AN EMPIRICAL INVESTIGATION OF THE EFFECT OF ITEM SELECTION TECHNIQUES ON ACHIEVEMENT TEST CONSTRUCTION

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

Richard Clair Cox

1964

This is to certify that the

thesis entitled

AN EMPIRICAL INVESTIGATION OF THE EFFECT
OF ITEM SELECTION TECHNIQUES ON
ACHIEVEMENT TEST CONSTRUCTION

presented by

RICHARD CLAIR COX

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education

Willard D. Harrington
Major professor

Date 8-3-64

O-169

ABSTRACT

AN EMPIRICAL INVESTIGATION OF THE EFFECT
OF ITEM SELECTION TECHNIQUES ON
ACHIEVEMENT TEST CONSTRUCTION

by Richard Clair Cox

Problem

The subject-matter content and instructional objectives to be evaluated are identified by the test constructor. Items which measure these instructional objectives in each content area are created by the item writer. Usually more items than will be used in the final form of the instrument are written. Items for the final form of the instrument are commonly selected on the basis of the statistical analysis of the item pool. In order that the final form of the evaluation instrument validly measures the objectives identified in the original item pool the method of item selection should not have an appreciable effect on the structure of the final instrument as compared with the structure of the item pool from which the test items were selected. This study investigates the effect that the statistical item selection has on the structure of the final form of a test as compared with the original item pool.

Procedure

An item pool of 379 multiple-choice natural science items was identified. This item pool was described with consideration to the average difficulty and discrimination levels of the items computed for a male and female tryout group. The classification of each item according to the instructional objective being measured was also examined. From this item pool the 100 most discriminating items identified by the Davis index and the 100 most discriminating items identified by the Difference index were selected to form two 100 item tests. This was done separately for males and females. The entire procedure was repeated using data obtained from high and low achieving male and female tryout groups.

The 100 item tests were compared to the total item pool with respect to the classification of items according to the instructional objective being measured. The comparisons were made separately for the tests constructed using the data from the male and female groups and the high and low achieving groups.

Findings

1. Statistical selection of items from the total item pool has a biasing effect on the selected tests. The proportion of items in the selected tests which measure certain instructional objectives is unlike the proportion of items in the total item pool which measure the same objectives.

2. Statistical selection of items from the total item pool appears to operate differentially for male and female groups. The structure of the selected tests as indicated by the taxonomical structure of the items differs for male and female groups.

3. Statistical selection of items from the total item pool appears to operate differentially for high and low achieving tryout groups, both male and female. The structure of the selected tests as indicated by the taxonomical classification of items differs for high and low achieving groups.

4. The statistical selection of items from the total item pool operates similarly with respect to the taxonomical structure of the items selected no matter which of the two discrimination indices are used as the criterion for selection.

AN EMPIRICAL INVESTIGATION OF THE EFFECT OF ITEM SELECTION TECHNIQUES ON ACHIEVEMENT TEST CONSTRUCTION

By

Richard Clair Cox

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

COLLEGE OF EDUCATION

Department of Foundations of Education

1964

ACKNOWLEDGEMENTS

The investigator would like to acknowledge his gratitude to Dr. Willard Warrington, the thesis director, and to Dr. Bernard Corman, the guidance committee chairman, for their assistance in the preparation of the thesis and for their helpful guidance through the doctoral program. Appreciation is also extended to Dr.Jean M. LePere and Dr. Paul Bakan for their assistance and helpful criticism.

The investigator is indebted to the Office of Evaluation Services, Michigan State University, which not only made possible the collection of data,but also assisted the investigator in his efforts.

The investigator is particularly grateful to his wife, Cynthia, and son, Kevin, for their moral support and patience and for the occasional neglect they have endured.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The validity of educational evaluation depends on the extent of correspondence between educational objectives and the instruments intended to evaluate these objectives. It is desirable that there be a relationship between these educational objectives and evaluation procedures. It is the purpose of this investigation to examine some commonly used test construction procedures which could have some serious consequences for this relationship.

Educational Achievement Test Construction

According to Lindquist, the construction of an educational achievement test consists of the following five major steps:

1. Planning the test
2. Writing the test exercises
3. Trying out the test in preliminary form and assembling the finished test after tryout
4. Determining the procedures and preparing the manuals for administering and scoring the test
5. Reproducing the test and accessory materials (33:119)

This investigation will be concerned with only the first three steps outlined above. These are the stages in test construction where the correspondence between instructional objectives and

the objectives measured by the final form of the test could be diminished.

Planning the Test

Initial planning of an educational achievement test involves the identification of the subject-matter content and the instructional objectives to be tested. A two-way table of specifications is often utilized at this stage to insure that these two aspects of test content are represented. (48:161-162) (24:50) Nelson (40:117-119) presents some examples of such two-way tables in which the Taxonomy of Educational Objectives (5) is used as the classification system for the instructional objectives.

In each cell of the two-way table the number of items to be written is specified. This provides the item writer with a guide to insure adequate representativeness of the content and instructional objectives which are to be tested. This investigation will be primarily concerned with the instructional objectives specified in the planning stage of test construction.

Instructional Objectives

The Taxonomy of Educational Objectives (5) is a comprehensive attempt to analyze cognitive objectives in a meaningful classification system. According to the authors "It is intended to provide for classification of the goals of our educational system. It is expected to be of general help to all teachers, administrators, professional specialists, and research workers who deal with curricular and evaluation problems." (5:1)

The Taxonomy of Educational Objectives arranges instructional objectives from the simple to the complex. The major categories of this classification system, arranged in order of increasing complexity, are designated as Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation.* The Taxonomy of Educational Objectives thus provides a framework for the classification of the instructional objectives which can be utilized in the planning stage of test construction.

Writing the Test Item

As discussed above, the two-way specification table is used in the test planning stage to insure that the composition of the test is representative of the instructional objectives and the content areas to be evaluated. The specification table provides the item writer with a guide to the number of items to be written for measuring the instructional objectives in each content area. Within this framework the item writer can create test items using a variety of test forms.

Ebel (15:193-204) describes four commonly employed forms of "objective"** test items: the short-answer, the true-false, the multiple-choice, and the matching forms. The multiple choice form,***

---

*For a more detailed description of these categories see the Condensed Version of the Taxonomy of Educational Objectives.

**An objective test item is one which will be scored in the same manner regardless of the individual scoring it. This type of item is in contrast to an essay or free-response item which requires subjective judgment on the part of the scorer.

***The multiple-choice test item presents to the subject an introductory statement or question followed by several responses. From these responses the subject must select the most appropriate or best answer.

because of the demand for machine-scorable tests, is probably the most widely used type of test in educational achievement testing today.

Since the multiple-choice format is frequently used it has been the object of several recent attacks on testing. (51) (4) (26) Hoffmann (25) and Guilford (20) indicate that the multiple-choice format is not suitable to test certain types of educational objectives. In short, the critics feel that multiple-choice questions can measure little more than memory or recall of specific facts; they are not suited to the measurement of more complex objectives.

On the other hand, authorities in the test construction field have not only emphasized that the multiple-choice question can be more than a simple recall exercise but have also presented items which they consider examples of thought provoking and insightful questions. (37) (14) (18) In this study it will be assumed that it is possible to write multiple-choice test items which evaluate educational objectives of varying complexity.

The system of classification presented in the Taxonomy of Educational Objectives can be used to classify test items as well as educational objectives. (13) (43) (46) Such a classification requires a knowledge of the prior educational experiences of the individuals taking the test. In order to make an accurate classification of a test item it is necessary to know, or assume, the learning experiences which have preceded the administration of the test. (5:51) Stanley and Bolten (46) present evidence that the classification of test items into the major categories of the Taxonomy of Educational Objectives can be accomplished with a considerable degree of reliability.

The Taxonomy of Educational Objectives thus provides a framework for the classification of instructional objectives and the test items intended to measure these objectives. It is the task of the test item writer to attain a high degree of correspondence between the type of item he creates and the instructional objective to be measured. It will be assumed for this study that the item writer has conscientiously attempted to maintain this high degree of correspondence in the original item pool constructed.*

Experimental Tryout and Assembly of Final Test Form

The usual procedure in objective test construction is to prepare a larger pool of test items than will be used in the final form of the instrument. This large pool of items is administered to a tryout group in order to obtain certain information that will aid in the selection of items for the final form of the test. Conrad (9:250-251) lists seven purposes which may be served by the experimental tryout, several of which stress the statistical aspects of the test items. These statistical descriptions of test items are commonly called item analysis techniques or item selection techniques.

Item Selection Techniques

Guilford (19) states that the first major objective of an item analysis is to obtain objective information about the test

---

*This is a crucial assumption in the test construction procedure. The validity of the evaluation instrument depends to a great extent upon this correspondence. If the assumption is violated the final instrument could not possibly be a valid measure of the desired instrumental objectives. If the assumption is met, it is still possible that the measuring instrument is not valid. This later condition is the concern of the present study.

items which have been written for a test. This objective information indicates to the test constructor ways in which the quality of the final test form might be improved. Item analysis involves the computation of difficulty and discrimination indices for each item administered to the tryout group.

The difficulty index of an item is simply an indication of how difficult the particular item is for the group taking the test. The most obvious index is the per cent of the tryout group that passes the item. (12:267) If this percentage is high the item is not very difficult; if it is low the difficulty level of the item is high.

"A discrimination index is a measure of the extent to which students who are judged to be good in terms of some standard succeed on the item, and those who are judged to be poor on the same standard, fail it." (16) The "standard" in the above definition is often called the criterion. The criterion may be the total test score, in which case the discrimination index will reflect the extent to which an individual item measures the same characteristic as does the total test. (12:286) The total test score is commonly used as the criterion for discrimination indices since it is readily available and because valid external criteria are often difficult to identify.

Suggested Item Discrimination Techniques*

At least fifty different techniques of item discrimination have been suggested. Several reviews of some of these techniques appear in the literature. Long and Sandiford (34) list and describe twenty-three item-validity techniques. Vernon (49) reviews eighteen discrimination indices which he classifies according to the way the criterion variable is treated. Davis (12) describes ten methods of expressing item discrimination power and also provides an excellent bibliography concerning techniques of item analysis. Guilford (19) describes twenty-five indices of item discrimination.

Classical probability theory provides the basis for the item selection techniques described in Solomon. (45) Statistical multivariate analyses and non-parametric solutions are treated in depth.

Other techniques have been suggested by Gulliksen,(22:380-382) Michael,(36) Webster, (50) Clemens, (7) Colver, (8) Levine and Lord, (32) Bryden, (6) and Scott. (44)

Studies Comparing Item Discrimination Methods

The use of a particular item discrimination technique often depends upon the assumptions the test constructor is willing to make.

---

*An extensive review of some suggested item discrimination techniques and studies concerning these techniques serves three purposes. First, it provides some indication of the multitude and variety of techniques available to the test constructor. Secondly, it serves to provide a current review of the literature pertinent to item discrimination techniques. Finally, the review will provide the rationale for selecting the particular techniques used in this study.

There are, however, many indices suited for identical situations. Trying to determine which is the best item discrimination index to use in any given situation has been the purpose of many investigations.

Barthelmess (3) evaluated ten item validity methods. The correlation ratio and biserial r had the highest average inter-correlations with all the other methods. When the methods were evaluated by correlation with two external criteria the correlation ratio, Long and McCall methods rank highest. The Vincent and biserial r were found to be the most reliable of the methods evaluated. Barthelmess (3) recommends the use of biserial r when a dichotomous method is called for and the correlation ratio when a multiple method should be used.

For difficult items, biserial r was found to be the most reliable and stable of five discrimination indices compared by Cook.(10) Cook's Index B was found to be the best overall index included in this study.

Lentz, et al, (31) compared the reliabilities and inter-correlations of four item discrimination indices. They conclude that the upper vs. lower third method is superior to the other three methods.

A comprehensive study by Long and Sandiford (34) evaluated thirteen techniques of item discrimination. They found that methods which select items of fifty per cent difficulty appear to be superior to those methods which do not; that the upper vs. lower technique is preferable when effectiveness and ease of computation are the primary considerations; and that the better techniques differ little in effectiveness.

Swineford (47) recommends the Holzinger and difference between means techniques on the basis of reliability and ease of computation. Eight methods were compared in this study.

Adkins (2) found little difference between ten discrimination techniques used to select test items. Scholastic indices were used as the validation criteria.

Pinter and Forlano (41) report high inter-correlations between nine methods of item selection methods. They conclude that it makes little difference which method is used to select the most internally consistent items.

Guilford and Lacey (21) compared six indices of item discrimination and found that test items were in approximately the same rank order no matter which index was used. The main difference was that the point biserial r and the Phi coefficient tend to select items of moderate difficulty while the other methods examined seemed uninfluenced by item difficulty.

Humphreys (27) has compared the Phi coefficient, Flanagan's r, biserial r, point biserial r and the Tetrachoric coefficient. Point biserial r was found to be the most representative index on the basis of intercorrelations with the other indices. With the exception of the Tetrachoric coefficient the reliabilities of the methods were comparable.

Lawshe and Mayer (30) found that a high proportion of the same items were selected by the Davis index and D-Values when these two indices were used as the criteria for selection.

The reliabilities of items selected by D-Values and Phi coefficients were compared by Mason.(35) No significant differences between the two techniques were found.

Vernon (49) computed rank order correlations between six methods of item discrimination. All correlations were fairly high. D-Values and the Davis index were almost identical. All of the indices except the per cent method had comparable reliabilities.

Kuang (29) compared the reliabilities of biserial r, the Davis index and probit analysis. He found that the three methods select a high percentage of the same items. He concludes that any differences between the three methods are insignificant.

Adams (1) investigated the effect of item difficulty level upon the reliability of nine item analysis techniques. He concludes that the upper vs. lower twenty-seven per cent, Phi, biserial r, point biserial r, and t-ratio and the Davis index all have comparable reliabilities at all difficulty levels. Intercorrelation of the nine methods revealed that point biserial r and the t-ratio were most representative of the techniques.

The stability of item discrimination indices for groups of different ability levels was investigated by Hall. (23) The Flanagan index, Davis index, Gulliksen index and the difference index were all sensitive to changes in the ability level of the sample. Hall concluded that experimental items should be administered

to subjects whose mean ability level is similar to the population for whom the test is intended.

In general these comparative studies indicate that there are few differences in the end results obtained by using any of the better item discrimination techniques. The main difference seems to be that some techniques favor items of moderate difficulty while others are not influenced by the difficulty level of the item. Ease of computation is often suggested as the criterion for determining which method to use.

Assembly of Final Test Form

After the test has been planned, the items written and statistically analyzed using a tryout group, the task for the test constructor now becomes that of selecting items for the final form of the test.

The difficulty and discrimination indices described above are often used for the selection of items for the final test form. There is not general agreement as to how these indices should be employed.

One important consideration is the nature and the purpose of the test under consideration. Many educational achievement tests are power tests in which the examiner is interested in how many items the subject is able to answer correctly and not how rapidly he can work. These tests are designed to rank the subjects on some specified characteristic;thus, the difficulty and discrimination

level of the items becomes a crucial consideration. For other types of tests difficulty and discrimination levels become less important.

After a review of relevant literature Adams (1) concludes that selection of items at the fifty per cent level of difficulty yields items of maximum validity and discrimination. Saupe (42) indicates that in a practical situation it would be difficult to find a set of items all of which cluster around the fifty per cent level of difficulty. Faced with this problem the test constructor will select those items which are the most discriminating for inclusion in the final form of the test.*

Statement of the Problem

The major steps of test construction up to and including the assembly of the final instrument have been discussed. The subject-matter content and the instructional objectives to be tested are identified by the test constructor. Items which measure these instructional objectives in each content area are created by the item writer. Usually more items than will be used in the final form of the instrument are written. Items for the final form of the instrument are commonly selected on the basis of the statistical analysis of the item pool after administration to a tryout group.

---

*Because of the relationship between item difficulty and certain discrimination indices the selection of the most discriminating items will in general yield items in the middle difficulty ranges.

Many item discrimination indices which are used in the selection of test items for the final instrument have been suggested. Considerable research has been directed to the comparison of these item discrimination techniques with respect to reliability, difficulty level and ease of computation. The effect of item selection techniques on the intended structure of a test as indicated by the composition of the original item pool has apparently received little attention.

The Taxonomy of Educational Objectives (5) provides a framework for the classification of test items with respect to the instructional objectives the items are intended to measure. It is possible, therefore, to classify the items in the original item pool according to the instructional objectives they are designed to measure. In order that the final form of the evaluation instrument validly measures the objectives identified in the original item pool the method of item selection should not appreciably alter the structure of the original item pool. The structure of the final form of a test constructed by selection of items from the original item pool should be similar to the structure of the original item pool. If the item selection procedure biases the final test form by disproportionate selection of items which measure certain instructional objectives the final form of the test will not validly evaluate the objectives measured by the total item pool.

It is the purpose of this study to investigate the effect of statistical item selection on the structure of the final evaluation instrument as compared with the structure of the original item pool. This effect of statistical item selection will be examined with respect to the sex and achievement level of the tryout group.

While there has been considerable research comparing item discrimination techniques with respect to reliability, difficulty level and ease of computation, little evidence is available concerning the effect that the sex of the tryout group might have on these techniques. If, for the items in the original item pool, the values assigned by a discrimination index vary for male and female groups the items selected for the final form of a test could differ depending on the composition of the tryout group. The possibility that males or females do better on certain types of items and that consequently the items discriminate differently for males and females has been overlooked in the comparisons of various item discrimination techniques.

Another variable to be examined is the achievement level of the tryout group. If the values assigned by a discrimination index vary for high and low achieving groups the items selected for the final form of a test could vary if the tryout group was composed of a disproportionate number of high or low achieving students.

The possibility that differences in sex or achievement level have an effect on the values assigned by discrimination indices could have far reaching implications for test construction. A differential effect on discrimination indices by either sex or achievement level would bias the final form of the test in favor of a particular group. For this reason the present investigation will examine the item selection procedure taking into account the sex and achievement level of the tryout group.

CHAPTER II

PROCEDURE

The basic problem of this investigation is to examine the effect of item selection techniques on the structure of the final test form. The present chapter describes the examinations and the subjects used in the study. It also describes the classification of test items according to the instructional objectives they are designed to measure. Finally, the chapter describes the methods of item selection employed and the investigation of the basic problem.

Examinations and Subjects Used in the Study

The examinations used in this study are the end-of-term examinations used in the introductory natural science sequence at Michigan State University.* The fall and winter term examinations have 125 multiple-choice questions; the spring term examination has 129 such items. These three examinations were combined to form the total item pool under consideration in this study. This was deemed necessary in order to have enough items so that there

*Due to the security involved the examinations *per se* will not be presented in this study. Some of the types of items involved are presented in Nelson (38) (39) (40:117-119). These examinations were selected for two major reasons. First, a large part of the responsibility for the assembly of the examinations rests with Clarence H. Nelson, a national figure in science test writing. Secondly, the examinations are constructed with specific instructional objectives in mind.

would be an adequate representation of items classified as measuring each general instructional objective.*

Items for the examinations were submitted by the instructors in the natural science courses. An examining committee comprised of a group of these instructors reviewed the items and constructed the final examinations.

The reliability of each exam was computed using the Kuder-Richardson formula #21. The reliability coefficients for the three terms were .91, .86 and .85, respectively. These reliability coefficients are probably underestimates of the actual reliabilities due to the assumption of equal item difficulty made in the calculation of the formula.

Although a time limit was imposed, the examinations were essentially power tests and the majority of subjects responded to every item. The examinations were scored using the IBM 805 scoring machine. The test papers were first checked for overmarking, that is, more than one answer for each question. Those few papers with overmarks were excluded from the study. The remaining papers were scored and double-checked on the scoring machine. The score on a test was the number of correct responses, i.e. there was no penalty for guessing.

---

*See Table 1.

The subjects which were used in the study are those students who in the 1962-63 school year were enrolled in the introductory natural science sequence at Michigan State University and took the final examination at the end of all three terms. The majority of these students were college freshmen and sophomores.

The term-end examinations were taken by 5,028 students in the fall term, 4,249 students in the winter term and 3,369 students in the spring term. Of these, 3,150 students were identified as having taken all three term-end examinations. Those groups of students who were eliminated from the study either because they were not enrolled (and did not receive actual instruction but did take the final examinations for credit) or because they had not taken all three examinations were examined to see if their exclusion would have any biasing effect on the study.

One group of students eliminated from the present sample were those who had taken the examinations in order to obtain credit for the course without attending instructional periods. The effect of their exclusions was to eliminate several of the top scores of the total distribution since, for the most part, these students were of above average ability.

A second group eliminated from the sample were those students who did not complete the entire sequence due to academic failure. The effect of their exclusion was to eliminate several of the bottom scores of the total distribution. The overall effect of eliminating these two groups was to restrict the range of possible scores slightly.

The 3,150 students were separated by sex. From these groups 1,000 males and 1,000 females were selected at random. These are the two major groups used in the study.

Classification of Test Items*

The 379 items in the total test were classified using the categories of the Taxonomy of Educational Objectives.(5) Three judges worked independently on this classification using the examples presented in the Taxonomy of Educational Objectives and in Nelson. (40:117-119) There was agreement as to the category in which a particular item belonged for approximately eighty per cent of the items.

Such a classification of test items requires knowledge of the learning situations which have preceded the test. (5:51) Two of the judges were not completely aware of the actual presentation of subject matter to the subjects; hence, the discrepancies in the classification of test items. After consultation with the subject matter expert agreement was reached as to the proper classification of all 379 test items. Table 1 presents the number and percentage of test items classified in the major categories of the Taxonomy of Educational Objectives. None of the items were classified in the Synthesis and Evaluation categories. Multiple-choice items which fall in these categories are extremely rare.

---

*Some examples of the classification of natural science items in each of the four categories of the Taxonomy of Educational Objectives are presented in Appendix A.

TABLE 1. Frequency of Test Items Classified in Taxonomical Categories

Category	Number of Items	Per cent of Total
Knowledge	102	27
Comprehension	110	29
Application	91	24
Analysis	76	20
Total	379	100

Selection of Item Discrimination Indices

As discussed in Chapter I, a variety of item discrimination indices have been proposed. Some of these indices are no longer in use today, having been modified or discarded after critical examination. Studies which compare discrimination indices indicate that the major considerations when selecting discrimination indices should be the effect of item difficulty and the ease of computation.

The two indices to be used in this study are the Davis index and the Difference index. The Davis index is theoretically uninfluenced by item difficulty while the Difference index tends to assign maximum values to items of middle difficulty. Both are relatively easy to compute.

Davis Index*

The biserial r is a correlation coefficient between each test item and the total test score. Davis (11:9) states that biserial r is probably the most satisfactory measure of relationship available when the total test score is used as the criterion of discrimination.

Kelley (28) has demonstrated that the upper and lower twenty-seven per cent of the total test scores provide a good approximation to biserial r which involves considerably less computation. Flanagan (17) constructed a table from which the value of correlation coefficients between a test item and the total test score can be read directly. The table is entered by identifying the percentage of subjects in the upper and lower twenty-seven per cent who have passed the item. The tabled values are essentially uninfluenced by the difficulty of the item.

Davis transformed the coefficients in the Flanagan table into discrimination indices. (11:11-15) The Davis indices range from 0 (no discrimination) to 100 (perfect discrimination). Like the Flanagan coefficients, they are theoretically uninfluenced by item difficulty and can be read from the table with the knowledge of the percentage of subjects in the upper and lower twenty-seven per cent of the total test scores which passed the item.

---

*For a complete description of the Davis index see Davis (11).

The Difference Index

The Difference index is defined by Hall (23) as the difference between the percentage of subjects in the upper and lower per cent of the total test scores who have answered the item correctly. The percentage answering the item correctly in the lower group is subtracted from the percentage answering the item correctly in the upper group.

The Difference indices range from 0 (no discrimination) to 100% (perfect discrimination). Unlike the Davis indices, the Difference indices are influenced by item difficulty. Items of median difficulty (50%) will generally be assigned higher values than will either very easy or very difficult items.

Procedure

After the samples of 1,000 males and females were selected, the distributions of total test scores were tabulated. Table 2 presents these distributions for males and females indicating the number of scores that fall in each ten point interval.

TABLE 2. Distribution of Total Test Scores for Males and Females

Total Score	Males	Females
330-339	1	1
320-329	1	1
310-319	4	2
300-309	10	14
290-299	24	29
280-289	42	39
270-279	67	72
260-269	84	80
250-259	102	72
240-249	130	107
230-239	118	112
220-229	102	126
210-219	96	95
200-209	95	94
190-199	51	64
180-189	37	51
170-179	23	26
160-169	10	9
150-159	2	3
140-149	1	3
Total	1,000	1,000

A further comparison was made by computing the means and standard deviations for the two distributions. These figures are presented in Table 3.

TABLE 3. Means and Standard Deviations of Total Test Scores for Males and Females

Group	Mean	Standard Deviation
Males	235.86	31.03
Females	233.70	32.42

On the average, the males have achieved slightly higher scores on the combined tests. This is the usual pattern in the natural science examinations given at Michigan State University.

The upper and lower 270 subjects (27%) in each distribution were identified in order to compute indices of item difficulty and discrimination. The index of difficulty for a particular item was determined by the percentage of subjects in the upper and lower 27 per cent of the total test scores who passed the item. Thus, a difficulty index of 100 indicates that all the students in the combined upper and lower groups passed the item; an index of 50 indicates that half of the students in the combined upper and lower groups passed the item, etc.

The average difficulty level of items for males was 62.10 and for females was 61.79. The test items were on the average slightly easier for males. This was indicated previously by the slightly higher male achievement on the total test.

Davis indices and Difference indices were computed for all 379 items for both males and females. The average values of these indices for both groups appear in Table 4.

TABLE 4. Average Davis and Difference Indices by Sex

Item Discrimination Method	Males	Females
Davis Index	15.66	17.04
Difference Index	20.28	21.34

The average Difference indices are higher than the average Davis indices for both males and females. The Difference index will,in general, assign higher values to items near the middle ranges of difficulty than will the Davis index. The higher average values for females are also a reflection of difficulty levels.

In order to simulate the assembly of a final test form, the 100 items with the highest Davis indices were selected from the total item pool. The procedure was repeated using the Difference indices as the criteria for item selection. These procedures were followed for both the male and female groups.

The 100 items selected by each of the two techniques were compared with respect to the number of items classified according to instructional objectives, the average difficulty level and the average discrimination level. The same comparisons were made between the male and female groups and between the 100 item tests and the total 379 items.

As a supplement to the major analysis the entire procedure was repeated for samples of high and low achieving males and females as defined by the total test scores. Those males and females who had the top 270 (27%) scores on the total test were defined as high achievers. The bottom 270 were defined as low achievers. These groups are the upper and lower 27 per cent of the total test which were used in the computation of difficulty and discrimination indices for the total group.

The means and standard deviations of the test scores in the high and low achieving groups for males and females are presented in Table 5. The high achieving females have, on the average, achieved higher scores than high achieving males. The reverse is true for the low achieving groups. The higher overall achievement by males (see Table 3) seems to result from higher scores by males in the middle and lower ranges of the distribution.*

---

*In the discussion of students eliminated from the study it was indicated that the upper end of the total distribution was slightly restricted due to the elimination of those students who had taken the final examination without taking the actual course work. This is a provision allowed by the waiver system at Michigan State University. In the natural science sequence male students take greater advantage of this provision than do females. The upper end of the distribution of test scores for males is therefore more restricted than is that for females. This is true since these students often obtain the higher scores on the test.

TABLE 5. Means and Standard Deviations of High and Low Achievers by Sex

		Mean	Standard Deviation
High Achievers	Males	274.12	14.26
	Females	275.06	14.26
Low Achievers	Males	197.34	13.86
	Females	193.83	13.84

Difficulty and discrimination indices were computed for the high and low achieving groups in the manner described previously. The upper and lower 27 per cent of these high and low groups included 73 subjects. The average values for the difficulty and discrimination indices are presented in Table 6.

TABLE 6. Average Difficulty and Discrimination Indices for High and Low Achievers by Sex

		Difficulty Index	Davis Index	Difference Index
High Achievers	Males	72.73	8.83	9.03
	Females	72.98	9.09	9.01
Low Achievers	Males	51.22	6.38	8.83
	Females	50.59	6.47	8.91

Once again it is evident that the Davis and Difference indices differ with varying levels of average item difficulty. For the high achieving groups only slight differences between the two discrimination indices are apparent. For the low achieving groups, for which the average item difficulties are close to the 50 per cent level, the test items have, on the average, been assigned higher values by the Difference index.

As before, the 100 most discriminating items as indicated by the Davis and the Difference indices were selected from the total item pool. The same comparisons described above were made. The results of the analysis for the high and low achieving groups were compared with the results obtained using the entire distribution.

CHAPTER III

RESULTS AND DISCUSSION

The purpose of this investigation is to examine the effect of item selection techniques on the strucutre of the final form of a test. From the original item pool smaller tests were constructed by selecting the most discriminating items identified by two different item discrimination indices. These smaller tests represent the final form of the test which are to be compared with the original item pool. This chapter presents and discusses the results obtained from the comparisons made between the structure of the original item pool, as indicated by difficulty level, discrimination level and taxonomical classification of items, and the structure of the tests selected by the discrimination indices. The results are reported for males and females in general and for high and low achieving males and females.

Description of Total Item Pool for Males and Females*

The original item pool consists of 379 multiple-choice items. The number and percentage of items classified in each taxonomical category were presented in Table 1. Knowledge items accounted for

*In order to adequately discuss the results presented in this chapter it seems necessary at this point to summarize and elaborate upon the characteristics of the original item pool.

27 per cent of the total pool, Comprehension 29 per cent, Application 24 per cent and Analysis 20 per cent. These values should be closely approximated by the percentage of items classified in a similar manner for the tests selected by the discrimination indices if these selected tests are to be representative of the original item pool.

For males, the average score on the 379 items was 235.86 with the range of scores going from 146 to 330. Appendix B presents the difficulty and discrimination indices computed on each item for the male group. These values are presented separately for each taxonomical category. The average difficulty level of the items was 62.10. The average Davis index was 15.66 while the average Difference index was 20.28. Table 7 presents the average difficulty and discrimination indices by taxonomical category for the male group.

TABLE 7. Average Difficulty and Discrimination Indices by Taxonomical Category for Total Item Pool - Males

Taxonomical Category	Difficulty Index	Davis Index	Difference Index
Knowledge	65.98	15.25	19.28
Comprehension	63.74	17.16	22.32
Application	59.15	15.24	19.98
Analysis	58.04	14.54	19.01

The values of the average difficulty and discrimination indices differ with taxonomical category. On the average, Knowledge and Comprehension type items have a higher difficulty level than does the total item pool; average difficulty levels for Application and Analysis type items are smaller than the value for the total item pool. Knowledge and Comprehension items are easier on the average than all the items while Application and Analysis are more difficult.

The values of the average discrimination indices for males also differ with taxonomical category. Both the Davis and Difference indices indicate that Comprehension type items discriminate better on the average than the remaining types of items. Analysis items are the least discriminating as indicated by the average values.

For the female group the average score on the 379 items was 233.70 with a range of scores from 142 to 337. Appendix C presents the difficulty and discrimination indices computed for each item using the female scores. The average difficulty level was 61.79; the average Davis index was 17.04 and the average Difference index was 21.34. The 379 items were more difficult for the female group and subsequently discriminated better for females than for males. Table 8 presents the average difficulty and discrimination indices by taxonomical category for the female group.

TABLE 8. Average Difficulty and Discrimination Indices by Taxonomical Category for Total Item Pool - Females

Taxonomical Category	Difficulty Index	Davis Index	Difference Index
Knowledge	66.39	16.83	20.18
Comprehension	63.45	18.08	23.22
Application	58.37	17.08	21.31
Analysis	57.29	15.79	20.21

The pattern of the values presented in Table 8 is similar to the pattern for males. The average difficulty levels decrease with increasing complexity of taxonomical category. Knowledge type items are easiest while the Analysis type items are the most difficult. Both the Davis and Difference indices indicate that Comprehension type items discriminate better than the average. Analysis and Knowledge type items are least discriminating for females.

Results of Selection of Most Discriminating Items for Males and Females

From the total item pool a test was constructed by selecting the 100 items indicated as most discriminating by the Davis index. A second test was constructed by selecting those 100 items indicated as most discriminating by the Difference index. This procedure was followed for both the male and female groups.

For males, 80 of the 100 items identified as most discriminating by the Davis index were also selected by the Difference index. There were 79 common items selected by the two techniques for females.

The percentage of test items classified in each taxonomical category was computed for the 100 item tests. These values are presented in Table 9 along with the comparable values for the total item pool.

TABLE 9. Percentage of Most Discriminating Items Classified in Taxonomical Categories by Sex

Taxonomical Categories	Males		Females		
	Items Selected By: Davis Index	Difference Index	Items Selected By: Davis Index	Difference Index	Items In Total Item Pool
Knowledge	24	22	24	20	27
Comprehension	38	40	29	36	29
Application	18	21	31	27	24
Analysis	20	17	16	17	20
Total	100	100	100	100	100

None of the structures of the 100 item tests, as indicated by the percentage of items in each taxonomical category, closely correspond to the structure of the total item pool. In every case, the percentage of Knowledge items in the 100 item tests is less than the percentage of Knowledge items in the total item pool. Comprehension itemsare in general selected more often for inclusion in the 100 item tests than seems warranted by the structure of the larger item pool. The selection of Application items appears to operate

differentially for males and females. For males the 100 item tests reveal a smaller percentage of Application items than the total item pool. The reverse is true for females. Analysis items in general appear to be under-selected in the 100 item tests.

The tests composed of most discriminating items are not representative of the total item pool from which they were selected. These tests would not adequately measure the instructional objectives measured by the total item pool. In general, less emphasis is given to Knowledge and Analysis items and more emphasis is given to Comprehension items by the selected tests than is the case in the original item pool.

As cited above, 80 per cent of the items selected by the two discrimination indices for the male group were the same items. The two 100 item tests for males therefore have 80 items in common. The two selected tests for the female group have 79 common items. Table 10 presents the percentage of these common items classified in each taxonomical category.

TABLE 10. Taxonomical Classification of Items Identified as Most Discriminating by Both Discrimination Indices by Sex

Taxonomical Category	Common Items for Males	Common Items for Females
Knowledge	21	21
Comprehension	42	33
Application	18	31
Analysis	19	15
Total	100	100

The values in Table 10 are reported in order to compare the percentage of common items classified in each taxonomical category with the values presented in Table 9 for the entire 100 selected items and the total item pool. The values in Table 10 are similar to those yielded by the entire 100 item tests. In reference to the total item pool, Knowledge and Analysis items are underselected while Comprehension items are overselected. Application type items operate differentially by sex group. This pattern is identical to the pattern for the 100 item tests. The taxonomical pattern of common items selected by the two indices in both the male and female groups does not appear unlike the pattern for the entire 100 items tests.

Of the 100 items identified as most discriminating by Davis index for males, 67 of these items were also identified as most discriminating by the Davis index for females. Comparison of the 100 item tests selected by the Difference index indicates that 70 items selected for the male group were similarly identified for the female group. Table 11 presents the percentage of these common items classified in each taxonomical category.

TABLE 11. Taxonomical Classification of Items Identified as Most Discriminating for Males and Females by each of the Discrimination Indices

Taxonomical Category	Common Items by Davis Index	Common Items by Difference Index
Knowledge	25	23.5
Comprehension	37	40
Application	23	23.5
Analysis	15	15
Total	100	100

The values in Table 11 indicate that with respect to the taxonomical structure of the item it makes little difference which of the two indices are used in the selection procedure. Once again the percentage of Knowledge and Analysis items are not as large as the comparable percentages in the total item pool while the percentage of Comprehension items is larger. Those items which are not selected by each discrimination index for both males and females seem to have little effect on the results.

The discrepancies between the structure of the 100 item tests and the structure of the total item pool do not seem to be a function of the discrimination index employed in the selection process.

The average difficulty and discrimination indices for each 100 item test are presented by taxonomical category in Tables 12 and 13. Similar values computed on the total item pool appear in Tables 7 and 8.

As in the total item pool, the average Difference index values are higher than the average Davis index values for both males and females. The values of each discrimination index do not vary with taxonomical category. This similarity across taxonomical category is a function of the selection procedure. The most discriminating items regardless of taxonomical category were selected for inclusion in the 100 item tests. In order to be selected an item in any category would have to discriminate as well as item in all the other categories. Thus, the average discrimination values for a particular index will differ only slightly across taxonomical categories.

TABLE 12. Average Difficulty and Discrimination Indices by Taxonomical Category for 100 Item Tests - Males

Taxonomical Category	Test Selected By Davis Index		Test Selected By Difference Index	
	Difficulty	Discrimination	Difficulty	Discrimination
Knowledge	71.17	24.83	62.00	33.50
Comprehension	63.32	23.79	59.20	32.58
Application	69.67	23.89	60.48	32.05
Analysis	68.30	24.15	62.00	33.24

TABLE 13. Average Difficulty and Discrimination Indices by Taxonomical Category for 100 Item Tests - Females

Taxonomical Category	Test Selected By Davis Index		Test Selected By Difference Index	
	Difficulty	Discrimination	Difficulty	Discrimination
Knowledge	70.96	27.71	70.10	34.75
Comprehension	63.07	26.55	57.06	34.25
Application	64.84	26.74	58.59	35.81
Analysis	67.75	26.75	59.82	34.76

The average difficulty values on the total item pool (Tables 7 and 8) decreased with increasing complexity of taxonomical category, i.e. Knowledge items are easiest while Analysis items are most difficult. This pattern does not appear when the average difficulty indices for the 100 item tests are examined.

For males the average difficulty levels for both 100 item tests indicate that Comprehension type items are the most difficult. Knowledge items are easiest in the 100 item test selected using the Davis index while for the 100 item test selected by the Difference index Knowledge and Analysis items are equally least difficult.

For females, the average difficulty levels for both 100 item tests indicate that Comprehension type items are the most difficult and Knowledge type items are the easiest.

The items selected for the 100 item tests are not typical of the items in the total item pool with respect to taxonomical classification or difficulty levels. A good indication that this would occur is given by the average discrimination indices for the total item pool (Tables 7 and 8). Since certain types of items are more discriminating than others it would be expected that these items would be selected for the 100 item tests more often than types of items which are less discriminating.

For males the average Davis index of the original 379 items was 15.66; the average Difference index was 20.28. The average Davis and Difference indices for males on the 100 item tests were 24.13 and 32.78 respectively. For females the average Davis index of the original 379 items was 17.04; the average Difference index was 21.34. The average Davis and Difference indices for females on the 100 item tests were 26.92 and 34.86 respectively. These figures illustrate the increase in average discrimination values resulting from the selection procedure.

The average difficulty index of the original 379 items for males was 62.10 and was 61.79 for females. The same values computed on the 100 item tests for males are 67.34 for the test selected using the Davis index and 60.56 for the test selected using the Difference index. For females the average difficulty level of the 100 item test selected using the Davis index was 66.26 and was 60.55 for the 100 item test selected using the Difference index. The average difficulty levels of the tests selected using the Davis index increased as the most discriminating items were selected from the item pool. The reverse is true for the tests selected using the Difference index.

Within each taxonomical category a similar pattern appears for both the male and female groups. Those types of items which have the highest discrimination indices in the total item pool are selected more often than the other types of items for the 100 item tests and have the lowest average difficulty level of the 100 item tests, i.e. they are the most difficult group of items selected. The Comprehension items follow this pattern. Those types of items in the total item pool which have the lowest average discrimination indices (Knowledge and Analysis items) are selected for the 100 item tests less often and are, in general, the easiest group of items selected.

Results of Selection of Most Discriminating Items by Achievement Level

Average difficulty and discrimination indices were computed on the total item pool using the high and low achieving male and female groups described in Chapter II. These values are presented in Tables 14 to 17.

The average difficulty indices for high and low achieving males decrease with increasing complexity of taxonomical category. This was also true for males in general (Table 7). The discrepancy in average difficulty levels for the high and low achieving groups is a result of defining these groups by their scores on the total test.

TABLE 14. Average Difficulty and Discrimination Indices by Taxonomical Category for Total Item Pool - High Achieving Males

Taxonomical Category	Difficulty Index	Davis Index	Difference Index
Knowledge	75.56	10.08	9.82
Comprehension	75.53	9.40	9.48
Application	69.94	9.36	10.23
Analysis	68.22	5.67	5.87

TABLE 15. Average Difficulty and Discrimination Indices by Taxonomical Category for Total Item Pool - Low Achieving Males

Taxonomical Category	Difficulty Index	Davis Index	Difference Index
Knowledge	55.29	6.35	8.55
Comprehension	51.92	6.04	8.41
Application	48.98	6.91	9.68
Analysis	47.43	6.28	8.82

TABLE 16. Average Difficulty and Discrimination Indices by Taxonomical Category for Total Item Pool - High Achieving Females

Taxonomical Category	Difficulty Index	Davis Index	Difference Index
Knowledge	76.76	8.74	8.11
Comprehension	75.89	10.39	10.95
Application	69.20	8.99	9.19
Analysis	68.21	7.79	7.20

TABLE 17. Average Difficulty and Discrimination Indices by Taxonomical Category for Total Item Pool - Low Achieving Females

Taxonomical Category	Difficulty Index	Davis Index	Difference Index
Knowledge	55.92	6.48	8.97
Comprehension	51.81	7.33	9.93
Application	46.35	6.57	9.24
Analysis	46.76	5.12	6.96

In the total item pool for high achieving males Analysis items are least discriminating. When the average Davis index is considered, Knowledge type items are most discriminating. Applications items are indicated as most discriminating by the Difference index.

For low achieving males Comprehension items are least discriminating. Application items are indicated as most discriminating by both the Davis and Difference index for low achieving males.

For males in general (Table 7) Analysis items were least discriminating while Comprehension items were most discriminating. The values for high and low achieving males show considerable deviation from this pattern. Analysis items were least discriminating in the high achieving male group but in no case were Comprehension items indicated as most discriminating for the high and low achieving groups.

The average difficulty indices for high and low achieving females are similar to those values for high and low achieving males. Knowledge items are the easiest for these groups while Analysis items are in general the most difficult. This was also true for the total female group (Table 8).

In the total item pool for both high and low achieving females Comprehension items are, on the average, the most discriminating type items. Analysis items are the least discriminating. This pattern is consistent with that for the total female group (Table 8).

The average discrimination indices computed on the total item pool indicate in general that Comprehension items are most discriminating and Analysis items are least discriminating for the total female group and for the high and low achieving groups. High and low achieving female groups are more like the total female group than high and low achieving males are like the total male group.

From the total item pool two tests were constructed by selecting the 100 items indicated as most discriminating by the Davis and Difference indices. This procedure was followed for the high and low achieving male and female groups. Thus,for each of the four groups two tests were constructed - one using high Davis indices as the criteria for selection of items and the other using Difference indices.

The percentage of test items classified in each taxonomical category was computed for each of the 100 item tests. These values are presented in Tables 18 and 19 along with the comparable values for the total item pool.

TABLE 18. Percentage of Most Discriminating Items Classified in Taxonomical Categories for High and Low Achieving Males

Taxonomical Category	High Achievers: Items Selected By: Davis Index	High Achievers: Items Selected By: Difference Index	Low Achievers: Items Selected By: Davis Index	Low Achievers: Items Selected By: Difference Index	Items in Total Item Pool
Knowledge	32	32	23	26	27
Comprehension	31	29	28	29	29
Application	27	29	27	26	24
Analysis	10	10	22	19	20
Total	100	100	100	100	100

TABLE 19. Percentage of Most Discriminating Items Classified in Taxonomical Categories for High and Low Achieving Females

Taxonomical Category	High Achievers: Items Selected By: Davis Index	High Achievers: Items Selected By: Difference Index	Low Achievers: Items Selected By: Davis Index	Low Achievers: Items Selected By: Difference Index	Items in Total Item Pool
Knowledge	23	23	30	30	27
Comprehension	35	41	33	33	29
Application	26	23	21	21	24
Analysis	16	13	16	16	20
Total	100	100	100	100	100

Once again the structures of the 100 item tests, as indicated by the percentage of items in each taxonomical category, differ from the structure of the total item pool from which the 100 item tests were selected. The structure of the 100 item test for low achieving males is most like the structure of the total item pool. Knowledge items are not selected as often for inclusion in the 100 item tests while Application items are selected more often than seems warranted by the structure of the total item pool. These differences are slight when compared with the differences between the taxonomical structures for high achieving males and the total item pool.

For high achieving males the 100 item tests reveal a larger percentage of Knowledge items and a smaller percentage of Analysis items than the total item pool. The differences between the high and low achieving males are even more striking for these two taxonomical categories.

The percentage of Comprehension type items for both high and low achieving females is higher than the percentage of Comprehension items in the total item pool. Analysis items are selected less often for inclusion in the 100 item tests for high and low achieving females than seems warranted by the percentage of Analysis items in the total item pool. The major difference between high and low achieving females appears in the percentage of Knowledge items selected for the 100 item tests. The percentage of items falling in the Knowledge category for low achieving females is higher than that for the total item pool while the reverse is true for high achieving females.

The relationship between the average discrimination indices computed in each taxonomical category for the total item pool and the subsequent selection of items for the 100 item tests is once again apparent in this data. For high achieving males the average discrimination indices computed on the total item pool are lowest for Analysis type items. The Analysis items are selected least often for inclusion in the 100 item tests. For low achieving males the average discrimination indices computed on the total item pool are all similar with the Application items discriminating slightly better than the rest. The Application items are the only type items selected more often for inclusion in the 100 item tests for low achieving males than seems warranted by the structure of the item pool.

The average discrimination indices for both high and low achieving females indicate that Comprehension items are most discriminating and Analysis items are least discriminating. The percentage of Comprehension is higher in the 100 item tests than it is in the total item pool; the percentage of Analysis items is lower.

As for the total male and female groups, the tests composed of most discriminating items for low and high achieving males and females are not, in general, representative of the total item pool from which they were selected. These selected tests could not adequately evaluate the instructional objectives measured by the total item pool.

Summary of Major Findings

In the total item pool 27 per cent of the items were classified as Knowledge type items. In the 100 item tests selected from the total item pool using the Davis and Difference indices computed for the male group, the percentage of Knowledge items decreased. The same result was obtained in the 100 item tests selected for the female group.

In the 100 item tests for high achieving males and low achieving females the percentage of Knowledge items was larger than that value for the total item pool. The opposite was true for the tests constructed using the low achieving male and the high achieving female groups.

In the total item pool 29 per cent of the items were classified as Comprehension type items. In the 100 item tests selected from the total item pool using the two indices computed for the male group, the percentage of Comprehension items increased. The same result was obtained in the 100 item test selected by the Difference index for the female group.

In the 100 item tests for high and low achieving males the percentage of Comprehension items does not appreciably differ from the percentage in the total item pool. The 100 item tests for high and low achieving females have a higher percentage of Comprehension items than does the total item pool.

In the total item pool 24 per cent of the items were classified as Application type items. The 100 item tests for the male group has a smaller percentage of Application items while for the female group the 100 item tests have a larger percentage of Application items than does the total item pool.

In the 100 item tests for high and low achieving males the percentage of Application items is greater than the percentage in the total item pool. For the high and low achieving female groups the trend is toward fewer Application items than the total item pool.

In the total item pool 20 per cent of the items were classified as Analysis type items. In the 100 item tests for both the male and female groups the percentage of Analysis items decreased with one exception where the value remained the same. A similar result was obtained for the tests selected for the high and low achieving female groups and for the high achieving male group. The 100 item test selected from the total item pool using the Davis index computed for low achieving males was the only 100 item test in which the percentage of Analysis items exceeds that in the total item pool.

In summary, the 100 item tests selected from the total item pool using discrimination indices computed for males were structurally different from those computed for females. The same results were obtained for high and low achieving males and females. None of the 100 item tests closely correspond to the taxonomical structure of the total item pool.

CHAPTER IV

SUMMARY, CONCLUSIONS AND IMPLICATIONS

This study sought to investigate the effect of statistical item selection on the structure of the final form of a test. An item pool of 379 multiple-choice natural science items was identified. This item pool was described with consideration being given to the average difficulty and discrimination levels of the items computed for male and female tryout groups. The classification of the items according to the instructional objectives being measured was also examined for the total item pool.

From this large item pool the 100 most discriminating items identified by the Davis index and the 100 most discriminating items identified by the Difference index were selected to form two 100 item tests. This procedure was followed using the data from the male and female tryout groups separately.

The entire procedure was repeated using data obtained from high and low achieving male and female tryout groups. These groups were selected from the total male and female groups by identification of subjects who had either high or low scores on the 379 items.

High discrimination values were used as the criteria for the selection of items for the smaller tests in order to approximate a commonly used test construction procedure. The final form of a test is often constructed by the selection of items from a larger item pool using high discrimination indices as a guide.

The 100 item tests were compared to the total item pool with respect to the classification of items according to the instructional objectives being measured by each item. In order for a 100 item test to validly measure the instructional objectives measured by the total item pool the item selection procedure should not disproportionally select items from these categories of classified items.

Some comparisons were also made between the average difficulty and discrimination values for the 100 item tests and the total item pool in order to further examine the effect of statistical item selection.

Conclusions

The major conclusions of this study are as follows:

1. Statistical selection of items from the total item pool has a biasing effect on the selected tests. The proportion of items in the selected tests which measure certain instructional objectives is unlike the proportion of items in the total item pool which measure the same objectives. The selected tests are not representative of the total item pool in this respect.

2. Statistical selection of items from the total item pool appears to operate differentially for male and female groups. When the statistical data obtained from the female tryout group is used to select tests from the total item pool the results differ from those obtained using the male tryout group. The structure of the selected tests as indicated by the taxonomical structure of the items differs for male and female groups. Application items are selected from the total item pool more often using the discrimination indices computed for the female tryout group than for the male tryout group. In general, Application type items discriminate better for females.

3. Statistical selection of items from the total item pool appears to operate differentailly for high and low achieving tryout groups, both male and female. The structure of the selected tests as indicated by the taxonomical classification of items differs for high and low achieving groups. The major differences between the tests selected using indices computed for high and low achieving male tryout groups appear in the selection of Knowledge and Analysis items. For high and low achieving female groups the major difference in the selection of items appears for Knowledge items.

4. The statistical selection of items from the total item pool operates similarly with respect to the taxonomical structure of the items selected no matter which of the two discrimination indices are used as the criterion for selection. The taxonomical structure of the tests selected using the Davis index as the criterion for selection does not differ appreciably from the structure of the tests selected using the Difference index as the selection criterion.

The general results of pertinent research reviewed earlier in this study (p. 11) indicate that there are few differences in the end results obtained by using different discrimination indices. This finding holds true with respect to the taxonomical structure of the tests selected in this study.

The major difference between the selection of items by the two indices is in the difficulty level of the items selected. The Davis index tends to select easier items than does the Difference index. This would be expected due to the difference between the two indices previously described (p. 20).

Implications

In the assembly of an item pool the test constructor includes items which measure the instructional objectives to be evaluated in the final form of the test. The items for the final form of the test are commonly selected from the item pool on the basis of the statistical analysis of the items using data obtained from a tryout group. The results of this study would suggest that the structure of a test constructed in this manner, as indicated by the proportion of items measuring each instructional objective, may be very much unlike the structure of the total item pool. Consequently, the constructed test may not evaluate the instructional objectives in the same proportion as would the original item pool.

The practice of statistical selection of items for the final form of an evaluation instrument is seriously questioned by the results of this study. Statistical item selection alone is not sufficient; other variables should be considered.

It has been shown that in the total item pool the average discrimination values differ for the four major categories of items classified according to the instructional objective being measured. This should indicate to the test constructur that selection of items from the item pool on the basis of these discrimination indices will be biased in favor of the group of items which have the highest average discrimination values. This suggests the possibility of selecting the most discriminating items within a particular taxonomical category rather than selecting the most discriminating items from the total item pool, disregarding the taxonomical structure of the items. In the planning stages the test constructor would specify the number of items measuring each instructional objective to be included in the final form of the test. The items for the item pool would be written accordingly. Then, within each category a specified number of items would be selected for inclusion in the final form of the test. In this manner the test constructor would be sure that the final form of the test will validly evaluate the instructional objectives as indicated in the planning stage.

This study has also indicated that the sex and achievement level of the tryout group has an effect on the statistical item selection. It is a well-known principle that the tryout group should be essentially similar to the group for which the test is to be used. The results of this study clearly indicate that the sex and achievement level of the tryout group make a difference in the selection of certain types of items for inclusion in the final test form. If, for example, Application items discriminate better for females than for males, then the test constructed using the discrimination indices computed using the female tryout group for item selection will include more Application items than will the test constructed using the data from the male tryout group. If the proportion of males and females in the group or groups for which the test is intended is not similar to the proportion of males and females in the tryout group the test would not validly measure the instructional objectives specified in the planning stage. This could be a critical consideration in the construction or use of a test with an all male or female group. The test constructor might consider the possibility of computing item discrimination indices separately by sex as well as by taxonomical category and selecting items for the final form of the test accordingly.

The use of item selection using discrimination indices computed for high and low achieving males and females indicate similar implications. A test constructed using data from the entire tryout group might be invalid when used with high or low achieving students. This could be a crucial consideration if the test were to be used to evaluate a group of high or low achieving students, e.g. scholarship testing.

The results and implications of this study must be tempered in light of the following limitations inherent in the study:

1. The specific population in this or any study may have an effect on the obtained results. In order for the results to be meaningful for a larger population the study should be replicated with subjects of varying age and grade levels.

2. The test items used in this study may have a biasing effect on the results. Natural science items were used in this study. The procedure should be replicated using items from various subject-matter areas.

As evidenced by the discrimination indices presented in Appendix B and C, the test items used in this study were in general high quality items. The study should be replicated using test items of varying quality, from the highly discriminating to the negative discriminating.

The total item pool identified in this study was comprised of items from three tests given at different times. It is not evident how this time interval between test administration has affected the results obtained. It would be desirable to replicate the study using items all of which were administered to the tryout group at one time.

In the opinion of the investigator the limitation imposed by the quality of the test items might be more of an asset to the study rather than a liability. It is highly possible that a replication of the study using a more typical item pool with varying quality items would yield even more discrepancies between the selected tests and the original item pool from which the tests are selected.

3. Only two discrimination indices were employed in the study. Although these were considered representative of the wide variety of indices available, there is no assurance that the use of other indices would yield similar results. This suggests replications using a variety of discrimination indices as the criteria for the selection of items.

In order for the results of this study to have far-reaching implications for educational achievement test construction further study is required. The study has suggested, however, that the type of statistical data on test items commonly used in item selection might be only one of many considerations in the selection procedure. The test constructor should consider the taxonomical structure of the test items and the sex and achievement level of the group for which discrimination indices are to be computed. The test constructor should not overlook these aspects of the test items and the discrimination indices computed for a particular group if a valid instrument is desired.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Adams, James F. "An Evaluation of the Effect of Level of Item Difficulty on Various Indices of Item-Discrimination." Unpublished Ph.D. thesis, State College of Washington, Pullman, Washington, 1960.

2. Adkins, Dorothy C. "A Comparative Study of Methods of Selecting Test Items." Unpublished Ph.D. thesis, The Ohio State University, Columbus, Ohio, 1937.

3. Barthelmess, Harriet M. "The Validity of Intelligence Test Elements," Contributions to Education, # 505, New York: Bureau of Publications, Teachers College, Columbia University, 1931.

4. Barzun, Jacques. The House of Intellect. New York: Harper, 1959.

5. Bloom, Benjamin S. (ed.) Taxonomy of Educational Objectives. New York: David McKay Company, Inc., 1961.

6. Bryden, M. P. A Non-Parametric Method of Item and Test Scaling, Educational and Psychological Measurement, 1960, 20, No. 2, 311-315.

7. Clemens, William. An Index of Item Criterion Relationship. Educational and Psychological Measurement, 1958, 18, 167-172.

8. Colver, Robert M. Estimating Item Indices by Nomographs. Psychometrika, 1959, 24, 179-185.

9. Conrad, Herbert S. The Experimental Tryout of Test Materials. In E. F. Lindquist (ed.) Educational Measurement. Washington: American Council on Education, 1961, 250-265.

10. Cook, W. W. The Measurement of General Spelling Ability Involving Controlled Comparisons Between Techniques. University of Iowa Studies in Education, 1932, 16, No. 6.

11. Davis, F. B. Item-Analysis Data: Their Computation, Interpretation, and Use in Test Construction. Harvard Education Papers, 1946, No. 2.

12. Davis, F. B. Item Selection Techniques. In E. F. Lindquist (ed.) Educational Measurement. Washington: American Council on Education, 1961, 266-328.

13. Dressel, Paul L., and Nelson, Clarence H. Questions and Problems in Science, Item Folio No. 1. Princeton: Educational Testing Service, 1956.

14. Dyer, Henry S. Measuring Creativity and Intelligence. In The Behavioral Sciences and Education, No. 10, College Admissions. Princeton: College Entrance Examination Board, 1963, 46-65.

15. Ebel, Robert L. Writing the Test Item. In E. F. Lindquist (ed.) Educational Measurement. Washington: American Council on Education, 1961, 185-249.

16. Ebel, Robert L. Measuring Educational Achievement in School and College Classrooms. East Lansing: Michigan State University Book Store, 1964. (Mimeographed)

17. Flanagan, J. C. General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient From Data at the Talis of the Distribution. Journal of Educational Psychology, 1939, 30, 674-680.

18. Gerberich, J. R. Specimen Objective Test Items: A Guide to Achievement Test Construction. New York: Longmans, Green and Company, 1956.

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

20. Guilford, J. P. The Nature of Intellectual Activity. In The Behavioral Sciences and Education, No. 10, College Admissions. Princeton: College Entrance Examination Board, 1963, 65-73.

21. Guilford, J. P. and Lacey, J. I. Printed Classification Tests: Aviation Psychology Research Report No. 5. Washington: Government Printing Office, 1947.

22. Gulliksen, H. Theory of Mental Tests. New York: John Wiley and Sons, 1950.

23. Hall, Alfred E. "An Empirical Study of the Stability of Four Item-Discrimination Indices Over Groups of Different Average Ability." Unpublished Master's thesis, State University of Iowa, Iowa City, Iowa, 1960.

24. Hill, Walker H.and Dressel, Paul L. The Objectives of Instruction. In Paul L. Dressel, Evaluation in Higher Education. Boston: Houghton Mifflin Company, 1961.

25. Hoffmann, Banesh. The Tyranny of Multiple-Choice Tests. Harper's Magazine, 1961, 222, 37-44.

26. Hoffmann, Banesh. The Tyranny of Testing. New York: The Crowell-Collier Press, 1962.

27. Humphreys, L. G. Commonly Used Statistical Procedures. In J. P. Guilford (ed.) Printed Classification Tests, Aviation Psychology Research Report No. 5. Washington: Government Printing Office, 1947.

28. Kelley, T. L. The Selection of Upper and Lower Groups for the Validation of Test Items. Journal of Educational Psychology, 1939, 30, 17-24.

29. Kuang, H. P. A Critical Evaluation of the Relative Efficiency of Three Techniques in Item Analysis. Educational and Psychological Measurement, 1952, 12, 248-266.

30. Lawshe, C. H. Jr. and Mayer, J. S. Studies in Item Analysis: 1. The Effect of Two Methods of Item Validation on Test Reliability. Journal of Applied Psychology, 1947, 31, 271-277.

31. Lentz, T. F. Hirshstein, Bertha, and Finch, J. H. Evaluation of Methods of Evaluating Test Items. Journal of Educational Psychology, 1932, 23, 344-350.

32. Levine, Richard and Lord, Frederic M. An Index of the Discrimination Power of a Test at Different Parts of the Score Range. Educational and Psychological Measurement, 1959, 19, No. 4, 497-503.

33. Lindquist, E. F. Preliminary Considerations in Objective Test Construction. In E. F. Lindquist (ed.) Educational Measurement. Washington: American Council on Education, 1961, 119-158.

34. Long, John A. and Sandiford, Peter. The Validation of Test Items. Toronto: Department of Educational Research, University of Toronto, 1935.

35. Mason, C. F. A Comparison of Two Methods of Item Analysis on the Basis of Reliability of Selected Tests. Unpublished Master's thesis, Purdue University, Lafayette, Indiana, 1947.

36. Michael, W. B. Development of Statistical Methods Especially Useful in Test Construction and Evaluation. Review of Educational Research, 1956, 26, 89-109.

37. Multiple-Choice Questions: A Close Look. Princeton: Educational Testing Service, 1963.

38. Nelson, Clarence H. Examining In Biological Science. In Comprehensive Examinations in a Program of General Education. East Lansing, Michigan: Michigan State College Press, 1949, 44-64.

39. Nelson, Clarence H. Evaluation of Objectives of Science Teaching. Science Education, 1959, 43, No. 1, 20-27.

40. Nelson, Clarence H. Evaluation in the Natural Sciences. In P. L. Dressel, Evaluation in Higher Education. Boston: Houghton Mifflin Company, 1961.

41. Pinter, R. and Forlano, G. A Comparison of Methods of Item Selection for a Personality Test. Journal of Applied Psychology, 1937. 21, 643-652.

42. Saupe, Joe L. Some Useful Estimates of the K-R Formula No. 20 Reliability Coefficient, Educational and Psychological Measurement, 1961, 21, No. 1, 63-71.

43. Scannell, Dale P. and Stellwagen, Walter R. Teaching and Testing for Degrees of Understanding. California Journal for Instructional Improvement, 1960, 3, No. 1, 8-14.

44. Scott, William A. Measures of Test Homogeneity. Educational and Psychological Measurement, 1962, 22, No. 4, 751-757.

45. Solomon, H. (ed.) Studies in Item Analysis and Prediction. Stanford: Stanford University Press, 1961.

46. Stanley, J. C. and Bolten, D. Book Review in Book Review Section, William B. Michael, (ed.) Educational and Psychological Measurement, 1957, 17, 631-634.

47. Swineford, F. Validity of Test Items. Journal of Educational Psychology, 1936, 27, 68-78.

48. Vaughn, K. W. Planning the Objective Test. In E. F. Lindquist (ed.) Educational Measurement. Washington: American Council on Education, 1961, 159,184.

49. Vernon, P. E. Indices of Item Consistency and Validity. British Journal of Psychology, Statistical Section, 1948, 1, 152-165.

50. Webster, Harold. Item Selection Methods for Increasing Test Homogeneity. Psychometrika, 1957, 22, 395-403.

51. Whyte, William H. The Organization Man. New York: Simon and Schuster, 1956.

APPENDIX A

EXAMPLES* OF NATURAL SCIENCE TEST ITEMS CLASSIFIED IN THE TAXONOMICAL CATEGORIES

KNOWLEDGE

Use the following key to answer Items 1-3

Key:
1. The statement is false under the condition stated.
2. The statement is false regardless of the condition.
3. The statement is true under the condition stated.
4. The statement is true regardless of the condition stated.
5. Impossible to determine without more data.

Statement	
1. A negatively charged particle repels a positively charged particle	if the negative particle has the larger charge. (2)
2. At constant temperature the pressure and volume of a gass are inversely proportional	if the pressure is expressed in mm. of mercury and the volume is expressed in cubic centimeters. (4)
3. Two charged objects repel each other	if both attract similarly charged objects. (3)

COMPREHENSION

For Itesm 4-7 use the following key:

Key:
1. Statement A, which is empirical, is explained by Statement B, which is theoretical.
2. Statement A, which is theoretical, explains Statement B, which is empirical.
3. Statement A is empirical and Statement B is theoretical, but there is no explanatory relationship between them.
4. Statement A is theoretical and Statement B is empirical, but there is no explanatory relationship between them.
5. Either both statements are theoretical or both are empirical.

Statement A	Statement B
4. Crossing in the human population gives a progeny sex ratio of 1 : 1.	Sex chromosomes segregate to different cells during the formation of gametes. (5)

*Examples from 40:124-149

5. A certain monohybrid cross gives a 3 : 1 ratio in a population.	Each parent carries contrasting alleles for the characteristic in question. (1)
6. The hereditary determinants are carried on the chromosomes.	An individual usually resembles his father about as much as he resembles his mother. (4)
7. Genes occur in a constant linear order on the chromosomes.	The same kinds of crosses always give about the same crossover frequencies. (2)

APPLICATION

For Items 8-11 select the most appropriate answers from the following key:

Key:
1. A fact based on empirical observation.
2. An assumption basic to the solution of the problem.
3. A conclusion that is contradicted by the evidence.
4. A conclusion that is justified by the data.
5. Insufficient evidence to make this judgment.

8. In the P generation the brightest rat made fewer than ten errors while the dullest rat made more than 200 errors. (1)

9. The extent to which each rat is able to learn to run a maze can be used as a measure of the rat's intelligence. (2)

10. In the F_1 generation the least number of errors made by any B_1 individual was 12. (1)

11. If the number of errors made in maze running is a criterion of intelligence, then the ablest member of the D_1 group was more intelligent than the ablest member of the B_1 group in the F_1 generation. (4)

ANALYSIS

Items 12-15 are concerned with the following situation:

A certain species of fish comes in various colors. Some are red, some gray; some profusely spotted, some without spots. In an attempt to analyze the inheritance of color in these fish a geneticist has worked out the following:

Red color is due to a recessive gene borne on the X-chromosome. Gray color is due to the dominant allele of this gene.

The gene for plain is dominant in the female and is not sex-linked.

The gene for spotted is dominant in the male and is not sex-linked.
(The gene for spotted is allelic to the gene for plain.)

Let: X represent a sex chromosome which bears the gene for gray color.
x represent a sex chromosome which bears the gene for red color.
P represent the gene for plain (dominant in the female).
p represent the gene for spotted (dominant in the male).

For items 12-15 use the following key:

Key:
1. XYPP x XxPp.
2. XYPP x XxPp.
3. xYPp x xxPp.
4. XYPP x xxPp.
5. None of the above.

12. The theoretical yield of this cross will be
25% red plain males,
25% red spotted males,
50% gray plain females. (4)

13. The theoretical yield of this cross will be
12.5% gray plain males,
12.5% gray spotted males,
12.5% red plain males,
12.5% red spotted males,
50% gray plain females. (1)

14. The theoretical yield of this cross will be
25% gray plain males,
25% red plain males,
50% gray plain females. (2)

15. The theoretical yield of this cross will be
37.5% red spotted males,
37.5% red plain females,
12.5% red plain males,
12.5% red spotted females. (3)

APPENDIX B

FALL TERM EXAMINATION - MALES

Knowledge Items

Item Number	Item Difficulty	Davis Index	Difference Index
1	67	20	29
3	88	19	14
5	48	7	11
6	59	9	7
7	71	24	32
8	71	25	33
9	73	30	35
18	82	11	12
19	65	26	37
20	62	34	47
21	67	10	14
22	82	16	17
23	81	7	8
24	55	19	30
25	75	12	15
26	35	4	6
27	73	16	21
28	44	18	27
29	68	15	21
30	33	14	21
31	44	8	12
32	78	31	32
33	9	-2	-1
34	51	23	36
35	42	2	3
36	49	19	30
37	64	21	30
38	82	29	27

FALL TERM EXAMINATION - MALES

Comprehension Items

Item Number	Item Difficulty	Davis Index	Difference Index
2	64	5	8
4	69	11	15
10	85	22	19
11	57	14	22
12	68	26	35
13	61	14	22
14	47	23	34
15	72	12	16
16	61	23	34
17	72	25	31
49	65	26	37
50	60	24	35
51	67	25	34
52	75	19	23
53	46	7	11
54	70	21	28
55	70	22	29
56	22	25	28
57	64	21	31
58	56	20	32
115	46	16	25
116	76	12	15
117	57	24	37
118	30	10	14
119	79	26	28
120	64	8	12

FALL TERM EXAMINATION - MALES

Application Items

Item Number	Item Difficulty	Davis Index	Difference Index
39	49	15	24
40	69	27	36
41	44	8	12
42	83	21	19
43	77	9	11
44	57	17	47
45	55	24	37
46	73	13	17
47	67	12	17
48	69	20	27
59	59	10	15
60	81	13	14
61	73	16	21
84	46	12	19
85	74	25	30
86	71	20	27
87	70	17	24
88	29	16	22
89	78	25	28
90	63	20	29
91	79	14	16
92	59	21	31
93	69	17	24
94	56	18	27
95	48	11	18
96	43	11	18
97	55	18	27
98	39	18	27
99	37	11	18
100	41	19	29
101	47	14	23
102	78	26	20
103	58	11	18
104	71	18	25
105	71	15	20
106	75	13	17
107	83	14	13
108	75	19	23
121	72	23	29
122	18	1	1
123	74	19	23
124	79	27	28
125	70	21	28

FALL TERM EXAMINATION - MALES

Analysis Items

Item Number	Item Difficulty	Davis Index	Difference Index
62	63	6	9
63	90	10	7
64	52	15	14
65	95	5	2
66	68	13	18
67	68	11	15
68	67	11	15
69	75	15	19
70	68	11	16
71	62	14	22
72	28	7	9
73	68	11	15
74	84	22	19
75	70	20	27
76	58	8	12
77	30	11	15
78	90	13	9
79	60	10	16
80	52	15	24
81	88	12	9
82	20	-2	-2
83	33	10	14
109	68	26	35
110	71	22	29
111	71	20	27
112	68	22	30
113	69	22	29
114	66	21	30

WINTER TERM EXAMINATION - MALE

Knowledge Items

Item Number	Item Difficulty	Davis Index	Difference Index
1	88	10	8
2	53	3	5
3	72	17	22
4	71	16	22
5	39	11	18
6	67	6	9
7	59	7	10
8	82	15	15
14	83	14	13
15	40	16	24
16	73	19	25
17	89	21	14
18	69	14	19
19	69	15	21
20	76	14	17
21	75	12	15
22	90	10	7
23	80	15	17
24	73	14	18
25	76	16	19
25	85	22	19
27	77	23	27
28	73	19	25
29	70	16	22
69	64	26	37
70	73	13	17
71	78	17	20
72	65	23	32
73	81	13	14
74	40	9	13
75	58	13	22
76	64	23	34
77	51	20	32

WINTER TERM EXAMINATION - MALES

Comprehension Items

Item Number	Item Difficulty	Davis Index	Difference Index
9	88	16	12
10	75	6	7
11	92	19	10
12	72	14	18
13	60	15	23
30	73	19	24
31	84	17	16
32	74	14	18
33	43	19	29
34	43	24	37
35	75	27	32
36	72	23	30
37	70	24	32
38	63	20	29
39	82	19	19
40	66	8	12
41	52	20	31
54	75	16	20
55	61	17	26
56	73	15	19
57	80	20	21
58	48	11	17
59	73	27	33
60	73	9	12
61	70	15	21
62	61	14	21
63	27	11	15
81	31	8	11
82	60	23	34
83	74	10	13
84	44	19	29
85	73	7	10
86	83	14	14
101	38	21	32
102	60	32	46
103	53	21	33
104	40	17	26
105	70	21	28
106	46	19	30
107	38	11	17
108	94	0	0
109	94	12	5
110	77	15	18
111	72	13	17
112	65	13	20
113	76	8	10
114	27	3	4

WINTER TERM EXAMINATION - MALES

Application Items

Item Number	Item Difficulty	Davis Index	Difference Index
42	56	18	27
43	54	20	31
44	60	20	30
45	26	18	23
46	51	13	22
47	43	4	6
48	63	5	7
49	43	7	13
50	62	2	4
51	66	11	16
52	62	10	16
53	65	9	13
64	69	17	24
65	89	26	16
66	92	19	10
67	74	18	22
68	71	21	28
94	83	15	15
95	29	15	21
96	19	20	20
97	58	23	35
98	55	14	22
99	27	10	10
100	30	0	0

WINTER TERM EXAMINATION - MALES

Analysis Items

Item Number	Item Difficulty	Davis Index	Difference Index
78	69	29	38
79	61	21	32
80	44	12	19
87	39	2	3
88	43	15	23
89	38	27	39
90	44	14	21
91	80	18	19
92	77	21	24
93	29	2	3
115	66	30	41
116	70	8	12
117	48	21	33
118	53	14	23
119	70	18	25
120	38	9	13
121	52	19	29
122	61	17	26
123	17	7	7
124	64	15	23
125	55	15	23

SPRING TERM EXAMINATION - MALES

Knowledge Items

Item Number	Item Difficulty	Davis Index	Difference Index
30	54	10	16
32	88	17	13
29 -	85	18	16
31	80	18	19
25	67	23	32
24	74	18	22
26	48	1	1
22	71	13	18
21	71	22	29
20	61	14	21
19	71	19	26
27	70	16	22
23	89	4	3
28	47	7	11
11	99	7	2
10	64	33	46
5	90	13	9
7	75	19	23
4	95	24	9
6	87	22	17
13	54	12	19
9	50	16	25
12	70	23	31
92	81	11	12
87	93	9	5
88	35	19	28
86	67	21	30
98	77	15	18
76	58	18	27
81	86	23	19
33	40	24	35
34	38	9	13
35	33	0	0
36	34	7	10
121	43	11	18
117	64	14	22
120	64	8	12
122	50	9	15
124	57	9	19
119	36	-4	-6
118	63	15	23

SPRING TERM EXAMINATION - MALES

Comprehension Items

Item Number	Item Difficulty	Davis Index	Difference Index
2	78	17	19
3	89	15	11
1	95	12	5
8	42	21	31
14	63	16	24
15	85	20	17
18	46	7	12
16	69	23	31
69	69	20	27
70	65	23	31
99	79	16	18
97	39	16	24
93	86	30	22
94	63	26	38
79	74	14	18
78	98	22	17
80	49	20	31
75	54	11	18
45	64	23	33
48	27	16	20
44	63	21	31
47	79	27	29
43	36	8	12
46	65	21	30
42	45	5	8
83	48	12	18
85	98	14	4
82	72	19	25
84	77	22	25
96	96	19	5
95	36	28	40
128	77	17	20
129	57	15	24
37	60	17	26
38	53	21	33
39	44	17	26
123	39	10	15

SPRING TERM EXAMINATION - MALES

Application Items

Item Number	Item Difficulty	Davis Index	Difference Index
17	36	10	16
74	54	20	31
73	73	15	19
71	31	-3	-5
89	13	15	12
90	17	0	0
91	43	8	12
55	80	20	21
65	74	11	14
61	74	21	26
57	16	19	17
67	27	5	6
49	31	-1	-1
59	66	12	17
53	79	20	22
63	30	12	17
51	78	12	14
100	99	7	1
101	93	16	8
102	76	22	26
103	64	26	37
104	85	15	13
105	73	20	26
106	42	31	46
-			

SPRING TERM EXAMINATION - MALES

Analysis Items

Item Number	Item Difficulty	Davis Index	Difference Index
72	29	15	21
77	87	12	10
56	53	11	17
66	90	15	10
62	79	24	26
58	31	4	6
68	47	6	9
50	11	5	4
60	62	18	27
54	70	24	31
64	42	11	18
52	90	15	10
125	38	20	30
126	58	27	40
127	58	27	40
107	83	21	19
108	76	28	31
109	63	10	16
110	78	25	28
111	3	-3	-1
112	66	16	19
113	79	23	25
114	15	0	0
115	82	11	12
116	33	8	11
40	35	11	17
41	33	6	9

APPENDIX C

FALL TERM EXAMINATION - FEMALES

Knowledge Items

Item Number	Item Difficulty	Davis Index	Difference Index
1	68	21	29
3	88	17	13
5	51	13	21
6	90	15	10
7	78	31	32
8	75	35	38
9	79	29	30
18	87	12	10
19	67	29	39
20	64	36	49
21	73	6	7
22	86	19	16
23	88	-2	-2
24	59	29	43
25	73	12	16
26	31	12	17
27	72	9	12
28	42	15	23
29	73	18	23
30	35	16	24
31	43	6	9
32	81	38	33
33	11	9	7
34	49	22	34
35	51	8	13
36	50	23	35
37	68	18	25
38	84	33	26

FALL TERM EXAMINATION - FEMALES

Comprehension Items

Item Number	Item Difficulty	Davis Index	Difference Index
2	60	17	25
4	67	12	18
10	88	13	11
11	60	23	34
12	73	30	35
13	60	11	18
14	53	31	47
15	68	11	15
16	66	31	42
17	77	13	16
49	70	35	43
50	53	26	41
51	71	27	35
52	82	19	19
53	53	16	25
54	69	29	38
55	73	24	30
56	28	26	33
57	64	28	40
58	52	22	34
115	47	18	27
116	82	17	17
118	32	21	29
119	78	28	30
120	62	6	9

FALL TERM EXAMINATION - FEMALES

Application Items

Item Number	Item Difficulty	Davis Index	Difference Index
39	57	21	32
40	69	29	38
41	51	18	27
42	78	25	28
43	77	11	14
44	62	37	52
45	64	24	35
46	76	19	22
47	69	18	25
48	67	26	36
59	60	11	18
60	83	21	19
61	74	18	22
84	52	20	32
85	83	22	20
86	73	18	23
87	75	15	19
88	40	24	35
89	87	34	23
90	69	29	37
91	84	15	14
92	4	22	32
93	74	22	27
94	54	14	22
95	50	13	21
96	45	15	24
97	49	16	25
98	43	23	36
99	41	21	32
100	45	26	39
101	52	19	24
102	77	35	36
103	65	12	18
104	80	25	25
105	74	20	25
106	78	17	20
107	86	29	21
108	79	25	27
121	77	32	33
122	18	1	1
123	77	32	33
124	79	29	30
125	67	24	33

FALL TERM EXAMINATION - FEMALES

Analysis Items

Item Number	Item Difficulty	Davis Index	Difference Index
62	68	6	9
63	92	20	11
64	55	9	14
65	95	13	5
66	76	14	17
67	62	20	29
68	65	12	18
69	74	13	17
70	70	19	26
71	61	21	32
72	34	13	19
73	70	10	14
74	91	25	14
75	81	29	27
87	52	11	17
77	37	15	23
78	90	29	17
79	63	14	21
80	47	22	34
81	91	15	9
82	21	5	5
83	38	13	20
109	75	29	34
110	75	23	28
111	77	32	34
112	73	33	39
113	73	20	36
114	73	22	28

WINTER TERM EXAMINATION - FEMALES

Knowledge Items

Item Number	Item Difficulty	Davis Index	Difference Index
1	86	4	4
2	58	5	8
3	70	20	27
4	70	16	22
5	36	8	11
6	68	8	12
7	64	8	11
8	84	6	6
14	86	23	18
15	45	19	24
16	80	21	22
17	92	24	13
18	80	17	18
19	71	18	24
20	78	25	27
21	81	22	22
22	92	19	10
23	82	21	21
24	77	23	26
25	80	14	15
26	84	22	19
27	72	29	36
28	65	25	35
29	69	17	23
69	62	20	30
70	70	16	22
71	85	15	13
72	71	23	30
73	85	4	4
72	44	9	11
75	58	10	15
76	60	23	34
77	52	15	24

WINTER TERM EXAMINATION - FEMALES

Comprehension Items

Item Number	Item Difficulty	Davis Index	Difference Index
9	90	13	9
10	75	8	9
11	92	20	11
12	73	9	12
13	57	9	14
30	68	16	22
31	85	18	16
32	75	14	17
33	44	20	31
34	32	18	25
35	68	26	35
36	69	22	30
37	75	29	34
38	66	18	26
39	86	12	10
40	76	11	13
41	46	18	27
54	75	17	21
55	68	20	28
56	75	14	17
57	85	20	17
58	56	15	24
59	78	19	22
60	78	12	15
61	68	14	19
62	65	13	19
63	29	11	15
81	38	10	16
82	60	18	27
83	74	14	18
84	48	19	30
85	75	13	16
86	86	14	12
101	43	24	37
102	66	25	35
103	54	25	38
104	39	20	30
105	73	19	24
106	50	18	27
107	27	19	24
108	93	4	2
109	95	17	7
110	75	16	20
111	78	12	14
112	66	16	24
113	78	9	11
114	30	11	16

WINTER TERM EXAMINATION - FEMALES

Application Items

Iten Number	Item Difficulty	Davis Index	Difference Index
42	54	11	18
43	37	18	27
44	47	23	36
45	18	7	8
56	34	10	14
47	39	4	6
48	55	3	5
49	42	8	12
50	60	8	11
51	66	10	14
52	63	9	13
53	67	7	10
64	70	16	22
65	91	20	11
66	94	27	10
67	81	20	20
68	70	19	26
94	82	18	18
95	24	12	15
96	8	9	5
97	47	30	40
98	45	27	41
99	24	-3	-3
100	36	2	4

WINTER TERM EXAMINATION - FEMALES

Analysis Items

Item Number	Item Difficulty	Davis Index	Difference Index
78	68	20	27
79	71	13	17
80	44	10	16
87	39	9	14
88	46	18	28
89	39	22	33
90	49	7	12
91	82	21	21
92	77	19	22
93	28	4	5
115	68	33	43
116	72	4	5
117	57	24	37
118	42	14	22
119	70	16	22
120	32	4	6
121	45	14	21
122	56	11	18
123	16	0	0
124	61	11	17
125	51	12	20

SPRING TERM EXAMINATION - FEMALES

Knowledge Items

Item Number	Item Difficulty	Davis Index	Difference Index
30	47	11	18
32	86	17	15
29	85	16	15
31	68	11	15
25	42	8	12
24	66	19	27
26	45	10	15
22	69	17	23
21	69	20	27
20	64	17	25
19	67	20	28
27	70	14	19
23	88	10	8
28	40	15	23
11	98	11	2
10	63	26	38
5	91	21	12
7	77	18	21
4	94	18	8
6	81	30	28
13	46	15	23
9	52	6	10
12	65	18	26
92	79	18	19
87	91	21	12
88	36	8	12
86	68	29	38
98	77	12	15
76	54	17	26
81	83	35	29
33	39	23	34
34	34	6	9
35	33	-4	-6
36	37	5	8
121	44	13	20
117	67	12	17
120	67	12	17
122	51	13	21
124	65	20	29
119	36	11	17
118	65	20	29

SPRING TERM EXAMINATION - FEMALES

Comprehension Items

Iten Number	Item Difficulty	Davis Index	Difference Index
2	79	15	17
3	89	18	12
1	94	18	8
8	41	25	38
14	57	19	29
15	84	23	20
18	51	7	12
16	62	25	36
69	68	21	29
70	56	19	29
99	74	14	18
97	17	5	5
93	79	20	21
94	57	27	39
79	69	17	23
78	80	28	27
80	43	19	29
75	46	14	22
45	58	16	25
45	58	16	25
48	31	9	13
44	51	18	27
47	73	28	34
43	32	1	2
46	64	24	34
42	52	2	4
83	44	15	24
85	96	22	8
82	71	22	29
84	79	17	18
85	96	22	8
82	71	22	29
84	79	16	18
96	91	33	16
95	28	24	31
128	68	22	31
129	52	20	32
37	53	19	29
38	48	19	29
39	45	18	27
123	43	12	19

SPRING TERM EXAMINATION - FEMALES

Application Items

Item Number	Item Difficulty	Davis Index	Difference Index
17	34	8	12
74	41	22	34
73	58	22	34
71	31	-4	-6
89	10	5	4
90	15	5	5
91	43	-3	-5
55	76	20	23
65	71	15	20
61	67	15	22
57	9	22	13
67	30	3	5
49	21	-2	-2
59	68	11	15
53	73	24	29
63	17	-1	-1
51	78	19	22
100	97	7	2
101	90	13	9
102	65	22	31
103	62	25	36
104	84	9	9
105	59	25	38
106	35	30	41

SPRING TERM EXAMINATION - FEMALES

Analysis Items

Item Number	Item Difficulty	Davis Index	Difference Index
72	21	5	6
77	79	21	23
56	47	19	30
66	86	20	17
62	71	17	23
58	27	9	12
68	44	10	15
50	8	0	0
60	58	15	24
54	61	20	30
64	31	4	6
52	88	21	15
125	36	10	16
126	51	23	36
127	42	24	37
107	80	21	22
108	70	26	35
109	56	22	34
110	71	25	32
111	3	-4	-1
112	60	29	43
113	78	19	22
114	16	3	3
115	80	12	13
116	29	16	22
40	39	8	12
41	30	12	17