysis

**Hypothesis Four**

Hypothesis four states that three methods will yield tests with similar \( P_o \) values when the consistency of
mastery/nonmastery classification is measured. The results of testing this hypothesis are presented in Table 8 for the 15-item tests, Table 10 for the 25-item tests, and Table 12 for the 35-item tests.

Table 8

Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for $P_0$ in the 15-Item Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>2</td>
<td>0.0008</td>
<td>0.00004</td>
<td>0.73</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>0.0184</td>
<td>0.00055</td>
<td></td>
</tr>
<tr>
<td>Within Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>2</td>
<td>0.2105</td>
<td>0.10526</td>
<td>470.09 **</td>
</tr>
<tr>
<td>Method * Sample Size</td>
<td>4</td>
<td>0.0005</td>
<td>0.00013</td>
<td>0.58</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>0.0121</td>
<td>0.000224</td>
<td></td>
</tr>
</tbody>
</table>

** Significant at $p < .01$.

The data in Table 8 show that for the 15-item tests, a significant effect due to item selection method was found. In order to test the mean differences of $P_0$ values among item selection methods, Scheffé post hoc comparisons were performed. The results of these analyses are presented in Table 9.

The data in Table 9 show that the mean of $P_0$ values using the agreement approach is significantly different from the mean of $P_0$ values using the phi coefficient approach and
the random selection method at the .05 level. Also, the mean of $P_o$ values using the phi coefficient approach is significantly different from that using the random selection method.

Table 9

* Scheffé Post Hoc Comparisons for $P_o$ Mean Differences Among Item Selection Methods in the 15-Item Tests*

<table>
<thead>
<tr>
<th>Mean</th>
<th>Method</th>
<th>$P(X_c)$</th>
<th>$\phi$</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.84</td>
<td>$P(X_c)$</td>
<td>0.03 *</td>
<td>0.11 *</td>
<td></td>
</tr>
<tr>
<td>0.81</td>
<td>$\phi$</td>
<td></td>
<td>0.08 *</td>
<td></td>
</tr>
<tr>
<td>0.73</td>
<td>RM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant at $p < .05$.

The data in Table 10 show that for the 25-item tests, the effect of item selection method is significant at the .01 level. Scheffé post hoc comparisons were performed to further examine the mean differences of $P_o$ values among item selection methods. The results of these analyses are presented in Table 11.

The data in Table 11 show that the mean of $P_o$ values using the agreement approach is significantly different from the mean of $P_o$ values using the phi coefficient approach and the random selection method at the .05 level. Also, the
### Table 10

**Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for Pq in the 25-Item Tests**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>2</td>
<td>0.0016</td>
<td>0.0008</td>
<td>2.35</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>0.0092</td>
<td>0.00034</td>
<td></td>
</tr>
<tr>
<td>Within Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>2</td>
<td>0.1286</td>
<td>0.06430</td>
<td>315.20 **</td>
</tr>
<tr>
<td>Method * Sample Size</td>
<td>4</td>
<td>0.0019</td>
<td>0.00048</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>0.0110</td>
<td>0.000204</td>
<td></td>
</tr>
</tbody>
</table>

** Significant at p < .01.

### Table 11

**Scheffé Post Hoc Comparisons for Pq Mean Differences Among Item Selection Methods in the 25-Item Tests**

<table>
<thead>
<tr>
<th>Mean</th>
<th>Method</th>
<th>$P(X_q)$</th>
<th>$\Phi$</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.86</td>
<td>$P(X_q)$</td>
<td></td>
<td>0.03 *</td>
<td>0.09 *</td>
</tr>
<tr>
<td>0.83</td>
<td>$\Phi$</td>
<td></td>
<td></td>
<td>0.06 *</td>
</tr>
<tr>
<td>0.77</td>
<td>RM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant at p < .05.

The mean of $P_q$ values using the phi coefficient approach is significantly different from the mean of $P_q$ values using the random selection method.
The data in Table 12 show that for the 35-item tests, the effect of item selection method is significant at the .01 level. Scheffé post hoc comparisons were performed to examine the mean differences of $P_o$ values among item selection methods. The results of these analyses are presented in Table 13.

Table 12

Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for $P_o$ in the 35-Item Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>29</td>
<td>0.00846</td>
<td>0.000115</td>
<td>0.38</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>0.00823</td>
<td>0.000305</td>
<td></td>
</tr>
<tr>
<td>Within Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>60</td>
<td>0.04924</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method * Sample Size</td>
<td>2</td>
<td>0.04550</td>
<td>0.02325</td>
<td>501.08 **</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>0.00251</td>
<td>0.0000464</td>
<td></td>
</tr>
</tbody>
</table>

** Significant at $p < .01$.

The data in Table 13 show that the mean of $P_o$ values using the agreement approach is significantly different from the mean of $P_o$ values using the phi coefficient approach and the random selection method at the .05 level. Also, the mean of $P_o$ values using the phi coefficient approach is significantly different from that using the random selection method.
Table 13

Scheffé Post Hoc Comparisons for \( P_\phi \) Mean Differences Among Item Selection Methods in the 35-Item Tests

<table>
<thead>
<tr>
<th>Mean</th>
<th>Method</th>
<th>( P(X_\phi) )</th>
<th>( \phi )</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.86</td>
<td>( P(X_\phi) )</td>
<td></td>
<td>0.02 *</td>
<td>0.06 *</td>
</tr>
<tr>
<td>0.84</td>
<td>( \phi )</td>
<td></td>
<td>0.04 *</td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>RM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant at \( p < .05 \).

**Hypothesis Five**

Hypothesis five states that the resultant tests have similar \( P_\phi \) values when three different sample sizes are considered. The data in Tables 8, 10, and 12 show that an overall significant main effect due to sample sizes was not found for 15-, 25- and 35-item tests, respectively.

With reference to the relationship between the test length and reliability \( P_\phi \), the data from Table 9, 11, and 13 reveal that as the test length increases, the \( P_\phi \) values also increase.

**Results of \( K \) (Kappa) Analysis**

**Hypothesis Six**

Hypothesis six states that the three item selection methods will yield tests with similar \( K \) values when the
consistency of mastery/nonmastery classification is measured. The results for testing this hypothesis are presented in Table 14 for the 15-item tests, Table 19 for the 25-item tests and Table 21 for the 35-item tests.

Table 14
Results of Split Plot Factorial Analysis of Variance for K (Kappa) in the 15-Item Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>2</td>
<td>0.05550</td>
<td>0.02775</td>
<td>7.97 **</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>0.09386</td>
<td>0.00348</td>
<td></td>
</tr>
<tr>
<td>Within Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>2</td>
<td>0.24590</td>
<td>0.12295</td>
<td>108.61 **</td>
</tr>
<tr>
<td>Method * Sample Size</td>
<td>4</td>
<td>0.01470</td>
<td>0.003675</td>
<td>3.25 *</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>0.06113</td>
<td>0.001132</td>
<td></td>
</tr>
</tbody>
</table>
** Significant at p < .01.
* Significant at p < .05.

The data in Table 14 show that for the 15-item tests, the effects due to item selection method and sample size are significant at the .01 level and the effect due to interaction of method-by-sample size is significant at the .05 level. According to Kirk (1982), the analysis for simple main effects should be performed when there is significant interaction effect. The results of analysis of simple main effects due to item selection methods are shown in Table 15.
The data in Table 15 show that effects due to item selection methods are significant at the .15 level across sample sizes (according to Kirk (1982), the significance level of the omnibus test is .05 (method effect) + .05 (sample size effect) + .05 (interaction effect) = .15).

Table 15

Analysis of Variance Table With Tests of Simple Main Effects due to Item Selection Methods

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method at 75-Sample Size</td>
<td>2</td>
<td>0.13994</td>
<td>0.06997</td>
<td>61.81*</td>
</tr>
<tr>
<td>Method at 150-Sample Size</td>
<td>2</td>
<td>0.07325</td>
<td>0.03663</td>
<td>32.35*</td>
</tr>
<tr>
<td>Method at 300-Sample Size</td>
<td>2</td>
<td>0.04742</td>
<td>0.02371</td>
<td>20.95*</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>0.06113</td>
<td>0.001132</td>
<td></td>
</tr>
</tbody>
</table>

* Significant at p < .15.

Because simple main effects were significant at the .15 level, Scheffé post hoc comparisons were performed to further examine the mean differences of K values among item selection methods in a given sample size. The results of these analyses are presented in Table 16 for the 75-sample size, Table 17 for the 150-sample size, and Table 18 for the 300-sample size.

The data in Table 16 show that the mean differences of K values between any two item selection methods are significant at the .15 level for the sample size of 75.
Table 16

Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods at the 75-Sample Size

<table>
<thead>
<tr>
<th>Mean</th>
<th>Method</th>
<th>Mean Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.59</td>
<td>( \Phi )</td>
<td>0.06 *</td>
</tr>
<tr>
<td>0.53</td>
<td>( P(X_c) )</td>
<td>0.16 *</td>
</tr>
<tr>
<td>0.43</td>
<td>( RM )</td>
<td>0.10 *</td>
</tr>
</tbody>
</table>

* Significant at \( p < .15 \).

The data in Table 17 show that the mean differences of K values between any two item selection methods are significant at the .15 level for the sample size of 150.

Table 17

Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods at the 150-Sample Size

<table>
<thead>
<tr>
<th>Mean</th>
<th>Method</th>
<th>Mean Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.52</td>
<td>( \Phi )</td>
<td>0.06 *</td>
</tr>
<tr>
<td>0.46</td>
<td>( P(X_c) )</td>
<td>0.12 *</td>
</tr>
<tr>
<td>0.40</td>
<td>( RM )</td>
<td>0.06 *</td>
</tr>
</tbody>
</table>

* Significant at \( p < .15 \).
The data in Table 18 show that mean difference of $k$ values between the phi coefficient approach and the random selection method is significant at the .15 level.

Table 18

Scheffé Post Hoc Comparisons for $k$ (Kappa) Mean Differences Among Item Selection Methods at the 300-Sample Size

<table>
<thead>
<tr>
<th>Mean</th>
<th>Method</th>
<th>$\Phi$</th>
<th>$P(X_{C})$</th>
<th>RM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.51</td>
<td>$\Phi$</td>
<td></td>
<td>0.05</td>
<td>0.09 *</td>
</tr>
<tr>
<td>0.46</td>
<td>$P(X_{C})$</td>
<td></td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>0.42</td>
<td>$\text{RM}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Significant at $p < .15$.

The graphic illustrations of Figure 3 and of Figure 4 provide the information about the interaction of method-by-sample size. These two figures illustrate that as sample size increases, the mean differences of $k$ values produced by these three item selection methods tend to decrease.

The data in Table 19 show that for the 25-item tests, the effect of item selection method is significant at the .01 level and there is no significant interaction effect. Scheffé post hoc comparisons were performed to investigate this effect further. The results of these analyses are presented in Table 20.

The data in Table 20 show that the mean of $k$ values using the phi coefficient approach is significantly
Figure 3. Illustration of the relationship between item selection methods and sample sizes in the 15-item tests.

Figure 4. Illustration of the relationship between sample sizes and item selection methods in the 15-item Tests.
Table 19

**Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for K (Kappa) in the 25-Item Tests**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>2</td>
<td>0.01558</td>
<td>0.00779</td>
<td>1.74</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>0.12087</td>
<td>0.00448</td>
<td></td>
</tr>
<tr>
<td>Within Factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>2</td>
<td>0.11683</td>
<td>0.058415</td>
<td>95.14**</td>
</tr>
<tr>
<td>Method * Sample Size</td>
<td>4</td>
<td>0.00456</td>
<td>0.001140</td>
<td>1.86</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>0.03315</td>
<td>0.000614</td>
<td></td>
</tr>
</tbody>
</table>

** Significant at p < .01.

different from the mean of K values using the agreement approach and the random selection method. Also, the mean of K values using the agreement approach is different from that using the random selection method.

Table 20

**Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods in the 25-Item Tests**

<table>
<thead>
<tr>
<th>Mean</th>
<th>Method</th>
<th>Mean Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.58</td>
<td>$\Phi$</td>
<td>$P(X_{\Phi})$ 0.06 * 0.09 *</td>
</tr>
<tr>
<td>0.52</td>
<td>$P(X_{\Phi})$</td>
<td></td>
</tr>
<tr>
<td>0.49</td>
<td>$RM$</td>
<td>0.03 *</td>
</tr>
</tbody>
</table>

* Significant at p < .05.
The data in Table 21 show that for the 35-item tests, the effect of item selection method is significantly different at the .01 level and there is no significant interaction effect. Scheffé post hoc comparisons for main effect due to item selection method were further performed. The results of these analyses are presented in Table 22.

Table 21

Results of Split Plot Factorial Analysis of Variance Among Item Selection Methods for K (Kappa) in the 35-Item Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between Factor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>29</td>
<td>0.06815</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>0.00113</td>
<td>0.000565</td>
<td>0.23</td>
</tr>
<tr>
<td><strong>Within Factor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>60</td>
<td>0.072063</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method * Sample Size</td>
<td>4</td>
<td>0.05864</td>
<td>0.02932</td>
<td>131.83 **</td>
</tr>
<tr>
<td>Error</td>
<td>54</td>
<td>0.001413</td>
<td>0.000353</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01201</td>
<td>0.0002224</td>
<td></td>
</tr>
</tbody>
</table>

** Significant at p < .01.

The data in Table 22 show that the mean of K values using the phi coefficient approach is significantly different from the mean of reliability K using the agreement approach and the random selection method at the .05 level. Also, the mean of K values using the agreement approach is different from that using the random selection method.
Table 22

Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods in 35-Item Tests

<table>
<thead>
<tr>
<th>Mean Differences</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \Phi )</td>
</tr>
<tr>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>

* Significant at \( p < .05 \).

The data in Table 23 reveal that as the test length increases, the K values also increase.

Table 23

Relationship Between the Mean of K (Kappa) Values and Test Length

<table>
<thead>
<tr>
<th>Methods</th>
<th>Test Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
</tr>
<tr>
<td>( \Phi )</td>
<td>0.54</td>
</tr>
<tr>
<td>( P(X_c) )</td>
<td>0.48</td>
</tr>
<tr>
<td>( RM )</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Hypothesis Seven

Hypothesis seven states that the resultant tests will have similar K values when different sample sizes are
considered. The results of testing this hypothesis indicate that a significant effect due to sample size was found in the 15-item tests, but not in the 25- and 35-item tests (see Tables 15, 19, and 21). Because there was a significant interaction effect in the 15-item tests, the test of simple main effects of sample size in a given item selection method was performed. The results are presented in Table 24.

Table 24

Analysis of Variance Table With Tests of Simple Main Effects due to Sample Sizes

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Square</th>
<th>Mean Square</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size at $P(X_\geq)$</td>
<td>2</td>
<td>0.03085</td>
<td>0.01542</td>
<td>8.06 **</td>
</tr>
<tr>
<td>Sample Size at $\phi$</td>
<td>2</td>
<td>0.03642</td>
<td>0.01821</td>
<td>9.52 **</td>
</tr>
<tr>
<td>Sample Size at RM</td>
<td>2</td>
<td>0.00294</td>
<td>0.00147</td>
<td>0.77</td>
</tr>
<tr>
<td>Error</td>
<td>81</td>
<td>0.15499</td>
<td>0.001914</td>
<td></td>
</tr>
</tbody>
</table>

* * Significant at $p < .15$.

The data in Table 24 show that there are significant simple main effects at the agreement and the phi coefficient approaches but not at the random selection method. Scheffé post hoc comparisons were performed to further examine the mean differences of $k$ values among the sample sizes. The results of these analyses are presented in Tables 25 and 26.
The data in Table 25 show that the mean differences of K values are significant between the 75- and 150-sample sizes and between the 75- and 300-sample sizes but not significant between the 150- and 300-sample sizes at the phi coefficient approach.

Table 25

Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Sample Sizes at the Phi Coefficient Approach

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sample Size</th>
<th>Mean Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.59</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>0.52</td>
<td>150</td>
<td>0.07 * 0.08 *</td>
</tr>
<tr>
<td>0.51</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

* Significant at $p < .15$.

Table 26

Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Sample Sizes at the Agreement Approach

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sample Size</th>
<th>Mean Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.53</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>0.46</td>
<td>150</td>
<td>0.07 * 0.07 *</td>
</tr>
<tr>
<td>0.46</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

* Significant at $p < .15$. 
The data in Table 26 show that the mean differences of K values are significant between the 75- and 150-sample sizes and between the 75- and 300-sample sizes but not between the 150- and 300-sample sizes in the agreement approach.
Chapter References

CHAPTER V

SUMMARY, FINDINGS, DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS FOR FURTHER RESEARCH

Summary

In the present study, the agreement approach, the phi coefficient approach, and the random selection method were compared in terms of the magnitude of agreement in selecting the best discriminating test items and the resultant test reliabilities. Test responses from 1,836 students were used in this study. The 50-item instrument was a district-developed criterion-referenced test and was used to measure the skills in physical science at high school level.

Three different sample sizes (75, 150, and 300) and three different numbers of items (15, 25, and 35) were examined, for a total 9 combinations. Each combination was replicated 10 times, for a total of 90 distinct data sets. The sample for each data set was randomly selected with replacement from the population. The agreement approach and the phi coefficient approach selected their individual "best" items on the basis of the size of discriminating value, while the random selection method selected items randomly and without replacement from the initial item pool. A cut-off score was set at 80% of the items correct.
In the agreement study, the best 15 discriminating items were used, then the average percentage of item overlap over 10 replications between any two of these item selection methods in each sample size was compared with the a priori criterion 67%. In any two methods, if 67% or more of the selected items were the same, these two methods were declared as selecting similar test items. In the reliability study, the coefficients $P_0$ and $K$ were computed over 90 distinct data sets. Huynh's beta-binomial method was used to compute these coefficients. A split plot SPF-3.3 factorial analysis of variance was performed, in which the sample size was the between factor and the item selection method was the within factor. The main effects due to item selection method, sample size, and interaction of method-by-sample size were tested individually. The factor of test length was not considered in the design because there was item dependency for the selected test items. However, the mean of coefficients $P_0$ and $K$ for each test length were compared to investigate whether longer tests generated higher reliability coefficients.

Findings

**Hypotheses One, Two and Three**

Hypotheses one, two, and three state that the three item selection methods will select the same test items that
best discriminate between masters and nonmasters in samples of 75, 150, and 300 subjects, respectively. The results of this study rejected these hypotheses. In this study, the random selection method failed to select items similar to either the agreement approach or the phi coefficient approach in any of the samples. As for the agreement and the phi coefficient approaches, they selected most of the items in common across sample sizes although (a) there were 6 of 30 replications in which the percentage of item overlap was lower than the a priori 67% criterion; and (b) one was even as low as 53% (see Tables 5, 6, and 7).

**Hypothesis Four**

Hypothesis four states that the three methods will yield tests with similar $P_0$ when the consistency of mastery/nonmastery classification is measured. For $P_0$ analysis, this hypothesis was rejected. Significant main effects due to item selection method were found at the .01 level across test lengths. Among these three-item selection methods, the agreement approach produced tests with the highest $P_0$ values, and the random selection method produced tests with the lowest ones.

**Hypothesis Five**

Hypothesis five states that the resultant tests will have similar $P_0$ values when three different sample sizes are
considered. For $P_0$ analysis, this hypothesis was retained because no significant sample size effects were found across three test lengths.

**Hypothesis Six**

Hypothesis six states that the three item selection methods will yield similar $K$ values when the consistency of mastery/nonmastery classifications is measured. For $K$ analysis, hypothesis six was rejected. However, the findings were different from those in $P_0$ analysis. In $K$ analysis, the phi coefficient approach produced tests with the highest $K$ values rather than the agreement approach.

Significant effects due to interaction of method-by-sample size across test lengths were not found in the $P_0$ analysis. For $K$ analysis, significant interaction effect was found in the 15-item tests but not in the 25- or 35-item tests. The analysis of simple main effects in the 15-item tests was performed. The results indicated that significant effects due to item selection method across three sample sizes were found. Scheffé post hoc comparisons were further examined. The results revealed that (a) at 75- and 150-sample sizes, the mean of $K$ values between any two item selection methods were significantly different from each other; and (b) at 300-sample size, only the mean of $K$ values in the phi coefficient approach was significantly different from that in the random selection method (see Tables 16, 17, and 18).
Hypothesis Seven

Hypothesis seven states that the resultant tests will have similar $K$ values when three sample sizes are considered. For $K$ analysis, the results of testing this hypothesis were different from those of $P_0$ analysis. For $P_0$ analysis, no significant sample size effects were found across three test lengths. For $K$ analysis, hypothesis seven was retained for the 25- and 35-test items because no significant sample size effects were found. However, for the 15-item tests, significant effects due to sample size and due to interaction of method-by-sample size were found. The analysis of simple main effects due to sample sizes was performed. The results indicated that significant effects due to sample size were found for the phi coefficient approach and the agreement approach but not for the random selection method. Scheffé post hoc comparisons were further examined. The results for the phi coefficient and the agreement approaches showed that the mean of $K$ values in the 75-sample size was significantly different from those in the 150- and 300-sample sizes (see Tables 25 and 26).

In addition, the relationship between the item selection method and test length was examined. The results in both the $P_0$ and $K$ analyses were the same; that is, as test length increased, the $P_0$ and $K$ values also increased in these item selection methods.
Discussion

Based on the results of the present study, several issues are discussed as follows.

**Discussion of the Random Selection Method**

Popham (1978), Hambleton (1982), and Hambleton and Gruijter (1983) suggested the use of the random selection method to construct a test since all test items in the criterion-referenced tests were developed on the basis of a carefully defined domain of tasks, and all items were homogeneous as well as interchangeable. The results of the present study revealed, however, that the random selection method did not produce tests with either $P_\theta$ or $K$ values similar or superior to those produced by the agreement approach or the phi coefficient approach. Similar findings were also found in the studies conducted by Crehan (1974), Haladyna and Roid (1983), and Smith (1978). These researchers had compared the random selection method with other methods, and they concluded that the effect of random selection method was not similar or superior to other methods. The reason, as some researchers have argued, may be that the items in the item pool still have some unwanted properties, which occur in the item writing stage and can only be detected through the use of item analysis (Saupe, 1966; Lange et al. 1967; Hsu, 1971; Smith, 1978; Haladyna & Roid, 1981; van de Linden, 1981).
**Discussion of Differential Results Between P₀ and K Analyses**

In this study, the results of \( P₀ \) analysis were different from those of \( K \) analysis in two ways: (a) the agreement approach produced tests with highest \( P₀ \) values in the \( P₀ \) analysis while the phi coefficient approach produced tests with highest \( K \) values in the \( K \) analysis; and (b) no significant effects due to sample size and due to interaction of method-by-sample size were found across test lengths in the \( P₀ \) analysis but significant effects due to sample size and due to interaction of method-by-sample size at the 15-item tests were found in the \( K \) analysis.

Examination of the literature showed that such differential results had been noted by researchers. Millman (1974), Berk (1984), and Subkoviak (1980, 1984) had stated that coefficients \( P₀ \) and \( K \) did not reflect the same aspect of the measurement of reliability. According to these researchers, \( P₀ \) and \( K \) are sensitive to different types of consistency and have a converse relationship. As noted earlier, \( P₀ \) is defined as the proportion of consistent classification (master/master and nonmaster/nonmaster) across two tests. According to Berk (1984) and Subkoviak (1980, 1984), coefficient \( P₀ \) measures overall consistency; that is, it measures the total consistency of mastery/nonmastery classifications across tests for "whatever reasons". \( K \) is defined as the proportion of consistent decisions corrected
for chance because the chance-expected proportions agreement (\(P_{\text{C}}\)) are subtracted from the overall proportions (\(P_{\text{O}}\)). Therefore, Berk (1984) and Subkoviak (1984) stated that \(K\) measures the test's contribution to the overall proportion of consistent classification. In other words, \(K\) reflects mastery/nonmastery decision consistency with the elimination of "expected chance".

Both \(P_{\text{C}}\) and \(K\) are sensitive to the selected cut-off score, test length, and score variability (Berk, 1984; Subkoviak, 1984). However, \(K\)'s sensitivity to the cut-off score is the reverse of \(P_{\text{C}}\) (Berk, 1984; Subkoviak, 1984). For \(P_{\text{C}}\), higher values of \(P_{\text{C}}\) correspond to cut-off scores at the tails of the score distribution and lower values correspond to the cut-off scores near the mean score. \(K\) does just the opposite. That is, higher values of \(K\) correspond to cut-off scores near the mean score, and lower values correspond to cut-off scores at the tail of the score distribution.

**Discussion of Selecting \(P_{\text{O}}\) Analysis or \(K\) Analysis**

Because coefficients \(P_{\text{O}}\) and \(K\) do not measure exactly the same thing and have a converse relationship to each other, researchers like Subkoviak (1984) and Berk (1984) suggested that the choice between \(P_{\text{O}}\) and \(K\) depend upon how the consistency is defined in a given decision and how the properties of the statistic are evaluated.
Berk (1984) said that $P_o$ should be used for tests when an absolute cut-off score was selected and for tests that had short subtests and/or produced low score variance. Tests developed for the classroom usually contain a small number of test items and yield low score variance; therefore, Berk (1984) recommended $P_o$ for use in classroom test construction and decision making.

Brennan and Prediger (1981) stated that the use of $K$ was appropriate only when the marginal frequencies were fixed a priori. When the marginal frequencies are free to vary, the "chance" term ($P_c$) in $K$ should be replaced by $1/n$ ($n$ is the number of categories in the contingency table); otherwise, the use of $K$ might lead to incorrect conclusions. According to Livingston and Wingersky (1979), when fixed marginals are used, the cut-off score represents a relative rather than an absolute value. In other words, the cut-off score is determined by the consequences of passing/failing a specified proportion of the examinees. Therefore, Berk (1984) claimed that $K$ should be used when a relative cut-off score was selected. For many competency tests developed at the district and state levels, the cut-off scores are adjusted in accordance with the consequences of passing particular number of students. Consequently, Berk (1984) recommended $K$ be used for test construction and decision making at the district and state levels.
Discussion of Effect of Sample Size

As noted in Chapter I, Swezey (1981) suggested that at least 50 percent more subjects were needed in the tryout sample than items in the initial item pool. Nunnally and Wilson (1975) said that at least 5 times as many subjects as items were necessary to avoid taking advantage of chance. In the present study, samples of 75, 150, and 300 subjects were used, representing 1 1/2, 3, and 7 times of the initial 50 items. Significant effects due to sample size were, however, not found except in the 15-item tests of the K analysis. When considering the factors affecting both the $P_0$ and $K$, it becomes obvious why no sample size effects, in general, were found in this study. That is, both $P_0$ and $K$ are sensitive to the selected cut-off score, test length, and score variability, but not to sample size (Berk, 1984; Subkoviak, 1984).

Conclusions

Based on the findings of this study, the random selection method (a) did not select "best" items in common with those selected by the agreement or the phi coefficient approaches, and (b) produced tests with the lowest reliability of mastery/nonmastery classifications, when compared with the agreement and the phi coefficient
approaches. Researchers like Hsu (1971), Smith (1978), and van de Linden (1981), for instance, stated that empirical item analysis is absolutely necessary for constructing criterion-referenced test items even though items were selected from a well-defined domain of tasks. The results of this study support this statement.

The results of $P_0$ and $K$ analyses were different in terms of recommending one selection method over another. These results are consistent with the statements provided by Berk (1984) and Subkoviak (1980, 1984). That is, $P_0$ analysis has a converse relationship with the $K$ analysis.

The phi coefficient approach produced tests with higher $K$ reliability coefficients than the agreement approach and the random selection method did. The significant effects due to sample size and interaction of method-by-sample size were found in the 15-item tests but not in the 25- and 35-item tests. Hambleton et al. (1983) said that test length could affect the reliability of mastery/nonmastery classifications. In this study, the results of $K$ analysis were consistent with these researchers' statement. That is, as test length increased, the values of $K$ also increased.

Researchers indicate that $K$ analysis is not appropriate when an absolute cut-off score is selected and the results may be misinterpreted when marginal frequencies are not fixed a prior (Berk, 1984; Brennan & Prediger, 1981).
The agreement approach produced tests with higher $P_0$ reliability coefficients than the phi coefficient approach and the random selection method did. Significant effects due to sample size and due to interaction of method-by-sample size were not found across test lengths. Similar to the results of $K$ analysis, there was a positive relationship between the test length and $P_0$ values, that is, longer tests had higher $P_0$ values. This result is also consistent with the statement provided by Hambleton et al. (1983).

As noted earlier, for $P_0$, the agreement approach is recommended over the phi coefficient approach; while for $K$, the phi coefficient approach is recommended over the agreement approach. The choice between the $P_0$ and $K$ depends on whether an absolute or relative cut-off score is set and/or whether the tests are administered at the classroom levels or at the district and state levels. When the tests are administered at the classroom settings and absolute cut-off scores are used, then $P_0$ should be used. In such a case, the agreement approach is recommended because this approach yields highest $P_0$ values in this study. Otherwise, the phi coefficient approach is recommended.

**Recommendations for Further Research**

Both $P_0$ and $K$ analyses are sensitive to the selected cut-off score, test length and score variability (Berk,
1984; Subkoviak, 1984). The present study only deals with one cut-off score (80% of the items correct) and one score variability (one test); as a result, the effects due to cut-off scores and/or score variabilities for recommending one method over another are unknown. It is, therefore, recommended that in future studies, more than one cut-off score and/or different score variabilities be used; thus, the statistical properties of these three item selection methods can be understood better.
Chapter References


APPENDIX A
January 18, 1988

Hui-Fen Lin
P.O. Box 7959 NTSU
Denton TX, 76203

Dear Ms. Lin:

This is in answer to your inquiry concerning why three replications were used in my dissertation study, and why 67% was used as my a priori criterion. To the best of my recollection, replications were needed to compute a within-cell error term for my design, and although more replications may have been desirable, three seemed adequate for my purposes. The 67% was somewhat arbitrary, but, besides being a common rule-of-thumb, it was thought that a 2 of 3 overlap was adequate for the reduction in computation intensity that would result from using P(X_o) over I(θ). The criterion could be set much higher if a larger proportion of the item pool were to be selected. I hope this is helpful.

I am enclosing an abstract from a Midwestern ERA convention that also dealt with this topic. I haven't seen it published anywhere, but you might write to the author for a copy if you are interested.

I would be interested in learning what you find out. Good luck!

Sincerely,

Deborah Harris
Research Psychometrician
Measurement Research
DH/sjl
Enclosure
program Tests (Input,Output,Infile,Outfile,Pofile1,Pofile2,  
    Pofile3,Kfile1,Kfile2,Kfile3);
Const  Students = 1835; (* total students *)
    Items  = 49;  (* the test items *)
    Set1  = 74; (* first set of students *)
    Set2  = 149; (* second set of students *)
    Set3  = 299; (* third set of students *)
    GoodItem = 34; (* the selected item *)
    MaxiSet = 299;
Type   Scores = array [0..Students, 0..Items] of char;
    Questions = array [0..Items] of integer;
Datatype = record
    Questions;
    integer;
    Answer : integer;
end;
SetOfItem = set of 0..49;
P_Phi_Record = record
    Item : integer;
    Value : real;
end;
P_Phi_Array = array [0..Items] of P_Phi_Record;
var
    Infile,  
    Outfile,  
    Pofile1,  
    Pofile2,  
    Pofile3,  
    Kfile1,  
    Kfile2,  
    Kfile3 : text;
    Seed,  
    PItems,  
    PhiItems,  
    I, J : integer;
    Test : Scores;
    SetData,  
    Master,  
    NonMaster : array [0..MaxiSet] of Datatype;
    CutScore : array [0..MaxiSet] of integer;
    Item : array [1..3] of integer;
    RanItem,  
    P,  
    Phi : P_Phi_Array;
    Swapped : boolean;
SetItem : array [0..Items] of integer;
SubsetScore: array [0..MaxiSet] of integer;
ItemSet : SetOfItem;
F_value : array [0..GoodItem+1,0..GoodItem+1]
of real;

(*-------------------------------------------------------------------------*)
(* Reads data from a input file. *)
(*-------------------------------------------------------------------------*)

procedure ReadData;
begin
  I := 0;
  while I <= Students do begin
    for J := 0 to Items do
      read (Infile, Test[I,J]);
    readln (Infile);
    I := I + 1;
  end;
end;

(*-------------------------------------------------------------------------*)
(* A function for generating random number. *)
(*-------------------------------------------------------------------------*)

function Random (Min,Max: integer): integer;
const Modulus = 65536;
var Range  : integer;
function Rannum : integer;
const Multiplier = 25173;
Increment = 13849;
begin
  Seed := (Multiplier * Seed + Increment) mod Modulus;
  Rannum := Seed;
end;
begin
  Range := Max - Min + 1;
  Random := Rannum * Range div Modulus + Min;
end;

(*-------------------------------------------------------------------------*)
(* Randomly select 75-, 150-, and 300-sample sizes *)
(*-------------------------------------------------------------------------*)

procedure RandomDataMatrix (Setnum: integer);
var Student : integer;
begin
  for I := 0 to Setnum do begin
    Student := Random (0,Students);
    SetData[I].Score := 0;
    for J := 0 to Items do
      if Test[Student,J] = '0' then
        SetData[I].Answer[J] := 0
      else begin
        SetData[I].Answer[J] := 1;
        SetData[I].Score := SetData[I].Score + 1;
      end;
  end;
procedure EachItemScore (Setnum: integer);
begin
  for I := 0 to Items do begin
    SetItem[I] := 0;
    for J := 0 to Setnum do
      if SetData[J].Answer[I] = 1 then
        SetItem[I] := SetItem[I] + 1;
  end;
end;

(* Calculates the cut off score and store the score to the master or nonmaster matrix. *)
procedure CutoffScore (Setnum: integer);
var K, L,
    Cutscore: integer;
begin
  writeln(Outfile,'*** Master & NonMaster');
  Cutscore := round((Items + 1) * 0.8);
  K := 0;
  L := 0;
  for I := 0 to Setnum do begin
    if SetData[I].Score >= Cutscore then begin
      for J := 0 to Items do
        Master[K].Answer[J] := SetData[I].Answer[J];
      Master[K].Score := SetData[I].Score;
      writeln(Outfile,'Master Score',master[K].score);
      K := K + 1;
    end
    else begin
      for J := 0 to Items do
        NonMaster[L].Answer[J] := SetData[I].Answer[J];
      NonMaster[L].Score := SetData[I].Score;
      writeln(Outfile,'NonMaster Score',nonmaster[L].score);
      L := L + 1;
    end;
    end;
  for J := 0 to Items do begin
    Master[K].Answer[J] := 9999;
    NonMaster[L].Answer[J] := 9999;
  end;
end;
(* Calculates the P(Xc) and Phi value. *)

procedure GetP_Phi_Value;
var K, L,
    N1, N2, N3, N4 : integer;
    Temp : real;
begin
    for I := 0 to Items do begin
        K := 0;
        L := 0;
        N1 := 0;
        N2 := 0;
        N3 := 0;
        N4 := 0;

        while Master[K].Answer[I] <> 9999 do begin
            if Master[K].Answer[I] = 1 then
                N1 := N1 + 1
            else
                N3 := N3 + 1;
            K := K + 1;
        end;

        while NonMaster[L].Answer[I] <> 9999 do begin
            if NonMaster[L].Answer[I] = 1 then
                N2 := N2 + 1
            else
                N4 := N4 + 1;
            L := L + 1;
        end;

        P[I].Item := I;
        P[I].Value := (N1 + N4) / (N1 + N2 + N3 + N4);
        Temp := sqrt((N1+N2) * (N3+N4) * (N1+N3) * (N2+N4));
        Phi[I].Item := I;
        if Temp = 0 then
            Phi[I].Value := -999
        else
            Phi[I].Value := ((N1*N4)-(N2*N3)) / Temp;
    end;
end;

(* Swaps the data. *)

procedure Swap (var Datal, Data2: P_Phi_Record);
var Temp1 : real;
    Temp2 : integer;
begin
    Temp1 := Datal.Value;
    Datal.Value := Data2.Value;
    Data2.Value := Temp1;
    Temp2 := Datal.Item;
    Datal.Item := Data2.Item;
end;
Data2.Item := Temp2;
end;

(*---------------------------------------------*)
(* Sorts the P(Xc) value in descending order.  *)
(*---------------------------------------------*)
procedure Sort_P_Value (SelectItem: integer);
begin
I := 0;
Swapped := true;
while (I <= Items) and Swapped do begin
  J := Items;
  Swapped := false;
  while J > I do begin
    if P[J].Value > P[J-1].Value then begin
      Swapped := true;
      Swap (P[J],P[J-1]);
    end;
    J := J - 1;
  end;
  I := I + 1;
end;
writeln (Outfile, ' The Item & Value of P');
writeln (Outfile, ' =====================•');
writeln (Outfile, ' Item  Value');
writeln (Outfile);
for J := 0 to SelectItem do
  writeln (Outfile, P[J].Item,'  ',P[J].Value);
end;

(*---------------------------------------------*)
(* Sorts Phi value in descending order.  *)
(*---------------------------------------------*)
procedure Sort_Phi_Value (SelectItem: integer);
begin
I := 0;
Swapped := true;
while (I <= Items) and Swapped do begin
  J := Items;
  Swapped := false;
  while J > I do begin
    if Phi[J].Value > Phi[J-1].Value then begin
      Swapped := true;
      Swap (Phi[J],Phi[J-1]);
    end;
    J := J - 1;
  end;
  I := I + 1;
end;
writeln (Outfile, ' The Item & Value of Phi');
writeln (Outfile, ' =====================•');
writeln (Outfile, ' Item  Value');
writeln (Outfile);
for J := 0 to SelectItem do
    writeln (Outfile, Phi[J].Item, ', Phi[J].Value);
end;
(*---------------------------------------------------------------------*)
(* The random selection method selects items. *)
(*---------------------------------------------------------------------*)
procedure RandomItem (SelectItem: integer);
var Item : integer;
begin
    Item := random (0,Items);
    I := 0;
    Itemset := [];
    while (I <= SelectItem) do begin
        while Item in Itemset do
            Item := random (0,Items);
        Itemset := Itemset + [Item];
        Ranitem[I].Item := Item;
        I := I + 1;
    end;
end;
(*---------------------------------------------------------------------*)
(* Compute the overlap among Phi, P(Xc), and random selection method.*)
(*---------------------------------------------------------------------*)
procedure ComputeOverlap (SelectItem: integer);
var Count1, Count2, Count3, Percent1, Percent2, Percent3 : integer;
begin
    writeln(Outfile, ' Compute overlap');
    writeln(Outfile, ' ===============');
    writeln(Outfile);
    writeln(Outfile, ' Item ', ' Phi Item ', ' Random Item');
    writeln(Outfile, ' ', ' ', ' ');
    for I := 0 to SelectItem do
        writeln (Outfile, P[I].Item, Phi[I].Item, Ranitem[I].Item);
    Count1 := 0;
    Count2 := 0;
    Count3 := 0;
    for I := 0 to SelectItem do begin
        J := 0;
        while (P[I].Item <> Phi[J].Item) and
            (J <= SelectItem) do
            J := J + 1;
        if P[I].Item = Phi[J].Item then
Count1 := Count1 + 1;
for I := 0 to SelectItem do
  if P[I].Item in Itemset then
    Count2 := Count2 + 1;
for I := 0 to SelectItem do
  if Phi[I].Item in Itemset then
    Count3 := Count3 + 1;
Percent1 := round((Count1 / (SelectItem+1)) * 100);
Percent2 := round((Count2 / (SelectItem+1)) * 100);
Percent3 := round((Count3 / (SelectItem+1)) * 100);
writeln(Outfile,' Overlap between p & Phi --
  Count# , countl,percent1,'%');
writeln(Outfile,' Overlap between p & random
  item ' , -- Count# , count2,percent2,'%');
writeln(Outfile,' Overlap between Phi & random
  item ', -- Count#, count3,percent3,'%');
end;
procedure Get_Po_K_Value (Student,SelectItem,A:integer;
Itemnum:P_Phi_Array);
procedure SelectedltemScore(Student,Selectltem:integer;
Itemnum:P_Phi_Array);
begin
go:
SubSetScore[I] := 0;
for J := 0 to SelectItem do begin
  K := Itemnum[J].Item;
  SubSetScore[I] := SubSetScore[I] +
    SetData[I].Answer[K];
end;
end;
procedure ComputeStatistic (Student,SelectItem: integer);
begin
  var Temp1,
      Mean,
      ScoreSquare,
      Alpha,
      Alpha21,
      Beta,
      Variance : real;
  Temp2,
  X, Y,
  TotalTestScore : integer;

for I := 0 to (GoodItem+1) do 
    for J := 0 to (GoodItem+1) do 
        F_value[I,J] := 0;
TotalTestScore := 0;
for I := 0 to Student do 
    TotalTestScore := TotalTestScore + SubSetScore[I];
Mean := TotalTestScore / (Student + 1);
writeln(Outfile,'Mean is ',mean);
ScoreSquare := 0;
for I := 0 to Student do 
    ScoreSquare := ScoreSquare + sqr(SubSetScore[I]);
Variance := (ScoreSquare / Student) -
    (sqr(TotalTestScore)) / 
    (Student * (Student+1));
writeln(Outfile,'Variance is ',variance);
Tempi := Mean * (SelectItem+1-Mean) / ((SelectItem+1) * Variance);
Alpha21 := ((SelectItem+1) / SelectItem) * (1-Tempi);
writeln(Outfile,'Alpha 21 is ',alpha21);
Alpha := (-1+1/Alpha21) * Mean;
writeln(Outfile,'Alpha is ',alpha);
Beta := -Alpha + ((SelectItem+1)/Alpha21) -
    (SelectItem+1);
writeln(Outfile,'Beta is ',beta);
Temp2 := 2 * (SelectItem+1);
F_value[0,0] := 1;
for I := 1 to Temp2 do 
    F_value[0,0] := F_value[0,0] * ((Temp2 + Beta - I) / 
    (Temp2 + Alpha + Beta - I));
for X := 0 to SelectItem do begin 
    F_value[X+1,0] := F_value[X,0] * (SelectItem+1-X) *
    (Alpha+X+0 / ((X+1) * (Temp2 + Beta - X - 0 - 1));
    F_value[0,X+1] := F_value[X+1,0];
end;
for Y := 1 to (SelectItem+1) do 
    for X := 0 to SelectItem do 
        F_value[X+1,Y] := F_value[X,Y] *
        (SelectItem+1-X) *
        (Alpha+X + Y / ((X + 1) *
        (Temp2 + Beta - X - Y - 1));
end;
(*-----------------------------------------------------------------------*)
(* Computes values of Po and K. *)
(*-----------------------------------------------------------------------*)
procedure ComputePo_K (SelectItem, A: integer); var Cutindex : integer;
Pnn, Pmm, Pna, Pnb,
Pma, 
Pmb, 
Po, 
Pc, 
K 
: real;
begin
Cutindex := round((SelectItem+1) * 0.8);
Pnn := 0;
Pmm := 0;
Pna := 0;
Pnb := 0;
Pma := 0;
Pmb := 0;
for I := 0 to (Cutindex-1) do 
  for J := 0 to (Cutindex-1) do 
    Pnn := Pnn + F_value[I,J];
for I := Cutindex to (SelectItem+1) do 
  for J := Cutindex to (SelectItem+1) do 
    Pmm := Pmm + F_value[I,J];
Po := Pnn + Pmm;
for I := 0 to (SelectItem+1) do 
  for J := 0 to (Cutindex-1) do 
    Pna := Pna + F_value[I,J];
for I := 0 to (SelectItem+1) do 
  for J := Cutindex to (SelectItem+1) do 
    Pma := Pma + F_value[I,J];
for I := 0 to (CutIndex-1) do 
  for J := 0 to (SelectItem+1) do 
    Pnb := Pnb + F_value[I,J];
for I := Cutindex to (SelectItem+1) do 
  for J := 0 to (SelectItem+1) do 
    Pmb := Pmb + F_value[I,J];
Pc := Pna * Pnb + Pma * Pmb;
K := (Po - Pc) / (1 - Pc);
writeln(Outfile, 'Po is ',po);
 writeln(Outfile, 'Pc is ',pc);
 writeln(Outfile, 'K is ',k);
Po := (round(P0 * 100)) / 100;
K := (round(K * 100)) / 100;
case A of
1: begin
write(Pofile1, P0:5:2);
write(Kfilel, K:5:2);
end;
2: begin
write(Pofile2, P0:5:2);
write(Kfile2, K:5:2);
end;
3: begin
write(Pofile3, P0:5:2);
write(Kfile3, K:5:2);
begin
  SelectedItemScore (Student, SelectItem, ItemNum);
  ComputeStatistic (Student, SelectItem);
  ComputePo_K (SelectItem, A);
end; (* Get Po_K Value *)
(* Procedure of writing results of Po and K to outfiles. *)
procedure EachItemSet (Student, SelectItem, A: integer;
                         Px, Phix, Ranx: P_Phi_Array);
begin
  ComputeOverlap (SelectItem);
  writeln (Outfile, ' The Po & K value of P');
  Get_Po_K_Value (Student, SelectItem, A, Px);
  writeln (Outfile, ' The Po & K value of Phi');
  Get_Po_K_Value (Student, SelectItem, A, Phix);
  writeln (Outfile, ' The Po & K value of Random Item');
  Get_Po_K_Value (Student, SelectItem, A, Ranx);
  case A of
    1: begin
       writeln(Pofile1);
       writeln(Kfile1);
    end;
    2: begin
       writeln(Pofile2);
       writeln(Kfile2);
    end;
    3: begin
       writeln(Pofile3);
       writeln(Kfile3);
    end;
  end;
end;
(* Overall procedure for samples of 75 subjects. *)
procedure SetlProcess;
var A, B : integer;
begin
  for A := 1 to 3 do begin
    for B := 1 to 10 do begin
      writeln (Outfile, 'Setl ', B);
      RandomDataMatrix (Setl);
      EachItemScore (Setl);
      CutoffScore (Setl);
      GetP_Phi_Value;
      Sort_P_Value (Item[A]);
      Sort_Phi_Value (Item[A]);
RandomItem (Item[A]);
EachItemSet (Set1, Item[A], A, P, Phi, RanItem);
end;
writeln(Pofile1);
writeln(Pofile2);
writeln(Pofile3);
writeln(Kfile1);
writeln(Kfile2);
writeln(Kfile3);
end;

(*---------------------------------------------------------------*)
(* Overall procedure for samples of 150 subjects.  *)
(*---------------------------------------------------------------*)

procedure Set2Process;
var A, B : integer;
begin
  for A := 1 to 3 do
    for B := 1 to 10 do begin
      writeln (Outfile, 'Set2 B');
      RandomDataMatrix (Set2);
      EachItemScore (Set2);
      CutOffScore (Set2);
      GetP_Phi_Value;
      Sort_P_Value (Item[A]);
      Sort_Phi_Value (Item[A]);
      RandomItem (Item[A]);
      EachItemSet (Set2, Item[A], A, P, Phi, RanItem);
    end;
  writeln(Pofile1);
  writeln(Pofile2);
  writeln(Pofile3);
  writeln(Kfile1);
  writeln(Kfile2);
  writeln(Kfile3);
end;

(*---------------------------------------------------------------*)
(* Overall procedure for samples of 300 subjects.  *)
(*---------------------------------------------------------------*)

procedure Set3Process;
var A, B : integer;
begin
  for A := 1 to 3 do
    for B := 1 to 10 do begin
      writeln (Outfile, 'Set3 B');
      RandomDataMatrix (Set3);
      EachItemScore (Set3);
      CutOffScore (Set3);
      GetP_Phi_Value;
      Sort_P_Value (Item[A]);
      Sort_Phi_Value (Item[A]);
      RandomItem (Item[A]);
EachItemSet (Set3, Item[A], A, P, Phi, RanItem):
    end;
    writeln(Pofile1);
    writeln(Pofile2);
    writeln(Pofile3);
    writeln(Kfile1);
    writeln(Kfile2);
    writeln(Kfile3);
end;

begin (* main program *)
    open (Infile, 'test.dat', history := old);
    Reset (Infile);
    Rewrite (Outfile);
    Rewrite (Pofile1);
    Rewrite (Pofile2);
    Rewrite (Pofile3);
    Rewrite (Kfile1);
    Rewrite (Kfile2);
    Rewrite (Kfile3);
    ReadData;
    Seed := 57832;
    Item[1] := 14;
    Item[3] := 34;
    Set1Process;
    Set2Process;
    Set3Process;
    end.

This program was written by Miss Wenjen Yu.
BIBLIOGRAPHY

Books


Articles


**ERIC Documents**


Unpublished Materials


