A COMPARISON OF THREE BASES FOR DETERMINING ITEM DISCRIMINATION

Dissertation for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
ERIC MARSHALL GORDON
1976

This is to certify that the

thesis entitled

A COMPARISON OF THREE BASES FOR DETERMINING ITEM DISCRIMINATION

presented by

ERIC MARSHALL GORDON

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education

Robert L. Ebel

Major professor

Date September 13, 1976

O-7639

ABSTRACT

A COMPARISON OF THREE BASES
FOR DETERMINING ITEM
DISCRIMINATION

By

Eric M. Gordon

The purpose of this study was to compare D, pre-post and expert-novice indices of discrimination with both true-false and multiple choice items. The three major questions that were formulated as research hypotheses were:

1. Is there a significant relationship between the three indices of discrimination for true-false and multiple choice items?

2. Using a large item pool and D, pre-post or expert-novice discrimination indices as the sole criterion for item selection, is there a significant number of common items in the final test forms for true-false and multiple choice items?

3. Are multiple choice and true-false achievement tests that were constructed using D, pre-post or expert-novice indices of discrimination as the sole criterion for item selection equally reliable?

To obtain a more complete understanding of the similarities and differences among these indices four

supplementary analyses were performed. The supplementary analyses were designed to answer the following research questions:

1. Are there significant differences in the mean discrimination index values for D, pre-post and expert-novice indices with true-false or multiple choice items?

2. Using Spearman Rho correlations, are there significant relationships among the three indices of discrimination which are different from the relationships obtained using Pearson's product moment correlations (thus investigating both monotonic and linear relationships)?

3. How similar are the indices as measured by the amount of shared variance obtained by squaring the correlational coefficients?

4. With true-false items, do any of the discrimination indices display a significant preference for selecting keyed true or keyed false items in developing achievement tests?

This investigation utilized a two hundred forty item true-false item pool and a sixty item multiple choice item pool. The true-false items were administered to two sections of an introductory course in teacher made tests and measurements as a pre test and also to two other sections of the same course as a post test. The multiple choice items were administered to three sections of an introductory course in standardized tests and measurements as both a pre and post test.

ad

in

mo

pu

it

in

fa

se

fi

bi

we

we

li

ic

an

th

ea

me

re

to

ni

A group of measurement and evaluation experts were administered both the true-false and multiple choice items.

For each of the items D, pre-post and expert-novice indices of discrimination was generated. Pearson product-moment and Spearman-Rho correlational coefficients were computed for each pairwise combination of indices across both item formats to test the degree of relationship between the indices. For each index the best one hundred twenty true-false items and the best thirty multiple choice items were selected to be considered the final test. Using phi-coefficients the amount of item overlap was tested for each combination of indices and item formats.

Kuder-Richardson Formula 20 reliability coefficients were calculated for each of the final tests. Comparisons were then made to test the differences in the obtained reliability coefficients.

To investigate whether any of the indices systematically provided higher or lower discrimination values, an analysis of variance was performed comparing the means of the indices. This analysis included all of the items in each of the pools.

The items in each of the final true-false achievement tests were dichotomized according to their keyed correct response. A chi square analysis was then performed to test whether any of the indices tended to select a significantly different number of either keyed true or keyed

false

hypot

tion

tion

item

indi

icar

bil

dis

tip

Bro

the

of

ite

th

th

us

to

no

false items then would be expected by chance.

The results associated with the three major research hypotheses and four supplemental analyses were:

1. With the exception of D and pre-post discrimination indices with multiple choice items, significant relationships were found among the three discrimination indices.

2. There was a significant amount of overlap in the items selected for inclusion in a final test by all of the indices tested with true-false items. There were no significant amounts of overlap obtained with multiple choice items.

3. There was no significant differences in reliability for tests where items were selected by any of the discrimination indices. When the reliabilities of the multiple choice items were adjusted utilizing the Spearman-Brown formula, no significant differences were found between the reliabilities of true-false and multiple choice tests.

4. There was no significant difference in the means of the indices for true-false items. With multiple choice items the mean value for the expert-novice index was greater than the means of either D or the pre-post indices.

5. Applying Spearman Rho correlational analysis to the data resulted in findings very similar to those obtained using Pearson's product-moment correlational analysis.

6. Squaring the obtained correlational coefficients to determine the amount of shared variance revealed that in no case was there more than 45% shared variance.

pool

pre

ite

pre

7. Dividing the true-false items into two separate pools based on the keyed correct answer it was found that pre-post and expert-novice indices significantly favor false items in their item selection process. D more adequately represented the proportions found in the item pool.

A COMPARISON OF THREE BASES FOR DETERMINING ITEM DISCRIMINATION

By

Eric Marshall Gordon

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Personnel
Services, and Educational Psychology

1976

tud

gui

in

gui

als

and

ass

sea

and

and

wif

the

occ

ACKNOWLEDGMENTS

The investigator would like to acknowledge his gratitude to Dr. Robert L. Ebel, the dissertation director and guidance committee chairman, for his assistance and support in the preparation of the disseration and for his helpful guidance throughout the doctoral program. Appreciation is also extended to Dr. Willard Warrington, Dr. Richard Johnson and Dr. Phillip Cusick, the dissertation committee for their assistance and helpful criticism.

The investigator is indebted to the Office of Research Consultation and especially to Dr. John Schweitzer and Dr. Robert Wilson for their assistance in the design and analysis portions of the dissertation.

The investigator is particularly grateful to his wife, Sarah, and children Jason Samuel and Sharon Lisa for their love, moral support and patience and for the more than occasional neglect they have endured.

LIS

LIS

Cha

II

TABLE OF CONTENTS

LIST OF TABLES

Tabl

4.1(

4.1

4.1

4.1

4.1

LIST OF APPENDICES

Intr

a gr

Ther

tion

the

item

mini

dete

of

dete

sub

tes

low

fro

swe

con

ind

the

val

CHAPTER I

THE PROBLEM

Introduction

The value of a testing instrument is determined to a great extent by the quality of the items it contains. Therefore, much emphasis should be placed on the construction, selection and analysis of the items that constitute the body of any test. Empirical analysis of a test and the items it contains cannot be performed until the test is administered to a pilot group. One commonly used tool for determining the quality of the test items is the D index of discrimination. Typically, the procedure utilized to determine this index has involved dividing a group of test subjects into upper and lower groups based on the total test score. On each item the proportion of people in the lower group who answered the item correctly is subtracted from the proportion of people in the upper group who answered the item correctly. This final proportion is then converted to a decimal fraction and becomes the item's index of discrimination. The greater the positive value the better the item.

Recently, a number of educators have questioned the value of this index of discrimination as a measure of item

quality for all tests. This has especially been the case with criterion-referenced achievement tests. Popham and Husek (1969), have suggested that with traditional indices of discrimination an item must contain variance to be classified as good, whereas with criterion-referenced tests an item might be very good even if all people correctly answered it. Thus they contend, traditional indices of discrimination are inappropriate for criterion-referenced tests. Carver (1974) has suggested that traditional item indices of discrimination are insensitive to change due to the effects of instruction and hence has recommended that in lieu of an upper-lower division of the scores, a pre-post grouping be utilized to select items that discriminate between those who have had instruction and those who have not yet had instruction. Ebel (1972) has proposed a third method of grouping whereby experts in the subject area constitute one group and novices constitute the second group. This provides an indication of which items discriminate between those who know the material and those who are not knowledgeable in the subject area.

The Usefulness of an Item Analysis

The worth of an item can be defined as its contribution to the evaluation we are trying to make. The worth of each item can be assessed in two ways. The first assessment is rationally determined and can be performed prior to the

adm

ing

ing

ite

pos

str

und

fai

eac

af

ce

te

ea

us

c

u

t

i

t

administration of the test. This assessment involves reviewing all items to determine first if they are written following proper item development principles and second, that the items adequately represent the objectives that they are supposed to measure. In this way items that are poorly constructed can be altered or eliminated. Changes can also be undertaken to insure that all objectives are adequately and fairly covered by the test items.

The second procedure for determining the worth of each item is empirically based. It can only be accomplished after the test has been administered and scored. This procedure is called item analysis. Item analysis provides the test constructor with an indication of which items are too easy or too hard. These extreme items yield little if any useful information. The item analysis also relates the discriminating power of each item. Thus an item analysis is a useful tool for selecting the best items to be included in the test. A second positive outcome of an item analysis is its usefulness in item revision.

Lange, Lehmann, and Mehrens (1967) have discussed the value of considering the indices of discrimination when revising items. They contend that in many situations it is easier to correct the fault in an item than to discard it or develop a new item to replace it.

Meas
an I

two
of t
leve
of
rem
mea
of
100
cor
cul
co
di
he
ar
ab
ti

t
d
i

Measures Obtained From an Item Analysis

An item analysis provides the test constructor with two useful measures for each item in the test. The first of these measures is commonly referred to as the difficulty level of the item. The difficulty level is the proportion of people who answered the item incorrectly. It should be remembered that because of the procedure for obtaining the measure, the more difficult an item the greater its level of difficulty. For example, if an item was administered to 100 people, and eighty of these people answered the item incorrectly, (a relatively difficult item) this item's difficulty level would be .8. Whereas if twenty people had incorrectly answered the item (a relatively easy item), the difficulty level would be .2. This statistic can be very helpful in analyzing the quality of the test. Items that are extremely difficult or extremely easy are of questionable worth as they add little to the information on relative achievement that is provided by the test score.

The second measure provided by the item analysis is the discrimination index. There are numerous methods of determining the discriminating power of an item, each yielding a different type of measure. There is, however, a convenient way of classifying the various methods for producing item discrimination indices into two distinct categories. One category consists of indices that are based on correlations and the other category consists of indices that are

base

indi

tal

tes

a c

the

min

tic

ite

cor

lik

it

re

me

ar

t

m

e

o

e

based on differences. The correlational methods provide an indication of the relationship between the item and the total score. Validity is defined as the correlation of the test to some criterion. If the test score is considered as a criterion measure, and each item is considered a test, then using a correlational method provides a means to determine the validity of each item. Item to total test correlations also provide an indication of the contribution of each item to the reliability of the total test. The greater the correlation between an item and the total test, the more likely the item is measuring the same factor as the other items and hence the more the item will contribute to the reliability of the test.

The difference methods (as well as the correlational methods) for determining the discriminating power of an item are based on the assumption that the function of a test is to discriminate between various levels of the trait being measured. Therefore, each item is analyzed to determine the extent to which it discriminates between the levels that are of interest to the test constructor. Basically the difference methods take the form:

$$D = \frac{R_1 - R_2}{f}$$ Where

D = Difference index.

R_1 = Number of people in one extreme group who answer item correctly. This group consists of those who are considered knowledgeable on the material being tested.

R_2 = Number of people in the other extreme group who answer item correctly. This group consists of those who are considered knowledgeable on the material being tested.

f = Number of examinees in either group.

The membership of R_1 and R_2 is determined by the groups that are to be compared such as upper and lower twenty-seven percent, pre-post, or expert-novice. The difference methods then provide the test constructor with a measure of which items discriminate best between the criterion groups. One step in the process of developing a difference index is to record the number of times each possible response to each item was chosen for each of the criterion groups. With this additional information the test constructor not only obtains an estimate of the discriminating ability of each item, but also a clue as to what the weaknesses are in each of the items. For example, he can see which of the distractors are being chosen too frequently by the upper group or he might discover which distractors are not working (few lower group people choosing the distractor). Thus, the test constructor now can not only use the item analysis data for item selection but also for item revision.

Reasons for the Popularity of the Difference Indices

Both from a theoretical and practical perspective, D is much easier to comprehend than the more sophisticated psychometric approaches to item discrimination. Thus, for

the m

posur

detei

D's

a se

diff

cons

not

of

the

mad

dis

wi

co

di

p

be

th

fu

q

ca

b

m

t

c

the majority of test constructors, who have had limited exposure to psychometric theory, D is a preferred method of determining an item's discrimination level.

Simplicity of calculation is a second reason for D's popularity among test constructors. The D values for a set of items can be computed much quicker and with less difficulty than can the correlational methods. Most test constructors, such as teachers and governmental workers do not have access to computers and also have a limited amount of time that can be devoted to test analysis. Therefore, the tremendous amount of time that can be saved using D has made it a very widely used index of discrimination.

A third reason for the popularity of the indices of discrimination that are based on differences is associated with a special property that is lacking in the majority of correlational approaches. D is biased in favor of middle difficulty items. Ebel (1972) has shown that only when $p = q$ can there be a D value of 1.00. It should be remembered that D values range from -1.00 to +1.00 and the greater the positive value, the more the item discriminates successfully between the groups. As the difference between p and q becomes greater, the less the maximum possible value D can possess. This does not mean that the closer p and q become, the greater the D value by definition. Rather the maximum value of D changes but the observed value reflects things other than just difficulty level such as ambiguity, chance factors and homogeneity of the group tested. When

one

ite

dif

clu

Cor
be

D U

mir

The

sho

int

cor

cor

gro

ob-

is

wor

pe:

th

Th

the

Th

as

le

sho

one uses D for item selection he need not be concerned with item difficulty for items that are extremely easy or extremely difficult will show low D values anyway and therefore be excluded from the final version of the test.

Contrasting Groups that Could be Used to Determine D

D Using Upper-Lower Extreme Groups

Typically the contrasting groups utilized to determine D has been some proportion of the group taking the test. There exists no hard and fast rule as to how the test group should be split. However, the test constructor should take into account two factors when determining the way he will construct the D group membership. The first factor the test constructor should consider is the size of the contrasting groups. The larger the groups the more reliable the results obtained. The smaller the group size the more weight that is assigned to each individual. Thus, an obvious solution would be to divide the test group in half. The top fifty percent of the scores would constitute the upper group and the remaining scores would be included in the lower group. This solution is in contradiction to the second factor that the test constructor should consider for D group membership. This second factor calls for the contrasting D groups to be as extreme as possible. The more extreme the groups the less chance of a person being placed in one group when he should have been placed in the other group. When a test

group is split at the top and bottom fifty percent mark, it is highly likely that a number of those people near the cutting score owe their group placement to some error factor. To the extent that this happens the validity of the groups is diminished. As a result of this misassignment, some upper students are classified as lower and vice-versa. However, if this same group had been split into groups consisting of the upper and lower 10% of the scores, it is highly unlikely that there would be any individuals placed in the upper group who should really be placed in the lower group and vice-versa. True, at each of the cut-off scores there will be some who should have been eliminated who were mistakenly included through errors of measurement. These misassignments are relatively less important, as these people really belong in the large middle category. Much research has gone into what proportions the groups ought to include. Kelley (1939) has shown that to maximize both the size and difference between groups, the upper and lower 27% should be contrasted. Aschenbrenner (1972) has demonstrated that when the extreme groups consisted of at least 100 members, the top and bottom 10% groups were more reliable than the top and bottom 27% groups. Thus for cases where a very large tryout is possible one can feel confident using more extreme groups than the upper and lower 27% as Kelley has advocated. This can provide a great time saving for the persons performing the analysis. Flanagan (1952) has suggested that a more reliable analysis would be obtained if additional weight

wer

wer

of

D U

gr

th

is

er

an

ge

ta

fe

me

cr

a

ti

me

th

ty

cc

ev

gi

ed

th

fo

were to be given to those subjects whose total test scores were extremely high or low and the response of the balance of the people in the contrasting groups counted just once.

D Using Pre-Post Extreme Groups

Carver (1974) has suggested that the division of a group into upper and lower contrasting units does not provide the test constructor with a criterion for item selection that is appropriate for all types of tests; i.e., criterion-referenced achievement tests. In lieu of this method of item analysis which he labels psychometric items analysis, he suggests the contrasting groups consist of students who have taken the item in both a pre and post test setting. He refers to this method of developing contrasting groups as edumetric item analysis. For achievement tests, especially criterion-referenced achievement tests, this method provides a way of selecting items that are most sensitive to instruction. Edumetric tests are given at least twice, once as a measure of baseline achievement and later as a measure of the change in the achievement level after instruction. This type of test is useful in the assessment of three educational concerns. First, an edumetric test is valuable in program evaluation as it provides an indication of how effective a given program is in altering student performance. Second, edumetric tests are useful in allowing a teacher to plot the growth of the students. Lastly, edumetric tests allow for student assessment in a pure form. With psychometric

tes

of

it

gro

den

nor

bas

fec

fec

tha

Goo

pre

ite

ric

for

ers

ear

gra

edu

st

ma

th

Th

de

fr

tests a person's achievement level is determined as a factor of the group that was tested. If the normed group is bright it is more difficult to be assessed highly than if the normed group is not so bright. With the edumetric tests the student's achievement is assessed without any reference to the normed group. Carver also suggests that selection of items based on psychometric item analysis will not produce an effective edumetric test. Rather, to develop a maximally effective edumetric test an item analysis be performed on items that have been used in both a pre and post test setting. Good items are those which are incorrectly answered on the pre test and correctly answered on the post test as these items are most sensitive to change.

Green, Nyquist and Griffore (1975) recommended edumetric tests be developed for two educational testing setting, formative evaluation and diagnostic testing. These researchers suggest that edumetrically sound items facilitate the early detection of a child in difficulty, or a faltering program. In the case of diagnostic testing, they suggest that edumetric tests would be of more value to "...prescribe instruction for particular areas of weakness within the subject matter for each individual student." Lastly they theorize that "...the two tests require different types of items." Thus the grouping basis by which the contrasting groups are determined may produce a different rank ordering of good items from the same general item pool.

Both edumetric item analysis and psychometric item analysis are techniques that are based on the theoretical definition of item analysis. That is, an item is good if it discriminates between those who know and those who do not know the concept being measured. In the case of the edumetric item discrimination index it is assumed that the scores on the post test reflect competence and the scores on the pre test reflect a lack of competence on the concept measured. Likewise, the psychometric item discrimination index assumes that those in the upper 27% of the group are competent and those in the lower 27% are not competent.

D Using Expert-Novice Extreme Groups

A third procedure for developing the contrasting groups to determine D is based on the theoretical definition of the discriminating power of an item. Lennon defines discriminating power as the ability of a test item to differentiate between persons possessing much of the same trait and those possessing little. To obtain a real measure of the discriminating power of an item the contrasting D groups might include subject matter experts and students taking a pre test. There are, of course, problems inherent in this approach to D calculations. Criteria to operationally define experts would have to be developed, and the services of the experts to take the pool of items would need to be secured. The only difference between this expert-novice approach and the upper-lower 27% approach lies in the

ass

gro

per

con

The

gro

gar

ass

oro

jec

et

nov

the

exp

ite

know

answ

know

novi

in a

Rela
to Te

relia

the e

is me

assumption regarding the knowledge level of the upper 27% group. The upper-lower 27% approach must assume that if a person is in the top 27% of his group he does in fact know considerably more than the people in the lower 27% group. The more the difference in score between the upper and lower groups the more confident the test constructor can be regarding this assumption. In the expert-novice grouping this assumption need not be of any concern as there can be rigorous criteria developed to insure that only high level subject matter experts be included in the upper group. Ebel et al. (1962) has argued for utilization of the expert-novice index using the argument that, "One indication of the effectiveness of an item is provided by the success of experts and of novices in answering it correctly. If an item is correctly keyed, unambiguous, and based on essential knowledge or ability, nearly all experts should be able to answer it correctly. If the item does test specialized knowledge or ability, and if it is free to irrelevant clues, novices lacking special training should answer it incorrectly in almost all cases."

Relationship of Item Analysis to Test Reliability

The most commonly used measure of test quality is the reliability coefficient. Reliability can be thought of as the extent to which the test measures accurately whatever it is measuring. The greater the reliability coefficient the

more accurate the test. The goal of any test constructor is to develop as reliable a test as he is capable of producing. Ebel (1972) has outlined a list of techniques that might be employed to increase the reliability of a test. He suggests that the reliability coefficient will be greater for scores:

1. from a longer test than from a shorter test,
2. from a test composed of more homogeneous items than from a more heterogeneous test,
3. from a test composed of more discriminating items than from a test composed of less discriminating items,
4. from a test whose items are of middle difficulty than from a test composed mainly of quite difficult or quite easy items,
5. from a group having a wide range of ability than from a group more homogeneous in ability,
6. from a speeded test than from one all examinees can complete in the time available.

Considering number three above, to maximize reliability, all other things equal, the test constructor should rank order his items according to their discriminating value, and select for his test only the top discriminating items. It should be pointed out here, that item selection should be based on a number of factors, i.e., balance and relevancy rather than just on the discriminating power of the items.

Purpose of the Study

It is the purpose of this study to investigate three grouping modes for determining an item's D index of

disc

lowe

The

sear

Know

fere

and

to t

dres

This

sult

spec

Over

prob

the

analy

discrimination. The three modes to be analyzed are upper-lower 27% groups, pre-post groups, and expert-novice groups. The investigation will focus on answering the following research questions:

1. Given a pool of items, will the three grouping modes order the items the same or differently?
2. Using discrimination index values as the sole criterion for item selection, to what extent will the three grouping modes select the same items for inclusion in a final test?
3. Will the reliability of tests developed using the three modes as the only selection criterion be the same or different?

Knowing whether the different modes select the same or different items will provide an empirical base to address Carver and Green's theory. This information will also be of value to those who teach measurement theory and practices by addressing the following question:

4. Need measurement specialists be concerned about which grouping mode is used for determining the index of discrimination or will the same results occur regardless of the grouping mode utilized?

This study will also address the issues of whether the results obtained are generalizable to different item formats, specifically true-false and multiple choice items.

Overview

In Chapter II the literature relating to the general problem is reviewed by topic area. The design of the study, the items selected, the sample taking the items, and the analysis are discussed in Chapter III. Chapter IV addresses

the results of the study and Chapter V contains a summary of the study, conclusions and implications of the analysis, the limitations of the study and suggestions for future research.

Int

com

nig

bee

tie

usa

shi

of

the

son

nat

of

of

str

D i

ina

ale

cur

bee

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

An abundance of research endeavors relating to the comparative advantages and disadvantages of various techniques for computing indices of item discrimination have been reported in the literature. Through the middle sixties and early seventies with the rise in desirability and usage of criterion-referenced tests the research emphasis shifted towards the logical justification and establishment of indices of item discrimination that are appropriate with these measures. In the middle seventies empirical comparisons of these methods with traditional indices of discrimination have been undertaken and reported. The first section of this chapter includes a general summary of the utility of employing indices of discrimination within a test construction model.

Section two reviews investigations establishing the D index as an acceptable and useful measure of item discrimination. Section three provides an examination of the rationale for inclusion of a pre-post grouping mode along with the current findings when psychometric and edumetric indices have been compared. The fourth section of this chapter deals with

t

r

i

a

i

W

p

m

b

a

a

d

t

v

d

s

m

c

t

m

p

f

the establishment of the expert-novice index as the theoretical representative grouping mode for establishing an item's index of discrimination. The final section provides a review of investigations dealing with comparisons of various item formats.

Writings and Investigations Regarding the Benefits of Determining an Index of Discrimination

Indices of discrimination have become a major component within the field of educational and psychological measurement and evaluation. In a review of current textbooks in the area of measurement all the authors devoted a considerable amount of attention describing the techniques and advantages of computing indices of discrimination when developing tests. Measurement specialists generally agree that the primary function of a test constructor is to develop as valid a test as he is capable of producing. After determining clearly just what the test is supposed to measure, the test constructor develops a pool of items that might be incorporated within the test. As a primary indicator of quality, he may have subject-matter experts review the item pool to insure the items do match the trait being measured. The next step in this process is to tryout or pilot the items. According to Conrad (1962) the seven functions of item tryout consists of the following:

Acc
tio
thr
ite
and
Lan
whe
evi
ite
inv

1. To identify weak or defective items and to reveal needed improvements. More specifically, to identify ambiguous items, indeterminate items, nonfunctioning or implausible distracters, overly difficult or overly easy items, and so forth.

2. To determine the difficulty of each individual item, in order that a selection of items may be made that will show a distribution of item difficulties appropriate to the purpose of the finished test.

3. To determine the discriminating power of each individual item, in order that all items selected may contribute to the central purpose of the finished test and together constitute an efficient measuring instrument.

4. To provide data needed to determine how many items should constitute the finished test.

5. To provide data needed to determine appropriate time limits for the finished test.

6. To discover weaknesses or needed improvements in the mechanics of test taking, in the directions to examiner and examinee, in the provisions for the responses, in the typographical format, and so forth.

7. To determine the intercorrelations among the items, in order to avoid overlap in item selection and to know how best to organize the items into subtests.

According to Marshall and Hales (1971) the item discrimination component of an item analysis is geared to investigate three of the aforementioned concerns, elimination of weak items, location of the sources of weakness in other items, and selection of items for inclusion in the final test. Lange, Lehmann and Mehrens (1967) were interested in seeing whether it was more efficient to rewrite poor items, as evidenced by low discrimination values, or to write new items to improve the discrimination power of a test. These investigators compared the time necessary to write new tests

with

item

intro

State

it re

does

that

items

concl

of t

disc

sele

rive

gest

on

ical

ere

con

ite

cer

dev

iti

lis

vel

if

Gre

with the time necessary to improve the existing items. The item pool consisted of 14 multiple choice items used for an introductory source in educational psychology at Michigan State University. The results of this study indicated that it requires five times longer to write new items than it does to correct existing items. This study also demonstrated that the revised items were more discriminating than the items that were newly generated. Thus these researchers concluded that benefits can be accrued from complete usage of the data provided by the establishment of the index of discrimination in the form of item revision as well as item selection.

Ebel (1951) has advanced an additional benefit derived by performing an item analysis on test items. He suggests that having test constructors conduct an item analysis on their own test items tends to cause them to become critical of the other aspects of item quality that is not covered by the item analysis itself. Thus not only will test constructors improve the quality of their items from an item analysis perspective but also from other quality concerns, hence improving the overall quality of the tests they develop. Although Ebel does not enumerate these other qualities in this article, in a later textbook (1972) he does list ten quality measures that should be considered when developing tests. Although no empirical evidences is available, if Ebel's theory is correct, performing an item analysis will greatly improve the overall quality of tests constructed.

of pe
consi
is th
gated
stude
of th
ale
many
(und
an i
the
obta
the
incl
inde
of
stu
Ano
of
twe
par
wit
the
Wit
duc
Cab

Indices of discrimination are based on the sample of people administered the items. Therefore, a serious consideration when generating indices of discrimination is the reliability of these indices. Ebel (1951) investigated this concern using the upper-lower index with college students. One of the concerns of this study was reliability of the index as a function of the sample size. The rationale underlying this issue stemmed from the concern that since many college courses contain a small number of students (under 50), conceivably the labor necessary to calculate an index of discrimination might not be worth expending if the reliability of the index is extremely low. The data obtained in the study indicated that the reliability of the index did increase when a larger sample of students was included. Thus, for reliable data to be obtained in an index of discrimination, it is best to use a large sample of people. Ebel used both vocabulary and math items in his study and found the same results in both content areas. Another concern of this study was whether the reliability of the index would increase if a more extreme group than twenty-seven was employed. To investigate this, Ebel compared an upper-lower twenty-seven percent grouping mode with an upper-lower ten percent grouping mode. Once again the analysis was performed on vocabulary and math items. With both sets of items, the more extreme division produced a more reliable index. However, in the case of vocabulary items the difference in the obtained reliabilities

was .

Divid

treme

the e

inde

was

two

mini

in e

the

The

inat

sub

.82

was

Unf

dif

Bas

tha

uat

Lev

not

pil

dis

ina

not

was .03 well within a questionable range of true differences. Dividing a group into the upper and lower ten percent extreme groups will not guarantee a more reliable index than the extreme twenty-seven percent groups.

Pyrczak (1973) investigated the reliability of an index of discrimination to determine how stable the index was across similar subjects. In his study, Pyrczak utilized two parallel forms of an arithmetic-reasoning test and administered the instruments to students who planned to major in education. The subjects were administered one form of the tests and then given the remaining form one week later. The index (biserial correlations converted to Davis Discrimination Indices) was determined on a random division of the subjects. The obtained correlations for the two forms were .82 and .47. Thus, in one case the reliability of the index was fairly high and in the other case the index was low. Unfortunately, Pyrczak did not further investigate these differences to discover why such a discrepancy existed. Based however on the results of this study, it appeared that the index of discrimination would under certain situations be accepted as a reliable measure of item quality. Levine (1976) suggests that indices of discrimination should not be generalized much beyond the group who originally piloted the items. He sights the situation of items that discriminate well with younger children being poor discriminators for older children. However, this investigator did not find any justification for not generalizing indices of

discrimination from one group of students taking a course to another group of students taking the same course, if it appears that similar students are contained in both sections.

Diederich (1960), discussed the practical advantage of teachers performing item analysis on their test items. He stressed that with item analysis on classroom tests, "...teachers can build up a file of test items that have worked well in the past or have been revised to eliminate faults that appeared in earlier forms. This file will both reduce the work of constructing tests and improve the tests. If the file is large (as it very soon will be), students seldom learn what questions to expect. Examiners report very little tendency for old items to get "easier" as the years roll on." This idea has also been addressed by Ebel (1972) as a way for teachers to reduce the time necessary to prepare tests and also insure more reliable measures of classroom achievement.

Considering the investigation by Lange, Lehmann and Mehrens (1967) and the discussions Diederich (1960) and Ebel (1972) one might easily misinterpret the role of item analysis in regard to item selection. Granted item analysis is useful for selecting the items to be incorporated into a test, however, other considerations must be included to insure a valid assessment of achievement.

Cox (1964) utilized a pool of items that were categorized according to Bloom's Taxonomy of Educational Objectives. In his research Cox discovered that using only an

index of discrimination to select items from the pool, the final test did not reflect equally the objectives labeled according to the taxonomy. Kwansa (1974) obtained similar results based on a mathematics item pool where items were classified according to the skills they measured. He found some skills were omitted in the test while other skills were greatly overrepresented.

These studies were presented to more adequately represent the role of item analysis in selection of test items. Item analysis is a useful tool, but final item selection should be based on a number of factors to insure that the content measured by the instrument indeed represents the domain of interest.

Writings and Investigations Regarding the Acceptance and Usefulness of D as an Index of Discrimination

D as an index of discrimination is based on the difference between some proportion of the top or upper group on a test who correctly answered the item and some proportion of the bottom or lower group on the test who correctly answered the item. The most commonly accepted proportion is based on the work of Kelley (1939). Using various high level mathematical procedures Kelley concluded that dividing the test scores into the highest and lowest twenty-seven percent, results in extreme groups that maximize the following two conditions which are desired for the most valid results:

1. The extreme groups are as different as possible.

2. The extreme groups are as large as possible.

Kelley's findings have been so acceptable that they have also been applied to certain correlational item discrimination procedures such as Flanagan's index.

There have been other suggestions for group division based on different sized upper and lower proportions. Ebel (1951) discusses a study conducted by Aschenbrenner (1949) which demonstrated that the reliability of discrimination indices is greater from a division of test scores into upper and lower ten percent groups than from upper and lower twenty-seven percent groups when the extreme groups contained more than one hundred members each. Thus, when extremely large subject groups are used to pilot a set of items, the analysis can be satisfactorily undertaken with less scores and therefore time and effort can be reduced. Considering the factors of size of extreme groups and difference of extreme groups Aschenbrenner's findings intuitively seem logical. Extreme ten percent groups are indeed more divergent than extreme twenty-seven percent groups. Correspondingly, with a minimum of one hundred test papers per group, the extreme groups are quite large.

In a classroom situation, rarely if ever would there be sufficient numbers of students to comply with the diversion advocated by Aschenbrenner. Teachers, however, are quite busy and are concerned with ways to reduce the time necessary to perform more instructional activities. Diederich

(1960) recommends division of the group into halves and performance of the item analysis as a class activity after the test has been completed and scored. This procedure, according to Diederich, allows the teacher to save vast amounts of time and provide the students with a better understanding of the test, and provides a basis for class discussion relative to the items that were troublesome. Diederich suggests the division of the scores into upper and lower halves to allow all students an active role in the process. Mehrens and Lehmann (1973) also advocate division of classrooms into upper and lower halves. They contend that in testing settings where under forty students are administered the test, the upper-lower twenty-seven percent division produces indices that are not very reliable.

Brennan (1972) suggests that the division of scores based on extreme twenty-seven percent groups is appropriate only if the test score distribution is normal or at least symmetrical. In non-symmetrical distributions, Brennan recommends an upper-lower division based on different proportions in the extreme groups. One problem in following the model established by Brennan is that the numerical values of the indices obtained must be tested for statistical significance. Thus an item with a discrimination value of .45 could be significant while another item with a discrimination value of .45 might not be statistically significant even though both of the items were evaluated on the bases of the same test data.

Flanagan (1962) has demonstrated that by differencially counting the scores in the extreme groups the reliability of the indices generated will be higher. He suggested (1952) that the responses in the most extreme nine percent of the groups be counted twice and that the next extreme twenty percent be counted once. This is especially useful in addressing Mehrens and Lehmanns' concern of the reliability of the indices with small groups. Although there has been concern regarding the size of the extreme groups, Kelley's suggestion of twenty-seven percent seems to be the most widely accepted cutoff point for grouping.

A number of investigations have been concerned with establishing the advantages of using D as an index of discrimination. Hales (1972) has suggested that there are in excess of sixty different methods of determining an item's index of discrimination. The question then becomes which index of discrimination is most advantageous to employ with a set of items?

Ebel (1967) demonstrated the relationship of D and total test variance. Algebraically he determined that knowing only the D values of the item in a test he could compute a very close approximation of the variance of the test using the formula:

$$\sigma^2 = \frac{(\Sigma D)^2}{6}$$

Hence, it becomes apparent that the greater the D values of test items the more the total test variance. Taking this

one step further, Ebel has shown that since a test's reliability typically increases with an increase in variability, the larger the D values of a test, the higher the test's reliability. Thus to maximize the reliability of a test, items should be chosen that reflect high D values. D therefore is a good criterion for item selection.

Oosterhof (1973) investigated twenty-four indices of discrimination to determine which index displayed the highest degree of stability. He was concerned about whether an item's index of discrimination changed as a factor of the other items it is paired with to form the final test. Oosterhof's concern was directed toward the idea of using items from a pool. If an item's index of discrimination changed drastically when it was paired with different combinations of items, then just viewing an item's index of discrimination would not provide an adequate selection criterion. The results of Oosterhof's research indicated that of the twenty-four indices investigated, D and Gulliksen's item reliability index were the most stable. Thus it appears that an item's D value is more a factor of pure discrimination than the other indices viewed which seem to be more affected by the other items that constitute the test.

In an attempt to relate item discrimination indices to test reliability, Hales (1972) utilized an item pool of 601 high school social studies items. Based on a tryout of the item pool to 2,891 tenth and eleventh grade students, Hales determined each item's D value, Flanagan's r, and

Flanagan's r corrected for guessing (r_c). Using only the magnitude of the indices of discrimination mentioned above, six tests were developed (tenth grade D, tenth grade r, tenth grade r_c, eleventh grade D, eleventh grade r and eleventh grade r_c). The final tests were then administered to students and KR_{20} and odd even reliability coefficients calculated for each test. The results indicated no significant differences in the reliabilities (KR_{20} or odd even) on tests where D, r or r_c was used as the sole selection criterion. Hales concluded that D was the most appropriate index as it involves much less time than the other two and yet provides for just as reliable a test.

Lentz Jr., Hershstein, and Finch (1932) investigated D with three other indices of discrimination to discover which index when used as a criterion for selection produces the most reliable test. As a second concern these researchers considered the amount of time necessary to complete the computations to determine the various indices. The results indicated D produced the most reliable test. The differences in the obtained coefficients were not tested for significance. There was a tremendous difference in the time necessary to compute each of the indices. D required one-third to two-thirds the time as the next quickest index and one-fourth to one-half the time as the slowest index.

Krang (1952) compared D to two other indices of discrimination as item selection criterion in producing reliable tests and also the time necessary to determine

each of the indices. In this study the researcher used graduate level educational statistics items. The results are similar to the other studies reviewed here. D as a selection criterion produces tests as reliable as those produced using other more complicated indices and yet D involves much less time spent on calculation. Oosterhof (1976) used a factor analysis to compare D and eighteen other indices of discrimination. The data consisted of fifty verbal analogy items from Form M of the Differential Aptitude Test. The subjects in this study consisted of 1,000 tenth grade students. The results of the research suggests that when any of the nineteen indices investigated are used to evaluate the discriminating power of an item, the results are basically identical. Thus, Oosterhof suggests "preference toward a particular index would more appropriately be based on convenience of calculation or intuitive preference. It is inappropriate to suggest that using any of the common indices included in the present study has an appreciable effect on the eventual outcome of an analysis." He further states "Findley's difference index (D) in its simplicity of calculation and interpretation has much to recommend it."

Davis (1952) in a review of the studies relating to indices of discrimination states "The writer knows of no studies that have yielded conclusive evidence that one type of discrimination index is superior to another when each is properly used for selecting items. In fact, it

seems like

will lead

mean that

meritoriou

convenient

quire far

comments a

necessary

it as an a

item selec

Writings
Regarding
of an Ind
tion Base
Post Test

P

taining t

utility

indices

The rese

ease of

ferent f

similar.

the vali

lying th

referred

(1972) h

seems likely that the use of different types of indices will lead to the selection of similar items. This does not mean that all sorts of discrimination indices are equally meritorious. Some are apparently less deceptive and more convenient to use than others. It is obvious that some require far less computational labor than others." From the comments and the results of the studies concerning the time necessary to determine the indices, D has much to recommend it as an appropriate index of discrimination for use in item selection.

Writings and Investigations Regarding the Establishment of an Index of Discrimination Based on Pre and Post Test Groups

Prior to the late 1960's, most of the emphasis pertaining to indices of item discrimination focused upon the utility of various methods, and the establishment of the indices as valid selection criterion for developing tests. The research comparing various indices demonstrated that ease of computation and time necessary to compute was different for the indices but the quality of the results was similar. There appeared to be little question regarding the validity of the classical measurement concepts underlying the use of an index of discrimination.

The late 1960's saw the rebirth of what has been referred to as criterion-referenced measurement. Ebel (1972) has expressed the idea that criterion-referenced

tests have been in use in education for many years. He advocates that the percent grading used in the schools and universities until about 1920 are examples of criterion-referenced measurement. Chase (1974) illustrates that criterion-referenced measurement has always continued to be incorporated, in a limited sense, as part of the evaluation process within the schools. He cites typing tests where certain levels of proficiency are required for grades and physical-fitness courses where performance of certain skills are necessary to obtain a given level of reward as criterion-referenced measures.

Advocates of the criterion-referenced measurement of the late 1960's have questioned the application of classical measurement principles in developing and judging the quality of these new tests. Pophan and Husek (1969) have suggested that classical measurement theory is based on variance. In reference to indices of discrimination, these educators have pointed out that for an item to obtain a high index of discrimination the item must contain variance across the students who are administered the item. To obtain variance, some students must correctly answer the item and other students must incorrectly answer the item. These researchers contend that on criterion-referenced measures variance is not a necessary condition and hence the index of discrimination is not a satisfactory criterion for item selection.

Anderson (1972) agrees with Pophan and Husek about the unsuitability of indices of discrimination for criterion-referenced measurement. Anderson argues that "...items selected because they discriminate between these two groups will tend to contain difficult vocabulary or require inferences which are not necessarily critical to an understanding of the concepts and principles being tested. A test constructed to maximize discriminating power will emphasize aptitude and deemphasize achievement."

Woodson (1974) disagreed with the concept of variance being unimportant in criterion-referenced testing. He has argued that since in the real world there is always degrees of knowledge and interest on any topic or concern, tests and items measuring this interest must have variance to provide any useful information and thus warrant consideration for inclusion within a test of that interest area.

Millman and Popham (1974) have taken issue with Woodson and have instead stressed that "When the construction and selection of criterion-referenced test items are tampered with to maximize the test's validity to discriminate between groups, which is the case when variability is required, then the defining character of criterion-referenced tests is destroyed."

Helmstadter (1972) and Roudabush (1973) have suggested an index of discrimination which is appropriate for criterion-referenced tests. Both of these researchers agree that variability is not a quality issue with

criterion-referenced tests. Rather they view criterion-referenced tests as being utilized in pre and post test design. According to these researchers a good item is one that is incorrectly answered prior to instruction (pre test administration) and correctly answered after instruction (post test administration). Thus, they envision a criterion-referenced index of discrimination as being similar to D except the contrasting groups are the pre and post test administrations. Both of these researchers empirically tested their theories comparing the pre-post index with conventional indices to determine the similarities of the indices.

Helmstadter (1972) compared the pre-post index with D and a third index of discrimination. The items pool consisted of fifty-nine multivariate statistics items administered to twenty-eight students. The correlation between the pre-post index and D was .47. Helmstadter concluded from this that "these data clearly confirm the contention of those who have argued caution in using traditional item analysis procedures in criterion-referenced situations." However, Helmstadter also reports that seventy-one percent of the items would have similarly been selected or rejected by both pre-post and D. This relatively high percent of overlap would indicate the indices are not as independent as Helmstadter would have us believe.

Roudabush (1973) compared a pre-post index of discrimination to the point biserial index. In this investigation, a large group of items ($n > 1600$) were involved.

There is no report, however, on the number of subjects utilized except for a statement that in many cases there were too few subjects to obtain a stable estimate of the index of discrimination. Roudabush reports that less than half of the items selected as good by one index was also selected by the other index. The conclusion is reached that the two indices tend to select different items and are, therefore, two completely different measures. The report was quite clear in describing how the process of item selection might proceed. The findings appear to be questionable to the limited number of people actually attempting the item. Ebel (1951) has shown that indices of discrimination are relatively unstable when small numbers of people are administered the items.

Crehan (1974) compared three nominal criterion-referenced indices of discrimination with the point biserial index, a teacher selection and a random selection of items for test inclusion. The concern of this study was to determine which method of item selection would generate the most reliable and valid criterion-referenced test. The results indicate that "there is little evidence that item selection method effected resultant test reliability." However, the Cox-Vargus (pre-post design) and Brennan methods (similar to D but uses different extreme groups) produced the most valid tests. Therefore, it appears that the pre-post index is a good selection criterion for criterion-referenced tests. It must be remembered, however, that in this study the

reliability and validity formulas were designed for criterion-referenced tests but were not considered to be the most appropriate measure to determine criterion-referenced reliability and validity.

Carver (1972) (1974) has argued against the use of classical measurement indices of discrimination for test improvement. Carver, although not a declared advocate of criterion-referenced tests is in favor of the pre-post index of discrimination. Carver views a good test as one that is sensitive to growth within an individual as a factor of instructional intervention. Therefore, a good test is one that consists of items missed prior to instruction and correctly answered after instruction. Carver labels this kind of measurement edumetric. He sees norm-referenced measurement as insensitive to change due to instruction and refers to this type of measurement as psychometric. According to Carver, edumetric item discrimination (pre-post) is both conceptually and practically different than psychometric indices of discrimination. He further states that the differences are so great that utilizing an edumetric index for item selection will produce a different test than if psychometric indices are employed.

Green, Nyquist and Griffore (1975) agree with Carver's beliefs and theorize that "For formative evaluation, edumetric tests are more appropriate than the traditional normative or psychometric type of test. These educators further theorize that diagnostic tests fit under the heading

of formative evaluation. Therefore, diagnostic tests should be developed using a pre-post index of discrimination for item selection.

To lend credence to Carver's theory, Thomas (1976) and Reynolds and Cobean (1976) compared the edumetric index of discrimination with psychometric indices. Thomas (1976) compared the pre-post index of discrimination with D. In his study Thomas included twenty college level statistics items administered to 192 students. The items were ranked by each index and a Spearman Rho correlation was computed to discover the relationship between the indices. A coefficient of .10 was obtained and Thomas concluded that the indices were relatively independent of each other. As part of the procedure, Thomas altered items from the pre test to the post test and therefore the results of this study are questionable as a strict interpretation of a pre-post index implies the identical items utilized on both administrations of the test. A further limitation of these findings lies in the fact there were so few items in the item pool. A larger item pool would have permitted a more valid assessment of the similarities and differences of the indices of interest.

Reynolds and Cobean (1976) compared two edumetric indices with a psychometric index. The psychometric index used in the investigation was the point biserial. The data was obtained from thirty-six items administered to seventy-seven students. The top ten items from each index was chosen to be considered the test. The researchers found that based

on a KR_{20} reliability coefficient, the psychometric test was superior to either of the edumetric tests. However, when a reliability estimate was determined based on Carver's theories, the edumetric tests were superior to the psychometric test. The edumetric tests were also considered more edumetrically valid than the psychometric test. Like the study conducted by Thomas, this research endeavor was based on an extremely small sample of items. The findings therefore need to replicated with a much larger pool of items.

Writings and Investigations Regarding the Establishment of an Index of Discrimination based on Expert-Novice Groups

A review of the literature indicated only one reference directly relating to the use of expert-novice groups when determining an item's ability to discriminate. This appeared rather disheartening to this writer as the expert-novice grouping is theoretically the most proper measure of discrimination. Lennon (undated) defines discriminatory power as the ability of a test item to differentiate between persons possessing much of the same trait and those possessing little. This definition implies that there is some way to classify those who know the material being tested from those who do not know the material. In the case of the psychometric indices of discrimination, both correlational and difference methods, the assumption is made that a high total test score reflects competence in the material tested

and a low total test score reflects little or no competence in the area tested. Edumetric indices assume that instruction does make a person competent and thus a pre test score reflects not knowing while a post score reflects knowledge in the area.

In both situations, edumetric and psychometric, the criteria of competence is a single measure. With all that is known regarding the error factors that can come to light in any single test administration it is reasonable to expect a more rigorous criterion for competence. Davis (1952) suggests that the first item analysis that was performed for selection purposes was conducted by Alfred Binet. Binet's procedure was simply to note the percentage of a sampling of children at different age levels who could pass the item. The idea being that the lowest age group who could pass the item was found, this would become the age appropriateness of the item. Binet's model was not exactly expert-novice grouping, however, it did involve a criterion other than just score on the test.

Ebel, et al (1962) conducted an investigation to obtain some empirical data regarding the expert-novice index. In this study, experts in the field of measurement and evaluation were asked to submit questions they had used in their courses along with any item discrimination information they possessed regarding the submitted items. Ten experts and ten novices then attempted all of the items. Unfortunately, no comparisons were tested between the

expert-novice grouping and the other indices of discrimination. Therefore, no conclusions can be discussed regarding the similarities of the indices relative to the expert-novice index.

Writings and Investigations Regarding Different Item Formats

Mehrens and Lehmann (1973) view true-false items as essentially multiple choice items with two responses. Other authors of measurement textbooks view true-false and multiple choice items as being more distinct. Marshall and Hales (1971) and Ebel (1972) devote separate chapters to each of these item formats. Chase (1974) provides separate sections within a single chapter to these types of item formats. The distinctions these authors point out refers to the construction, use and special properties of these types of items. Marshall and Hales (1971) suggest that multiple choice items require more time to answer and therefore in a given testing time provide a less thorough coverage of the content. These authors also state that because of the increased influence of chance on true-false items, and the lower discriminating ability of these items, multiple choice tests will be more reliable than true-false tests of comparable length. Ebel (1972) advises that despite occasional exceptions "...it seems safe to say that most aspects of educational achievement that can be tested using one of the two forms can also be tested satisfactorily using the other."

Ebel also suggests that true-false are easier to write and if multiple choice items are properly converted to true-false items, there might be an increase in the reliability of the test. Ebel (1969) showed algebraically that the number of alternatives affects the reliability of the test. Thus, mathematically, these items formats produce different items.

There have been empirical studies investigating the differences between various formats of multiple choice items. Ranos and Stern (1973) investigated test reliability and indices of discrimination between four and five alternative items. In this study the researchers compared coefficient alpha reliability estimates between five alternative French and Spanish reading examination items with the same items after the least popular incorrect alternative had been removed. The results indicated a significant decrease in reliability but not a significant change in discrimination as measured by point biserial correlations. The researchers concluded that "the use of four--rather than five--choice items in language test construction should result in gains in test development of efficiency and lower cost per item. It should be pointed out, however, that these gains in efficiency must be traded off against losses in test reliability and item discrimination."

Costin (1972) investigated test reliability and item discrimination for three and four option multiple choice items. In this study Costin randomly deleted one distracter from each four alternative items to obtain the three

alternative items. Costin reports that there was no significant change in reliability as determined by a KR_{20} of an item discrimination as determined by point biserial correlations between the two item formats.

Frisbie (1973) compared the reliabilities of true-false and multiple choice tests. The true-false items used in this study were developed from two different item conversion methods applied to the multiple choice items. The final sample of items included seventy multiple choice items and two seventy item true-false tests. The results indicate that the multiple choice test was significantly more reliable than either true-false test.

The mathematical findings of Ebel (1969) and the empirical studied by Ramos and Stern (1973) and Frisbie (1973) provide a basis for separate index of discrimination analysis for true-false and multiple choice items. It appears these different items produce tests of differing reliabilities even though the item content is basically equivalent. This together with the findings of Ebel (1967) who has shown that a relationship exists between test reliability and the D values of the items comprising it, leads the researcher to conclude that any study comparing indices of discrimination should be performed separately for true-false and multiple choice items.

CHAPTER III

DESIGN AND PROCEDURES

Introduction

This research was designed to examine the relationship among three indices of discrimination (D, pre-post and expert-novice), the amount of item overlap that occurs when these indices are utilized as item selection criteria and the reliabilities associated with tests developed by using discrimination indices as item selection criteria. Two types of item formats, true-false and multiple choice, were analyzed separately to determine if the results of the preceding analyses were format specific or generalizable across various objective formats.

Sample Items

The items utilized in this study consisted of two item pools for introductory college level educational measurement courses offered at Michigan State University. Each of the item pools consisted of questions using a different item format. The multiple choice item pool contained sixty items. Included within this pool was three, four and five option items. These items have been generated and utilized for the introductory course in standardized tests

and measurements. The multiple choice items were written to measure knowledge and general understanding of the basic concepts of measurement, program evaluation, test construction, function and selection with standardized testing instruments. The following items are typical of those used in the multiple choice item pool:

1. The correlation between test scores and a criterion is a measure of

    a. causation
    b. objectivity
    c. reliability
    *d. validity
    e. variability

2. Which of the four sets of data below will produce the highest reliability of difference scores?

    a. $r_{xx} = .80$, $r_{yy} = .80$, $r_{xy} = .80$
    b. $r_{xx} = .50$, $r_{yy} = .50$, $r_{xy} = .40$
    c. $r_{xx} = .80$, $r_{yy} = .80$, $r_{xy} = .40$
    *d. $r_{xx} = .70$, $r_{yy} = .90$, $r_{xy} = .00$

The true-false item pool consisted of two hundred forty items. The items were designed for an introductory course in teacher-made tests and measures. The true-false items were designed to measure knowledge and understanding of the basic concepts relating to item construction, item revision, quality control measures of tests and general educational measurement principles. The following items are typical of those used in the true-false item pool:

1. Test scores corrected for guessing tend to correlate highly with uncorrected scores on the same test. (True)

2. Triviality and ambiguity are inherent weaknesses of true-false test items. (False)

It is readily apparent that there is a discrepancy between the number of items in the true-false item pool and the multiple choice item pool. This was intentionally planned as less time is necessary for students to answer true-false items than to answer multiple choice items. Thus in a given testing period more true-false items can be used than multiple choice items. Hence, true-false tests tend to include more items and require a larger item pool than multiple choice items.

These two item pools were selected for inclusion within this study because they were similar in many ways. Both sets of items were constructed by measurement specialists who were highly skilled in the principles of item construction. A second similarity of these items was the subject matter content that was measured. Although the courses that these items were intended for were not identical, the basic concepts addressed in these courses was extremely similar with the primary difference being the emphasis towards standardized tests in one course and teacher made tests in the other course. Attempts to provide both true-false and multiple choice items for the same course would have greatly reduced the number of items in the item pools and would, therefore, have created serious threats to the validity of the findings.

Another area of similarity between these item pools was the groups who were administered these items. Both sets of items were designed to be used with senior level education undergraduates or masters level education graduate students.

In selecting item pools that were similar, comparisons between item formats could be accomplished without fear of confounding results attributable to item construction quality, subject matter coverage or grade level of the students taking the items. Therefore, this study provides an indication of the generalized ability of the findings to various objective formats.

Sample Subjects

The subjects in this investigation consisted of three groups of people. The first group were those administered only the multiple choice items. This group consisted of seventy-seven students enrolled in three sections of Ed 464, an introductory course in standardized tests and measurements at Michigan State University. These students were upper-level undergraduate students in education or masters level graduate students in education. The students in this group were administered the sixty multiple choice items at the first session of the class and then re-administered the same items as the mid-term examination. Thus all students were administered all items in both a pre

and post test setting. To restrict confounding due to memory the following two conditions were instituted:

1. All pre test papers were numbered, collected and verified after the administration of the pre test. In this way students were unable to use the pre tests to study for the post test.

2. Students were told that the pre tests were old Ed 464 exams and were used as a diagnostic tool for the instructor to determine the specific needs of the class. The implication was conveyed that the course exams would contain different items on similar content but not the identical items.

The second group of students utilized in this study were administered only the true-false items. These students were enrolled in Ed 465, the introductory course in teacher made tests and measures at Michigan State University. These students were upper level education undergraduates and masters level education graduate students.

The normal format of this course includes a pre-test examination administered at the first class session. The pre test is utilized for the following three reasons:

1. As a diagnostic measure for the instructor to discover the entry levels of the students.

2. As an example to the students of the measurement concerns the instructor considers important.

3. As a motivational device to stimulate class discussions regarding specific measurement concerns.

The pre test typically has been a previous class examination, however, the post test examination for a particular class does not consist of the items which that class attempted in the pre test. The pre test and post test items therefore

were administered to different students. In a discussion with Dr. Robert L. Ebel, the professor responsible for this course, he suggested that the classes over time were quite homogeneous and that performances of items and tests were sufficiently stable that measures could be obtained from different classes without much threat to the validity of the findings. This appeared to be a legitimate assumption as the background, experiences and professions of these students are very similar.

To obtain a sufficiently large item pool, the true-false items were acquired from four sections of the course. Two sections were administered the items in the pre test form and two sections were administered the items in the post test form. The true-false group consisted of one hundred seventy-six students being administered the pre test and three hundred five students being administered the post test.

The third group of subjects utilized in this research were administered both true-false and multiple choice items. This group had been classified as experts in the area of educational measurement and evaluation. For use in this investigation, experts were defined as meeting at least one of the following criteria:

1. Earned doctorate degree in measurement and evaluation.
2. Experience teaching a course in measurement and evaluation.

3. Employment in a position requiring doctoral level skills in measurement and evaluation.

It was determined for this investigation that a novice would be defined as someone who had not previously taken the course for which he/she was currently enrolled. Therefore the pre test scores were also classified as novice scores. Discussions with instructors of Ed 464 and Ed 465 revealed that only a very minute proportion of students ever take both courses. Considering the large number of subjects included in this investigation, there was a possibility that a few students had taken the other course in measurement previously, and hence could not be classified as novices. However, the few exceptions would not create any significant discrepancies in the determination of the index as the concepts measured are similar but not identical.

Concern was also rendered as the pre test subjects would also be the novice subjects. It was felt that this might create a problem of dependency. A review of the literature and subsequent discussions with various authorities provided no concrete evidence as to how this would nullify the research findings. Therefore, it was decided that the analysis could continue and be executed as initially proposed.

Twenty experts agreed to participate as subjects in this investigation. Due to the extremely large number of items under investigation (240 true-false and 60 multiple choice) two test forms were developed to be administered to

the expert group. Each test form consisted of one hundred twenty true-false and thirty multiple choice items. Therefore, each item was administered to a group of ten experts. Table 3.1 provides a graphical description of the sample incorporated into this investigation.

Table 3.1 Description of the sample administered the true-false and multiple choice items.

Type of Item	Form	Pre Test	Post Test	Total
True-False	A	95	120	215
	B	81	185	266
Total		176	305	481
Multiple Choice	A	77	77	154
Total		77	77	154
Combination	A		10	10
	B		10	10
Total			20	20

Construction of Indices

Upper-Lower.

The upper-lower index of discrimination (D) was determined using the following formula:

$$D = \frac{UR - LR}{f}$$ Where

UR = proportion of people in the upper 27% who correctly answered the item

LR = proportion of people in the lower 27% who correctly answered the item

f = frequency of people in either group (since both groups consist of 27% of the total amount of people who were administered the item $fU = fL$

The D values for each item were computed using only the post test scores of the students. The pre test scores were not included as D is designed to provide an indication of how well an item can discriminate between those who know the material and those who do not know the material being tested. It was assumed that since these items were written for introductory courses, the pre test scores reflect little more than chance variance and therefore inclusion of the pre test scores would provide little additional information as to the quality of the items. This is especially true as this model has typically been used for single administration, norm-referenced type tests. The expert group was also excluded as this index is practically concerned with discrimination between the students within the course. Addition of experts would tend to supply an extraneous source of variance which would normally not be present in the determination of the index.

Pre-Post.

The pre-post index was computed using the following formula:

$$PP = \frac{\text{Post R}}{f\ \text{post}} - \frac{\text{Pre R}}{f\ \text{pre}} \quad \text{where}$$

Post R = Number of people who correctly answered the item on the post test

Pre R = Number of people who correctly answered the item on the pre test

post = Number of people who were administered the post test

f post = Number of people who were administered the post test

f pre = Number of people who were administered the pre test

This formula is algebraically equivalent to the formula for obtaining D. The difference lies in that D uses the same number of people in the upper and lower groups while this formula for the pre-post index allows the pre test group to contain a different number of people than the post test. This alteration of the D formula was necessary to determine the pre-post index of the true-false items in this study as these items were administered to unequally sized groups.

The pre-post analysis is based upon the criterion-referenced assumption that a good item is one that is missed on the pre test and answered correctly on the post test.

Expert-Novice

The expert-novice index of discrimination was based on the following formula:

$$EN = \frac{ER}{f_E} - \frac{NR}{f_N} \qquad \text{where}$$

ER = Number of experts who correctly answered item

NR = Number of novices who correctly answered the item

f_E = Number of experts

f_N = Number of novices

This formula is essentially the same as that used to determine the pre-post index. The only difference between the

formulas is in the groups that are to be compared. It was necessary to use this formula as the number of experts and novices was different. This was the case for both the true-false and multiple choice items.

This index was based on a stricter interpretation of the theoretical definition of an index of discrimination than D or the pre-post index. D uses total test score as the criterion for knowing or not knowing the material tested. The pre-post index uses the criterion of having received instruction for determining those who know or do not know the material. The expert-novice index incorporates a more exhaustive criterion for group membership and hence more closely follows the idea of knowing and not knowing. Table 3.2 provides the number of subjects included in each of the extreme groups of the three indices under investigation.

Table 3.2 Description of the sample used to calculate the three indices of discrimination

		Type of Extreme Group					
Type of test	Form	Upper 27%	Lower 27%	Pre test	Post test	Expert	Novice
True-False	A	32	32	95	120	10	95
	B	50	50	81	185	10	81
Multiple Choice	A	21	21	77	77	10	77

Hypotheses

This investigation was focused upon three major hypotheses each containing six sub-hypotheses. Following is a listing of the hypotheses tested:

Major Hypothesis 1

There is no significant correlation among three indices of discrimination for true-false or multiple choice items.

Sub-Hypotheses

1a. There is no significant correlation between D and pre-post indices of discrimination for true-false items.

1b. There is no significant correlation between D and pre-post indices of discrimination for multiple choice items.

1c. There is no significant correlation between D and expert-novice indices of discrimination for true-false items.

1d. There is no significant correlation between D and expert-novice indices of discrimination for multiple choice items.

1e. There is no significant correlation between pre-post and expert-novice indices of discrimination for true-false items.

1f. There is no significant correlation between pre-post and expert-novice indices of discrimination for multiple choice items.

Major Hypothesis 2

There is no significant overlap of items selected among three indices of discriminazion for true-false or multiple choice items.

Sub-Hypotheses

2a. There is no significant overlap of items selected by D and pre-post indices of discrimination for true-false items.

2b. There is no significant overlap of items selected by D and pre-post indices of discrimination for multiple choice items.

2c. There is no significant overlap of items selected by D and expert-novice indices of discrimination for true-false items.

2d. There is no significant overlap of items selected by D and expert-novice indices of discrimination for multiple choice items.

2e. There is no significant overlap of items selected by pre-post and expert-novice indices of discrimination for true-false items.

2f. There is no significant overlap of items selected by pre-post or expert-novice indices of discrimination for multiple choice items.

Major Hypothesis 3

There is no significant difference in the reliabilities of tests where items are selected by three indices of discrimination for true-false and multiple choice items.

Sub-Hypotheses

3a. There is no significant difference in the reliability of true-false tests where items are selected by D and pre-post indices of discrimination.

3b. There is no significant difference in the reliability of multiple choice tests where items are selected by D and pre-post indices of discrimination.

3c. There is no significant difference in the reliability of true-false tests where items are selected by D and expert-novice indices of discrimination.

3d. There is no significant difference in the reliability of multiple choice tests where items are selected by D and expert-novice indices of discrimination.

3e. There is no significant difference in the reliability of true-false tests where items are selected by pre-post and expert-novice indices of discrimination.

3f. There is no significant difference in the reliability of multiple choice tests where items are selected by pre-post and expert-novice indices of discrimination.

Analysis

Major hypothesis one and its accompanying sub-hypotheses were analyzed using Pearson product-moment correlations. This analysis allows the investigator to compare two valuables at a time to determine if the obtained correlation was statistically different from zero (no correlation

or relationship between the variables). It is important to remember that when a large subject pool is involved as was the case in this investigation, a very small correlation can be shown to be statistically significant. To fully interpret the results of a correlational analysis, one should incorporate a two step process. The first step involves testing to see if the obtained correlation is statistically significant and the second step involves analyzing the magnitude of the coefficient to determine its meaningfulness. For example, a correlation of .3 might be significant, however, with such a small obtained coefficient the relationship between the variables is quite small and perhaps not meaningful. Following is the formula for obtaining the Pearson product-moment correlational coefficient:

$$r_{xy} = \frac{\Sigma(x-\bar{x})\ (y-\bar{y})}{NS_xS_y} \quad \text{where}$$

x = an item's score on the first index.

$\bar{x}$ = the mean score for the first index.

y = an item's score on the second index.

$\bar{y}$ = the mean score for the second index.

N = the number of items under investigation.

S_x = the standard deviation of the first index.

S_y = the standard deviation of the second index.

The test of statistical significant of the correlation has been described by Glass and Stanley (1970) as simply computing the coefficient between two variables and

comparing the obtained coefficient with a tabled coefficient with N-2 degrees of freedom. If the obtained coefficient is larger than the tabled coefficient the relationship is statistically significant.

This type of analysis allows for only pairwise comparisons of the variables. Therefore, to compare the three variables of interest in this investigation across the two types of item formats, six separate analyses had to be performed. This, however, created a problem referred to as inflation of the alpha level or chance factor in the testing. This problem occurs whenever multiple hypotheses are tested in a simple experiment. The alpha level refers to the error tolerance allowable in the analysis. Many if not most educational research endeavors have utilized an alpha level of .05. This means that to be significant the findings could be obtained through some chance factor no more than five times out of one hundred or five percent. In an investigation, however, the researcher must sum the alpha level across all tests to obtain a total or experimental alpha level. In dealing with major hypothesis one, the experimental alpha level would be .3 if all six sub-hypotheses were analyzed using an alpha of .05 (.05 x 6 = .3). To avoid this problem, all sub-hypotheses were analyzed using an alpha of .01 (less error tolerance on each analysis). The experimental alpha, therefore, was .06.

A more appropriate analysis of hypothesis one would have been to test the obtained correlational coefficient

against unity. This would have provided the answer to the question are the indices identical? Lord (1957), McNemar (1958) and Forsyth and Feldt (1969) (1970) have devised methods to test correlational coefficients different than unity. However, with the type of data available in this investigation these analyses could not be performed. In the limitation section of Chapter Five of this document a complete explanation of this problem is provided.

The second major hypothesis and its corresponding sub-hypotheses were concerned with the amount of overlap in item selection that occurs when the various indices of discrimination are utilized as the sole criteria for selection. For the true-false items, the best one hundred twenty items have been selected for each index of discrimination and for multiple choice, the best thirty items were selected. Analyses were then performed using a phi-coefficient to discover if the amount of overlap associated with the different indices was significant. Amount of overlap was defined for this investigation as the number of items that were jointly accepted or rejected by the indices under examination. The phi-coefficient has been described by Glass and Stanley (1970) as being "simply the Pearson product-moment coefficient of correlation for dichotomous data." In this investigation the dichotomy was being selected or not being selected. The formula for determining the phi-coefficient was given by Magnusson (1966):

$$r_{phi} = \frac{p_{ik} - p_i p_k}{\sqrt{p_i q_i p_k q_k}}$$ where

p_{ik} = the proportion of items selected by both indices

p_i = the proportion of items selected by the first index.

p_k = the proportion of items selected by the second index.

q_i = the proportion of items not selected by the first index.

q_k = the proportion of items not selected by the second index.

As the phi-coefficient was based on the Pearson product-moment correlational coefficient only pairwise comparisons are permitted. Therefore, to address the three indices and two item formats six analyses must be performed. To avoid an exceedingly inflated alpha level each of the sub-hypotheses were analyzed with an alpha of .01. Glass and Stanley (1970) demonstrate that when using the phi-coefficient with $n > 20$ the population mean is zero with a standard deviation of one. Therefore, to test significance the test statistic used was as follows:

$$Z = \sqrt{n}\ \phi$$ where

Z = the obtained figure to compare to the unit normal distribution.

n = the number of items used.

ϕ = the obtained phi coefficient.

The third major hypothesis and its accompanying sub-hypotheses was generated to answer the following question:

Using an index of discrimination as the only criteria for item selection, will different indices produce more reliable tests?

To empirically investigate this concern, the first step was to develop the tests using the three indices of discrimination. As was discussed in reference to major hypothesis two, using each index separately, the best one hundred twenty true-false items were selected to develop the final true-false tests and the best thirty multiple choice items were considered the final multiple choice tests. Kuder-Richardson Formula 20 reliabilities were then computed for each of the six tests. The formula for the Kuder-Richardson Formula 20 reliability is as follows:

$$r_{xx} = \frac{K}{K-1}\left[1 - \frac{\Sigma pq}{\sigma^2}\right] \quad \text{where}$$

r_{xx} = the reliability of the test.

K = the number of items in the test.

p = the proportion of people who correctly answered the item.

q = the proportion of people who incorrectly answered the item.

σ^2 = the test variance.

Due to a limitation in the data (the answer sheets for each student was not available), conventional computation of the test variance was impossible. Ebel (1967) demonstrated that test variance could be estimated using the following formula:

$$\sigma^2 = \frac{(\Sigma D)^2}{6}$$ where

D = upper-lower discrimination index for each item.

Thus using this approximation for test variance the reliabilities for each test could be estimated. The next step in the process was to empirically compare these coefficients. The following test statistic has been described by Glass and Stanley (1970) for comparisons of the differences in obtained coefficients:

$$Z = \frac{Z_{r_1} - Z_{r_2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$ where

Z = the calculated sample value of the unit normal distribution to be compared to the population unit normal distribution.

Z_{r_1} = the reliability of one of the tests converted to a z score using Fischer's z - transformation of r_{xy}.

Z_{r_2} = the reliability of the other test converted to a z score using Fischer's z - transformation of r_{xy}.

n_1 = number of items in the first test.

n_2 = number of items in the second test.

As was the case with the previous hypotheses this test statistic can only handle pair-wise comparisons. Thus six analyses needed to be completed to compare all indices and item formats. Each of the analyses was performed using an alpha of .01.

Summary

This investigation utilized a two hundred forty item true-false item pool and a sixty item multiple choice item pool. These item pools were selected on the basis of their sharing the following attributes:

1. Quality of the items.
2. Subject matter content.
3. Grade level of the students.

The true-false items were administered to two sections of an introductory course in teacher made tests and measurements as a pre test and also to two other sections of the same course as a post test. The multiple choice items were administered to three sections of an introductory course in standardized tests and measurements as both a pre and post test.

A group of measurement and evaluation experts were administered both the true-false and multiple choice items.

For each of the items D, pre-post, and expert-novice indices of discrimination was generated. Pearson product-moment correlational coefficients were computed for each pairwise combination of indices across both item formats to test the degree of relationship between the indices. For each index the best one hundred twenty true-false items and the best thirty multiple choice items were selected to be considered the final test. Using phi-coefficients the amount of item overlap was tested for each combination of indices and item formats.

Kuder-Richardson Formula 20 reliability coefficients were calculated for each of the final tests. Comparisons were then made to test the differences in the obtained reliability coefficients.

CHAPTER IV

RESULTS

Introduction

This chapter is divided into five major sections. The first section contains the results of correlational comparisons among the three discrimination indices of interest in this investigation.

The second section deals with the findings regarding the similarities of the indices relative to item selection. These similarities are reported jointly as phi-coefficients and as the percent of overlap between the indices.

Results that reflect the reliabilities associated with tests composed of items which used indices of discrimination as the sole criteria of selection are presented in the third section. The results of four supplementary analyses are presented in the fourth section. A final section, the chapter summary, follows.

Results Concerning Correlational Analyses of the Indices

Values for D, pre-post, and expert-novice indices of discrimination were calculated for each of the two hundred forty items in the true-false item pool and sixty items in the multiple choice item pool. Tables 4.1 and 4.2

TABLE 4.1.--D, Pre-Post, and Expert-Novice Values Associated With True-False Items.

	Indices		
Item Number	D	Pre-Post	Expert-Novice
1	.03	.17	.26
2	.20	.17	.42
3	.41	.24	.31
4	.07	.22	.28
5	.41	.50	.57
6	.26	.13	.27
7	-.05	.05	.17
8	.33	.07	-.19
9	.20	.17	.25
10	.06	-.02	.15
11	.26	.36	.34
12	.33	.35	.52
13	.08	.12	.12
14	.25	.48	.30
15	.36	.12	.33
16	.23	.28	.04
17	-.13	-.25	-.51
18	.51	.38	.64
19	.00	-.11	-.18
20	.02	-.04	-.40
21	.49	.44	.50
22	.31	.10	.23
23	.16	.25	.21
24	.31	.28	.24
25	.34	-.19	-.50
26	.25	.63	.57
27	.33	.48	.55
28	.41	.36	.21
29	.35	.28	.34
30	.54	.57	.35
31	.33	.54	.70
32	.41	.38	.28
33	.28	.25	.41
34	.23	.59	.70
35	.33	.40	.57
36	.46	.26	.14
37	.07	-.08	.18
38	.18	-.34	.05

TABL

Ite

TABLE 4.1.--Continued.

	Indices		
Item Number	D	Pre-Post	Expert-Novice
39	.28	.30	.31
40	.21	-.12	-.14
41	.43	.39	.52
42	.30	.04	.11
43	.31	-.01	.18
44	.23	.82	.90
45	.49	.34	.47
46	.03	-.16	.05
47	.13	.59	.42
48	-.13	-.49	-.10
49	.31	.58	.60
50	.33	.36	.60
51	.25	-.08	.05
52	.41	.48	.26
53	.28	-.07	.26
54	.07	.17	.31
55	.54	.18	-.31
56	.36	.53	.61
57	.46	.46	.63
58	.46	.68	.76
59	.13	.12	.21
60	-.31	-.15	.12
61	.08	-.31	.12
62	.25	.36	.07
63	.30	.44	.54
64	.36	.50	.34
65	-.20	-.32	-.10
66	.23	-.14	-.01
67	.36	.16	.49
68	.41	.39	.29
69	.08	.66	.69
70	.46	.56	.78
71	.31	.43	.18
72	.33	.51	.58
73	.07	.10	.23
74	.33	.12	.05
75	.43	.22	.26
76	.44	.04	.10
77	.06	-.07	.12
78	.54	.62	.47
79	.25	.28	.49
80	.23	.37	.18

TABLE 4.1.--Continued.

	Indices		
Item Number	D	Pre-Post	Expert-Novice
81	.69	.42	.81
82	.25	.30	.64
83	.25	.35	.38
84	.13	-.11	-.58
85	.26	.25	-.30
86	.25	-.03	.21
87	.16	.35	-.07
88	.28	.18	.14
89	.13	.34	.28
90	.49	.31	.40
91	.44	.41	.64
92	.51	.22	.30
93	.30	.48	.33
94	.10	.37	.45
95	.28	.42	.49
96	.18	.38	.09
97	.25	.26	.19
98	.20	.72	.46
99	.38	.73	.37
100	.08	.17	-.06
101	.23	.00	-.46
102	.36	.21	.30
103	.23	-.12	-.18
104	.33	.37	.35
105	.18	.17	.29
106	.33	.25	.51
107	.64	.61	.91
108	.51	.55	.63
109	.28	-.11	.30
110	.35	.59	.80
111	.31	.56	.37
112	.41	.24	.13
113	-.08	.00	.26
114	.35	.45	.71
115	.03	.05	.06
116	.10	.10	-.15
117	.10	.51	.30
118	.07	.08	-.09
119	.33	.33	.20
120	.39	.18	.19

TABLE 4.1.--Continued.

	Indices		
Item Number	D	Pre-Post	Expert-Novice
121	.32	.23	.24
122	.03	.03	.29
123	.04	-.04	.02
124	.22	.27	.08
125	.10	.01	.49
126	.17	.54	.48
127	.29	.35	.14
128	.16	.06	.35
129	.07	.31	-.05
130	.32	.18	.27
131	-.06	-.31	.07
132	.16	.06	.16
133	.03	.16	.02
134	.22	.55	.52
135	.10	.25	-.06
136	.03	.06	-.01
137	.48	.23	.51
138	.49	.40	.17
139	.39	.58	.36
140	.03	.50	.53
141	.10	-.04	.19
142	.16	.36	.38
143	.00	.50	.17
144	.10	.46	.54
145	.19	.50	.25
146	.03	.11	-.21
147	.10	.50	.14
148	.16	.17	.12
149	.13	.19	.25
150	.13	.12	.14
151	.20	.12	.07
152	.13	.10	.17
153	.03	.56	.57
154	.39	.49	.60
155	.26	.54	.33
156	.39	.37	.26
157	.35	.40	.33
158	.16	.15	.23
159	.68	.45	.31
160	.00	.35	.26

TABLE 4.1.--Continued

	Indices		
Item Number	D	Pre-Post	Expert-Novice
161	.13	.16	.18
162	.49	.32	.48
163	.10	.01	-.05
164	.07	.38	.26
165	.16	.15	-.38
166	.10	.68	.45
167	.10	.22	.26
168	.16	.30	.11
169	.32	.37	.45
170	.23	.18	.26
171	.10	.42	.19
172	.36	-.25	-.26
173	.20	.37	.69
174	.32	.48	.46
175	.39	.52	.36
176	.26	.09	.57
177	.16	.36	.15
178	-.06	.07	.16
179	.10	.15	.08
180	.35	.11	.16
181	.39	.18	.30
182	.09	.11	.11
183	.13	.59	.23
184	.13	.52	.21
185	.23	.57	.57
186	.03	.56	.49
187	.07	.30	.35
188	.10	.06	.15
189	.58	.45	.31
190	.00	.12	.12
191	.26	.14	.28
192	.46	.17	.37
193	.23	-.02	.05
194	.19	.44	.45
195	.26	.34	.37
196	.16	.18	.14
197	.33	.15	.59
198	.06	.51	.44
199	.55	.38	.23
200	.48	.30	.12

TABLE 4.1.--Continued.

	Indices		
Item Number	D	Pre-Post	Expert-Novice
201	.26	.12	-.04
202	.13	.47	.44
203	.36	.58	.79
204	.10	.25	.41
205	.16	.69	.59
206	.32	.40	.13
207	.03	.44	.48
208	.20	.08	.07
209	.16	.08	.23
210	.13	.45	.49
211	.26	.05	-.13
212	.10	.30	.25
213	-.04	.10	.43
214	.13	.05	.14
215	.16	-.16	-.36
216	.39	.53	.59
217	.04	.04	.15
218	.26	.02	.07
219	.32	.13	.12
220	.26	.49	.50
221	-.03	.12	.30
222	.26	-.01	.00
223	.00	.11	.11
224	.07	.13	.07
225	.10	-.05	.06
226	.23	.25	.33
227	.35	.22	.29
228	.23	-.01	.25
229	.32	.04	.23
230	.23	.23	.34
231	.10	.07	.12
232	.06	.03	.33
233	.07	.01	.03
234	.10	.02	.09
235	.13	.25	.33
236	.13	.03	.00
237	.03	.01	.19
238	.10	.37	-.47
239	.45	.29	.47
240	.29	.30	.57

TABLE 4.2.--D, Pre-Post, and Expert-Novice Values Associated With Multiple Choice Items.

	Indices		
Item Number	D	Pre-Post	Expert-Novice
1	.15	.06	.13
2	.45	.01	.29
3	.15	-.06	.10
4	.25	.16	.47
5	.30	.05	-.02
6	.10	.14	.59
7	-.05	.29	.19
8	.05	.11	.19
9	.20	.04	.60
10	.20	.20	.34
11	.15	.48	.46
12	.45	.19	.60
13	.25	.36	.48
14	.45	.24	.27
15	.10	.20	.18
16	.50	-.16	.30
17	.15	.24	.76
18	.40	.42	.21
19	.55	.23	.54
20	.30	.11	.65
21	.05	.03	-.03
22	.35	.09	.33
23	.35	.42	.68
24	.30	.35	.61
25	.35	.08	.54
26	.25	.24	.99
27	.40	.77	.85
28	.05	.44	.50
29	.40	.52	.69
30	.85	.16	.34
31	.65	.32	.84
32	.10	-.12	.26
33	.45	.43	.53
34	.25	.12	.51
35	.05	.19	.19
36	.35	.19	.25
37	-.20	-.06	.73
38	.30	.21	.15
39	.25	.31	.60

TABLE 4.2.--Continued.

	Indices		
Item Number	D	Pre-Post	Expert-Novice
40	.35	-.23	.49
41	.05	.12	.01
42	.50	.30	.43
43	.15	.47	.67
44	.35	.13	.70
45	.35	.24	.38
46	.30	.41	.62
47	.30	.44	.57
48	.65	.37	.65
49	.75	.27	.58
50	.15	.53	.42
51	.40	.32	.62
52	-.10	.26	.44
53	-.05	-.07	-.29
54	.30	.12	.71
55	.20	-.02	.64
56	.40	.28	.74
57	.25	.13	.50
58	.35	.18	.61
59	.00	-.10	-.11
60	.00	-.60	-.53

display the results of these computations. The items were identified by a code number. The listing of the items and the correct answer for each item can be found in appendices one and two.

The first hypothesis of interest that was stated in Chapter III was:

There is no significant correlation among three indices of discrimination for true-false or multiple choice items.

Sub-Hypotheses

1a. There is no significant correlation between D and pre-post indices of discrimination for true-false items.

1b. There is no significant correlation between D and pre-post indices of discrimination for multiple choice items.

1c. There is no significant correlation between D and expert-novice indices of discrimination for true-false items.

1d. There is no significant correlation between D and expert-novice indices of discrimination for multiple choice items.

1e. There is no significant correlation between pre-post and expert-novice indices of discrimination for true-false items.

1f. There is no significant correlation between pre-post and expert-novice indices of discrimination for multiple choice items.

Pearson product-moment correlations were calculated to test the pairwise combinations of the three indices and the two item formats. A visual inspection of the data reported in Table 4.3 indicated that all of the obtained correlational coefficients were significant at the .01 level except the correlation between D and pre-post indices for

TABLE 4.3.--Pearson Product-Moment Correlational Coefficients for D, Pre-Post, and Expert-Novice Indices of Discrimination With True-False and Multiple Choice Items.

Item Format	Correlated Indices	Obtained Correlation	Level of Significance
True-false	D, Pre-post	.4102	.001
	D, Expert-novice	.3681	.001
	Pre-post, Expert-novice	.6683	.001
Multiple choice	D, Pre-post	.2790	.015
	D, Expert-novice	.3758	.002
	Pre-post, Expert-novice	.5580	.001

multiple choice items. Based on the alpha level determined prior to the analysis sub-hypothesis 1b could not be rejected. It was therefore concluded that with the aforementioned exception, there was a significant relationship between the indices.

It is interesting to note that with both true-false and multiple choice items the greatest degree of relationship was found between the pre-post and expert-novice indices of discrimination. A Z-test of significance was employed to test the difference between the obtained correlational coefficients. This test statistic described by Glass and Stanley (1970), uses the following formula:

$$Z = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_1-3}}}$$

The results of this additional analysis indicate that with true-false items the correlation between pre-post and expert-novice indices of discrimination was significantly greater than the other true-false pairwise correlations but with multiple choice items there were no significant differences between any of the obtained correlational coefficients. This same test statistic was applied to test the difference between the indices across the two item formats. The results indicate that in all cases the obtained coefficient for the true-false items were not statistically different than the obtained coefficient for the multiple choice items. It can be concluded, therefore, that the relationship between the

indices is similar regardless of whether the items are true-false or multiple choice.

Results Concerning the Similarities of the Indices as Item Selection Criteria

The second hypothesis of this investigation was stated in Chapter III as:

There is no significant overlap of items selected among three indices of discrimination for true-false or multiple choice items.

Sub-Hypotheses

2a. There is no significant overlap of items selected by D and pre-post indices of discrimination for true-false items.

2b. There is no significant overlap of items selected by D and pre-post indices of discrimination for multiple choice items.

2c. There is no significant overlap of items selected by D and expert-novice indices of discrimination for true-false items.

2d. There is no significant overlap of items selected by D and expert-novice indices of discrimination for multiple choice items.

2e. There is no significant overlap of items selected by pre-post and expert-novice indices of discrimination for true-false items.

2f. There is no significant overlap of items selected by pre-post or expert-novice indices of discrimination for multiple choice items.

To test this hypothesis, the median value was determined for each index separately from the true-false and multiple choice item pools. Those items that had index values exceeding the median were considered as selected for the

final test and those items possessing index values below the median were rejected from inclusion within the final test. Tables 4.4 and 4.5 display a summary of which items were selected from each index. These tables also provide an indication of which items were selected by more than one index. These data were then placed into two by two contingency tables and phi-coefficients were calculated. The phi-coefficients were then tested for significance using the formula presented in Chapter III. The obtained phi-coefficients and the accompanying Z values are presented in Table 4.6

As was explained in Chapter III, all sub-hypotheses were conducted with an alpha level of .01. Thus for this analysis the critical value of z was 2.32. Any obtained z that exceeded 2.32 was considered to indicate a significant relationship. An inspection of Table 4.6 reveals that with true-false items there was a significant relationship between the indices as to the items they selected for inclusion in the final test. With multiple choice items, however, significant relationships were not found among the indices. It appeared that the indices as item selection criteria were related with true-false items but may not be related when dealing with multiple choice items.

Based on the results of these data, sub-hypotheses 2a, c, and e which dealt with true-false items were rejected and subhypotheses 2b, d, and f which dealt with multiple

TABLE 4.4.--Selection of True-False Items into a Final Test by D, Pre-Post, and Expert-Novice Indices of Discrimination.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
1	x							
2				x				
3						x		
4				x				
5								x
6				x				
7	x							
8		x						
9	x							
10	x							
11								x
12					x			
13	x							
14								x
15						x		
16					x			
17	x							
18								x
19	x							
20	x							
21								x
22		x						
23	x							
24					x			
25		x						

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
26								x
27								x
28					x			
29								x
30								x
31								x
32								x
33								x
34								x
35								x
36					x			
37	x							
38	x							
39								x
40	x							
41								x
42		x						
43		x						
44								x
45								x
46	x							
47							x	
48	x							
49								x
50								x

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
51		x						
52								x
53						x		
54				x				
55		x						
56								x
57								x
58								x
59	x							
60	x							
61	x							
62					x			
63								x
64								x
65	x							
66	x							
67						x		
68								x
69							x	
70								x
71					x			
72								x
73	x							
74		x						
75					x			

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
76		x						
77	x							
78								x
79								x
80			x					
81								x
82								x
83								x
84	x							
85		x						
86		x						
87			x					
88		x						
89							x	
90								x
91								x
92						x		
93								x
94							x	
95								x
96			x					
97					x			
98							x	
99								x
100	x							

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
101	x							
102						x		
103		x						
104								x
105				x				
106								x
107								x
108								x
109						x		
110								x
111								x
112		x						
113				x				
114								x
115	x							
116	x							
117							x	
118	x							
119					x			
120		x						
121		x						
122				x				
123	x							
124			x					
125				x				

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
126							x	
127					x			
128				x				
129			x					
130						x		
131	x							
132	x							
133	x							
134							x	
135	x							
136	x							
137						x		
138					x			
139								x
140							x	
141	x							
142							x	
143			x					
144							x	
145			x					
146	x							
147			x					
148	x							
149	x							
150	x							

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
151	x							
152	x							
153							x	
154								x
155								x
156								x
157								x
158	x							
159								x
160			x					
161	x							
162								x
163	x							
164							x	
165	x							
166							x	
167	x							
168			x					
169								x
170				x				
171			x					
172		x						
173							x	
174								x
175								x

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
176						x		
177			x					
178	x							
179	x							
180		x						
181						x		
182	x							
183			x					
184			x					
185								x
186							x	
187							x	
188	x							
189								x
190	x							
191						x		
192						x		
193	x							
194							x	
195								x
196	x							
197						x		
198							x	
199					x			
200					x			

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
201		x						
202							x	
203								x
204							x	
205							x	
206					x			
207							x	
208	x							
209	x							
210							x	
211		x						
212			x					
213				x				
214	x							
215	x							
216								x
217	x							
218		x						
219		x						
220								x
221				x				
222		x						
223	x							
224	x							
225	x							

TABLE 4.4.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
226							x	
227						x		
228	x							
229		x						
230						x		
231	x							
232		x						
233	x							
234	x							
235							x	
236	x							
237	x							
238			x					
239								x
240								x
TOTAL	66	25	16	12	14	17	26	64

TABLE 4.5.--Selection of Multiple Choice Items into a Final Test by D, Pre-Post, and Expert-Novice Indices of Discrimination.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
1	x							
2		x						
3	x							
4	x							
5	x							
6				x				
7			x					
8	x							
9				x				
10			x					
11			x					
12						x		
13			x					
14					x			
15			x					
16		x						
17							x	
18					x			
19								x
20						x		
21		x						
22	x							
23								x
24								x
25						x		

TABLE 4.5.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
26							x	
27								x
28							x	
29								x
30		x						
31								x
32	x							
33								x
34				x				
35	x							
36		x						
37				x				
38					x			
39							x	
40		x						
41	x							
42					x			
43							x	
44						x		
45					x			
46								x
47								x
48								x
49								x
50				x				

TABLE 4.5.--Continued.

Item Number	Selected By: None of the Indices	D Only	Pre-Post Only	Expert-Novice Only	D and Pre-Post	D and Expert-Novice	Pre-Post and Expert-Novice	All Indices
51								x
52			x					
53	x							
54						x		
55				x				
56								x
57				x				
58						x		
59	x							
60	x							
TOTAL	12	6	7	6	5	6	5	13

TABLE 4.6.--Phi-coefficients for True-False and Multiple Choice Items on Comparisons of Each Index of Discrimination.

Item Format	Indices Compared	ϕ	z
True-False	D, Pre-post	.315	4.880*
	D, Expert-novice	.354	5.484*
	Prepost, Expert-novice	.510	7.901*
Multiple Choice	D, Pre-post	.204	1.580
	D, Expert-novice	.277	2.146
	Pre-post, Expert-novice	.204	1.580

*Significant at .01

choice items could not be rejected. Therefore, it was concluded that for true-false items, the indices were related but for multiple choice no significant relationships were found. In item selection it appeared, therefore, that the indices behaved differently depending upon which item format was involved.

As a post hoc analysis the percent of overlap (proportion of items selected by both indices or rejected by both indices) in each pairwise combination of indices was determined. Table 4.7 provides the actual number of items that were jointly accepted or rejected and the percent of overlap associated with each pairwise combination of indices.

This analysis was not designed to test significant relationships between the indices. Rather, it was provided as an alternative method to determine how similarly the indices behaved as item selection criterion with the items in this study.

In all situations at least fifty percent of the items were jointly accepted or rejected (chance would dictate a twenty-five percent overlap). In the case of pre-post and expert-novice indices for true-false items, seventy-five percent of the items on the final test were common.

Results Concerning the Reliabilities of Tests Using Different Indices of Discrimination as Selection Criterion

Estimations of Kuder-Richardson Formula 20 reliability coefficients were computed for each of the final three

TABLE 4.7.--Percent of Overlap of Items Selected by D, Pre-Post and Expert-Novice Indices of Discrimination.

Format	Indices Compared	Number of Items Jointly Accepted or Rejected	Percent of Overlap
True-False	D, Pre-post	154	65
	D, Expert-novice	160	67
	Pre-post, Expert-novice	180	75
	All Indices	118	49
Multiple Choice	D, Pre-post	30	60
	D, Expert-novice	38	63
	Pre-post, Expert-novice	30	60
	All Indices	25	42

true-false and multiple choice tests. These coefficients can be found in Table 4.8. The estimations of the Kuder-Richardson Formula 20 reliability coefficients incorporated the formula $\frac{(\Sigma D)^2}{6}$ to estimate the total test variance. This procedure was necessary as the raw data did not contain the actual test sheets of the subjects tested for the pre or post versions of the true-false tests. To avoid any biasing effects due to this estimate of the test variance, all reliability coefficients were computed in the same way.

The hypothesis for which these coefficients were developed was stated in Chapter III as:

There is no significant differences in the reliabilities of tests where items were selected by three indices of discrimination for true-false and multiple choice items.

Sub-Hypotheses

3a. There is no significant difference in the reliability of true-false tests where items were selected by D and pre-post indices of discrimination.

3b. There is no significant difference in the reliability of multiple choice tests where items were selected by D and pre-post indices of discrimination.

3c. There is no significant difference in the reliability of true-false tests where items were selected by D and expert-novice indices of discrimination.

3d. There is no significant difference in the reliability of multiple choice tests where items were selected by D and expert-novice indices of discrimination.

3e. There is no significant difference in the reliability of true-false tests where items were selected by pre-post and expert-novice indices of discrimination.

3f. There is no significant difference in the reliability of multiple choice tests where items were selected by pre-post and expert-novices indices of discrimination.

TABLE 4.8.--Estimates of Kuder-Richardson Formula 20 Reliability Coefficients for Final Tests where Items were Selected Using Different Indices of Discrimination.

Test Format	Discrimination Index	Estimated Kuder-Richardson Formula 20	r to z Conversion
True-false	D	.935	1.697
	Pre-Post	.917	1.575
	Expert-novice	.910	1.528
Multiple choice	D	.790	1.071
	Pre-post	.438	.470
	Expert-novice	.599	.692
Adjusted multiple choice	D	.938	1.721
	Pre-post	.757	.990
	Expert-novice	.857	1.284

An inspection of the reliability coefficients showed that the true-false items, regardless of the index used for item selection, provides for a more reliable test than does the multiple choice items. However, this finding was not surprising as the number of items in the true-false tests were four times larger than the number of items in the multiple choice tests. Table 4.8 displays the reliability estimates for the multiple choice tests after they had been adjusted to match the length of the true-false tests. This equating procedure was based on the Spearman-Brown formula for estimating the reliability of a lengthened test. Following is the Spearman-Brown formula:

$$r_{xx_n} = \frac{nr_{xx}}{1 + (n-1)r_{xx}} \quad \text{where}$$

r_{xx_n} = The reliability of the test lengthened n times.

n = The number of times the test is increased in length.

r_{xx} = The estimate of reliability of the original test.

Adjustment of the multiple choice tests revealed that the reliability estimates were considerably closer. In one case (D) the multiple choice test actually exceeded the reliability of the true-false test. The test statistic used for this analysis was discussed in Chapter III. The first step in using this test statistic involved transforming the obtained reliability coefficients to z scores

using Fisher's r to z tables. Table 4.8 includes these converted z scores.

Examination of Tables 4.9 revealed that no significant differences in reliability regardless of which index of discrimination was employed as the sole criterion of item selection. These same results were found when comparisons were made across the different item formats. It was therefore concluded that reliability of true-false or multiple choice tests were not significantly different when D, pre-post or expert-novice indices were utilized as the sole criterion for item selection. Based on these data hypothesis three and all of its sub-hypotheses could not be rejected.

Supplementary Findings

This section contains additional findings that were investigated as a result of questions that arose during the primary analysis of the data. Although these inquiries were not previously discussed, they are reported here to provide further insights into the similarities and differences of the indices.

Supplementary Finding One

Hypothesis one was concerned with the amount of linear relationship between the indices. This type of analysis disregards the metric or absolute values of the variables involved and asks the question what happens to one

TABLE 4.9.--Comparisons of the Reliabilities of Tests Where Discrimination Indices Were Used as the Sole Criterion for Item Selection.

Tests Compared	Difference in Obtained Reliability	Differences in z Scores	Obtained z Score	Finding
True-false D, True-false pre-post	.018	.122	.137	NS*
True-false D, True-false expert-novice	.025	.169	.191	NS
True-false pre-post, True-false expert-novice	.007	.047	.054	NS
Multiple choice D, Multiple choice pre-post	.352	.601	1.293	NS
Multiple choice D, Multiple choice expert-novice	.191	.379	.701	NS
Multiple choice pre-post, Multiple choice expert-novice	.161	.222	.592	NS
True-false D, Adjusted multiple choice D	.003	.024	.023	NS
True-false pre-post, Adjusted multiple choice pre-post	.160	.585	1.224	NS
True-false expert-novice, Adjusted multiple choice expert-novice	.053	.244	.405	NS

*NS = Not Significant

variable when there is a change in the other variable? To discover if there were differences in the actual values assigned by each of the indices an analyses of variance was performed. This process involves first determining the means and standard deviations of each index for both true-false and multiple choice items. Table 4.10 provides means and standard deviations for all of the indices. The results of the analyses of variance is reported in Table 4.11. An inspection of the analysis indicated that for true-false items there was no significant differences between the means of any of the indices. Thus, it can be concluded that for true-false items the metric of the indices were similar.

There was a significant difference between the means of the multiple choice items. To discover where the differences existed a Tukey post hoc analysis was performed. This analysis revealed that the values assigned by the expert-novice index was significantly higher than either the D or pre-post values. There was no significant difference between D and pre-post values. The conclusion was reached that with multiple choice items, the expert-novice discrimination index produced values that were higher than the values produced by either D or pre-post discrimination indices.

Supplementary Finding Two

Referring back to Table 4.10, it appeared that for both the true-false and multiple choice items, D obtained

TABLE 4.10.--Means and Standard Deviations of the Discrimination Indices for True-False and Multiple Choice Items.

Item Format	Index of Discrimination		Standard Deviation
True-false	D	.2273	.1626
	Pre-post	.2466	.2313
	Expert-novice	.2577	.2632
Multiple choice	D	.2717	.2039
	Pre-post	.1925	.2165
	Expert-novice	.4290	.2858

TABLE 4.11.--Analysis of Variance Comparing D, Pre-Post and Expert-Novice Indices of Discrimination for True-False and Multiple Choice Items.

Source of Variance	M.S.	Df	F	Results
		True-False		
Between groups	.05680	2	1.1514	Not Significant
Within groups	.04933	717		
		Multiple Choice		
Between groups	.86948	2	15.2540	Significant at .001 level
Within groups	.05700	177		

a standard deviation lower than the other indices. This indicated a lesser degree of variation or spread in the raw scores than was found in pre-post or expert-novice indices. Correlations as were obtained for hypothesis one are attenuated when there is a restriction in range (little variance). It was believed that the obtained coefficients for D might have been effected by this problem. It was decided therefore to re-analyze this hypothesis using Spearman's Rho. This type of correlational analysis utilizes the rank order of each item rather than the actual value of the index. It tests for monotonic rather than linear relationships. Table 4.12 provides the Spearman Rho correlations obtained for the different tests. An inspection of these data indicate that although the coefficients were different than the Pearson product-moment correlational coefficients obtained in testing hypothesis one the results of the significance tests were identical.

Supplementary Finding Three

The conclusion of hypothesis one was that with one exception the correlations between the indices were statistically significant. To investigate further the degree of relationship between the indices each correlation was squared. This provided the degree of variance that was shared between the indices. The greater the r^2 of two variables, the more similar the variables. Table 4.13 provides the r^2 associated with each of the obtained coefficients. Viewing

TABLE 4.12.--Spearman Rho Correlational Coefficients for D, Pre-Post and Expert-Novice Indices of Discrimination with True-False and Multiple Choice Items.

Item Format		Obtained Correlations	Level of Significance
True-false	D, Pre-post	.3858	.001
	D, Expert-novice	.3899	.001
	Pre-post, Expert-novice	.6903	.001
Multiple choice	D, Pre-post	.2812	.015
	D, Expert-novice	.3605	.003
	Pre-post, Expert-novice	.4239	.001

TABLE 4.13.--Squared Correlational Coefficients for D, Pre-Post and Expert-Novice Indices of Discrimination With True-False and Multiple Choice Items.

Item Format	Correlated Indices	r^2
True-false	D, Pre-post	.1683
	D, Expert-novice	.1355
	Pre-post, Expert-novice	.4466
Multiple choice	D, Pre-post	.0778
	D, Expert-novice	.1412
	Pre-post, Expert-novice	.3114

this data it appeared that although the correlations were significant, the amount of shared variance was rather low (never exceeding 45%). Therefore, it was concluded that there was a significant relationship between the indices, however, these indices are far from being identical.

Supplementary Finding Four

To further investigate the relationship of the indices a chi square analysis was performed to determine if any of the indices favored keyed true or keyed false items for item selection. The results of this analysis can be found in Table 4.14. It appeared that both the pre-post and expert-novice indices displayed a preference towards keyed false items. This was not the case with D where the amount of keyed true and keyed false items selected for the final test more adequately represented the proportions existing in the item pool.

Summary

The results of the data analysis for this study were presented in this chapter. The findings concerning the three major research hypotheses and the supplementary analyses were:

1. There was significant relationships between D, pre-post and expert-novice discrimination indices for true-false items. D and pre-post indices were not significantly related for multiple choice items, however, all other

TABLE 4.14.--Chi Square Analysis for Keyed True and Keyed False Items Selected for Inclusion in a Final Test by D, Pre-Post and Expert-Novice Indices of Discrimination.

Indices Compared	Key Items	Observed Frequency	Expected* Frequency	x^2
D,	T	41	51	3.41**
	F	79	69	
Pre-post	T	34	51	10.86***
	F	86	69	
Expert-novice	T	36	51	7.67***
	F	84	69	

*There were 102 keyed true and 139 keyed false items in the pool. Therefore, by chance 51 true and 69 false items would be expected as the final test consists of half the items in the pool.

**Not significant

***Significant at .01

comparisons for multiple choice items provided significant relationships.

2. The three indices were significantly related in terms of item selection for true-false items. With multiple choice items there was no significant relationships found. Although not tested inferentially the percent of overlap in item selection was higher than chance for all indices both for true-false and multiple choice items.

3. There was no significant differences in reliability for tests where items were selected by D, pre-post, or expert-novice discrimination indices. When the reliabilities of the multiple choice items was adjusted utilizing the Spearman-Brown formula there were no significant differences found between the reliabilities of true-false and multiple choice tests, when the same index was used to select items.

4. There was no significant differences in the means of the indices for true-false items. With multiple choice items the mean value for the expert-novice index was greater the means of either D or the pre-post indices.

5. Applying Spearman Rho correlational analysis to the data resulted in findings very similar to those obtained using Pearson's product-moment correlational analysis. This was constant for both true-false and multiple choice items.

6. Squaring the obtained correlational coefficients to determine the amount of shared variance revealed that in

no case was there more than 45% shared variance. The indices were significantly related but were not identical.

7. Dividing the true-false items into two separate pools based on the keyed correct answer it was found that pre-post and expert-novice indices significantly favor false items in their item selection process. D more adequately represented the proportions found in the item pool.

CHAPTER V

SUMMARY AND CONCLUSIONS

Summary

The purpose of this study was to compare D, pre-post and expert-novice indices of discrimination with both true-false and multiple choice items. The three major questions that were formulated as research hypotheses were:

1. Is there a significant relationship between the three indices of discrimination for true-false and multiple choice items?

2. Using a large item pool and D, pre-post or expert-novice discrimination indices as the sole criterion for item selection, is there a significant number of common items in the final test forms for true-false and multiple choice items?

3. Are multiple choice and true-false achievement tests that were constructed using D, pre-post or expert-novice indices of discrimination as the sole criterion for item selection equally reliable?

To obtain a more complete understanding of the similarities and differences among these indices four supplementary analyses were performed. The supplementary analyses were designed to answer the following research questions:

1. Are there significance differences in the mean discrimination index values for D, pre-post and expert-novice indices with true-false or multiple choice items?

2. Using Spearman Rho correlations, are there significant relationships among the three indices of discrimination which are different from the relationships obtained using Pearson's product moment correlations (thus investigating both monotonic and linear relationships)?

3. How similar are the indices as measured by the amount of shared variance obtained by squaring the correlational coefficients?

4. With true-false items, do any of the discrimination indices display a significant preference for selecting keyed true or keyed false items in developing achievement tests?

A review of the literature revealed that there were a few studies comparing D or various other conventional indices of discrimination with a pre-post discrimination index. The findings of these studies indicate that the indices tended to rank order the items differently and hence would select different items from an item pool for inclusion within a final achievement test. However, these studies were limited in that they utilized rather small item pools and small subject pools. There were no studies available concerning empirical comparisons of the expert-novice discrimination index with any other discrimination index.

To obtain both a large item pool and a large subject pool, six samples were administered portions of the true-false and multiple choice items. The first sample consisted of 95 students who were administered 126 true-false items in a pre test setting. The second sample included 120 students who were administered these same true-false items as a post test. Samples three and four contained 81 students being administered the remaining 114 true-false items as a pre test and 185 students attempting the items as a post test. The fifth sample was comprised of 77 students who were administered the entire 60 item multiple choice pool as both a pre and post test. The last sample consisted of 20 experts in the field of measurement and evaluation who were divided into two groups each taking 120 true-false and 30 multiple choice items. Based upon these samples, D, pre-post and expert-novice discrimination index values were computed for each item in the true-false and multiple choice item pools.

Pearson product-moment and Spearman Rho correlational analyses were performed on the entire item pools. These analyses were designed to investigate the relationships among the three discrimination indices. The analyses were executed separately for each of the two item formats.

As a measure of the similarity of the indices as selection criterion, the best 120 true-false and 30 multiple choice items were determined and compared for each index. The comparisons were handled through phi-coefficients. As an added descriptive analysis of these similarities, the

percent of overlap was calculated and described for each comparison.

Estimations of the Kuder-Richardson Formula 20 reliability coefficients were calculated for each of the six final true-false and multiple choice tests. As all of the necessary components of the Kuder-Richardson Formula 20 reliability formula were not available with this data, an estimation of the Kuder-Richardson Formula 20 based upon the work of Ebel (1972) was applied to develop these coefficients. Statistical tests were performed to determine if the reliabilities obtained using the indices as the sole item selection criterion were significantly different.

To investigate whether any of the indices systematically provided higher or lower discrimination values, an analysis of variance was performed comparing the means of the indices. This analysis included all of the items in each of the pools. Where significant differences were reported, a Tukey post hoc analysis was performed.

Each of the obtained Pearson product-moment correlational coefficients were squared to provide for additional insights as to the meaningfulness of the existing relationships among the discrimination indices. This procedure allowed the investigator to determine the amount of shared variance of the variables compared.

The items in each of the final true-false achievement tests were dichotomized according to their keyed correct response. A chi square analysis was then performed to

test whether any of the indices tended to select a significantly different number of either keyed true or keyed false items than would be expected by chance. This analysis was designed to investigate whether certain indices preferred a particular type of keyed response item.

Conclusions

The conclusions associated with the three major research hypotheses and four supplemental analyses were:

1. With the exception of D and pre-post discrimination indices with multiple choice items, significant relationships were found among the three discrimination indices across both true-false and multiple choice items.

2. There was a significant amount of overlap in the items selected for inclusion in a final test by all of the indices tested with true-false items. There were no significant amounts of overlap obtained with multiple choice items.

3. There was no significant differences in reliability for tests where items were selected by D, pre-post or expert-novice discrimination indices. When the reliabilities of the multiple choice items were adjusted utilizing the Spearman-Brown formula there were no significant differences found between the reliabilities of true-false and multiple choice tests, when the same index was used to select items.

4. There was no significant differences in the means of the indices for true-false items. With multiple choice items the mean value for the expert-novice index was greater than the means of either D or the pre-post indices.

5. Applying Spearman Rho correlational analysis to the data resulted in findings very similar to those obtained using Pearson's product-moment correlational analysis. This was constant for both true-false and multiple choice items.

6. Squaring the obtained correlational coefficients to determine the amount of shared variance revealed that in no case was there more than 45% shared variance. The indices were significantly related but were not identical.

7. Dividing the true-false items into two separate pools based on the keyed correct answer it was found that pre-post and expert-novice indices significantly favor false items in their item selection process. D more adequately represented the porportions found in the item pool.

Discussion

The findings of this study are somewhat in agreement with the conclusions drawn by other researchers who compared D or another conventional index of discrimination with the pre-post method. None of the studies previously undertaken, however, incorporated the expert-novice discrimination index.

The correlational coefficients obtained in this study, demonstrate that a relationship does exist between the indices. This remains constant regardless of whether Pearson product-moment or Spearman Rho correlational coefficients are viewed. Based on the alpha level established for this study, all of obtained coefficients were statistically significant with the exception of D and pre-post indices for multiple choice items. The obvious explanation for this exception may be that with multiple choice items, D and pre-post indices are totally independent. Theoretically, these indices are defined differently. However, a significant relationship was found with the true-false items for these indices. Also, comparing the correlational coefficient for D and pre-post true-false items with the correlation coefficient for D and pre-post multiple choice items revealed no significant difference. As the true-false item pool was so much larger than the multiple choice item pool, it is possible that with a larger item pool these relationships might result in a significant correlations.

The results of the statistical testing of this hypothesis is probably not of paramount importance. The same kind of conclusion could be reached by an inspection of the obtained coefficients. In all cases, the coefficients are low to moderate. The squared correlations associated with these coefficients reveal that although significant relations exist there is little common variance between the indices.

From a meaningfulness sense it can be concluded that although the indices are related, the relationships are low enough as to be considered as not very strong. Referring then to the fourth research question posed by this study, it appears that the indices are sufficiently different to cause measurement specialists to be concerned regarding which index is employed with a set of items.

A comparison of the means of each index revealed that with true-false items the values assigned by the indices were not significantly different. However, with multiple choice items the mean expert-novice value was significantly greater than the mean pre-post value. The explanation for this may lie more with the item format than with differences in the indices. The previous discussion on the correlations among the indices revealed only one non-significant relationship. This non-significant relationship was found with the multiple choice items. The testing of the means also provided only one instance of inconsistency. This inconsistency existed within multiple choice items. This does not imply that the indices are identical with true-false items but rather the findings seem to be more consistent with true-false than multiple choice item formats.

An alternative explanation of this occurrence could be the number of items and subjects that were involved in this study. The true-false item pool was four times longer

and utilized more subjects than the multiple choice item pool. Consistency or reliability tends to increase when more items are included or when a more hetergeneous group of subjects is administered the items. It is possible therefore that the findings would be more consistent with the multiple choice items if a larger pool had been included or more subjects administered the items. This alternative explanation appears more valid when looking at the estimates of the Kuder Richardson Formula 20 reliability coefficients. The obtained coefficients for the multiple choice items were considerably lower than the coefficients that were adjusted by the Spearman Brown formula to equate the number of multiple choice items to the number of true-false items. This same kind of change might occur to the means and correlations of the multiple choice items if the number of items in the pool were increased.

The phi-coefficients calculated to test the amount of overlap in item selection where the indices of discrimination were employed as the sole selection criterion provided results that were consistent within each item format but inconsistent across item formats. In all cases there were a significant number of items jointly accepted or rejected by the indices for true-false items. The opposite was found for the multiple choice items. The obvious explanation for these findings is that for item selection the indices tend to function in a similar manner for true-false

items but function differently for multiple choice items. If this is correct then measurement specialists need be less concerned by which index is used for item selection with true-false tests but much more concerned when the test consists of multiple choice items. Knowing that a phi-coefficient is significant is similar to knowing that the Pearson product-moment correlational coefficient is significant. There is a statistical relationship but concern must be given to investigate if there is a meaningful difference. To address this concern a percent of overlap was computed for each phi-coefficient. This was provided for both the true-false and multiple choice items. For true-false items it revealed that at least sixty-five percent of the items would be identical regardless of the discrimination index used for selection. The indices appear to work very similarly for selecting items from a pool to constitute a final test form. The percent of multiple choice items that would be common regardless of the index employed was found to be at least fifty percent. Although this percentage is relatively high, it is not statistically different from what would be expected by just chance selection. However, it should be remembered that inability to reject the null hypothesis does not indicate that there is no relationship or similarity in the indices as item selection criterion. Rather, this inability to reject the null hypothesis indicates that if there was no relationship between the

indices, results such as were obtained in this study could be expected through chance factors more often than was specified by the alpha level determined prior to analysis.

Another explanation of why the phi-coefficients were statistically significant for the true-false achievement tests but not for the multiple choice achievement tests refers to a concern discussed earlier in this report. This concern is the number of items in the multiple choice item pool. Had there been a larger item pool it is possible that the obtained phi-coefficients might have produced significance. To address this alternative explanation of the findings, future research with multiple choice items might benefit from inclusion of a larger pool of multiple choice items.

The findings of this study suggest that regardless of the type of discrimination index utilized for item selection, the reliability of the final achievement tests will not be significantly different. The explanation for this finding may rest with the quality of items which constituted the item pools in this study. All of the items used in this study were constructed, piloted and revised by experts in the area of test construction. The reliability estimates obtained from any sample of these items would probably provide for a fairly reliable test. Also, considering the amount of item overlap in these final tests, they could almost be expected to provide for close estimations of total test reliability. This finding pertained

both to true-false or multiple choice achievement tests.

Supplementary finding four which dealt with the indices' preferences toward keyed true and keyed false items did reveal differences among the indices. Selection of items based on D values produced a test consisting of keyed true and keyed false items in proportions not significantly different than the proportions found in the item pool. Pre-post and expert-novice indices, however, displayed a significant preference for items that were keyed false. Thus, the use of either pre-post or expert-novice indices will result in achievement tests which contain a significantly larger ratio of false to true items than is represented in the item pool from which the items were selected.

The results of this study, when viewed in toto, suggest that the indices are in most cases related as indicated through Pearson product-moment and Spearman Rho correlational coefficients. Yet the indices are not identical as evidenced by investigation of the amount of shared variance. Also, in most cases the values assigned by the indices are not significantly different. As item selection criterion, the indices tend to select a significant amount of common true-false items, however, pre-post and expert-novice indices display a preference for keyed false items while D does not indicate any bias for particular keyed correct answers. Reliability does not seem to be significantly affected by any of the indices. The indices function similarly but they are not identical.

These findings tend to disagree with the theories posed by Carver (1974) and Green, Nyquist and Griffore (1975) who advocate that the indices are different and if used as item selection criterion will produce totally different achievement tests. However, the findings are very similar to the investigative efforts of Helmstadter (1972) and Crehan (1974) who have shown that there is a considerable amount of overlap between D and pre-post indices in the items they would select for inclusion in a final test and that the resultant test reliabilities are similar.

This study was not primarily concerned with making value judgments regarding any of the indices. However, considering the results of this study, some concern has arisen relating to the utility of employing the expert-novice discrimination index. This index more closely follows the theoretical definition of an index of discrimination. However, obtaining the cooperation of a group of experts is a rather difficult task. The amount of overlap between this index and D or pre-post was high, and the reliabilities of the final tests were not greatly different. Thus, from a practical perspective it appeared that the efforts put forth to obtain a sample of experts does not provide for tests that either contain greatly differing items or more reliable tests than if D or pre-post indices are utilized. For classroom teachers who have to develop a great number of tests for their students, D or pre-post indices are recommended.

The expert-novice index is useful in other situations such as in establishing objectivity of the test items and as a check for intrinsic ambiguity yet for item selection D or pre-post will probably be just as valuable.

Concerning the issue of true-false and multiple choice item formats, the findings of this study indicate that these formats perform differently relative to the indices of discrimination. However, as there was a large discrepancy between the numbers of true-false and multiple choice items it is possible that the differences may be more an indication of sampling differences than format differences. In the one situation (test reliability) where the number of items was equated, no significant differences were reported. Further study is necessary to clearly compare the item formats relative to the indices of discrimination.

This study concentrated on comparison of the indices as they relate to each other in the assignment of item quality values and item selection criterion. This study has not been designed to advocate the use of discrimination indices as the sole criterion for item selection. Discrimination indices are also useful for item revision. Item selection should be based on a number of other criterion including balance and relevance.

Limitations of the Study

The novices in this study were defined as those who had not taken the course previously. There was no way, however, to control for the earlier experiences of these subjects. It is possible that some of the items measured concepts that the subjects had been exposed to either in other courses or through professional experiences. To the extent that this occurred, the validity of the findings would be reduced. There was no method available to screen out the effects of past experiences and exposures.

The subjects who attempted the true-false items on the pre and post test sessions were different. The assumption was made that these subjects were basically equivalent over time. To the extent that this assumption is violated, the results obtained with this data would be questionable.

The items used in this study were designed, piloted and revised for inclusion in norm-referenced tests. Generalization to criterion-referenced tests are therefore contingent on the assumption that criterion-referenced items and norm-referenced items are basically identical in design and different only in interpretation of the total test scores. A review of the literature provided no evidence which would disprove this basic assumption.

All of these items were written by experts in the field of measurement and evaluation. Generalizations to items which are not so expertly constructed are left to the reader's discretion since the population of items used did

not reflect a great deal of variance in item construction quality.

The true-false and multiple choice items were limited to a rather specific subject area. It is questionable whether the findings should be generalized to vastly different subject areas.

As there was a discrepancy in the size of item pools utilized in this study, findings regarding the differences between item formats could be alternatively explained as resulting from sampling fluctuation. Large item pools such as was the case with the true-false items would be preferred to compare the item formats.

Limitations in the availability of the data restricted the types of analyses that could be performed. One such restriction resulted in testing all correlational coefficients different from zero. This analysis allows the researcher to state if the variables are related. A preferable analysis would consist of testing the coefficients different than unity. In that situation the researcher could investigate whether the variables are identical. Lord (1957) has developed such a procedure using a likelihood-ratio significance test. McNemar (1958) has provided a similar procedure using analysis of variance. Forsyth and Feldt (1969) (1970) have demonstrated a method of improving upon McNemar technique and have also generated a third measure based on the establishment of confidence intervals. These procedures are all based on the premise that two variables may be

perfectly correlated, however, because of unreliability in the measurement of each of the variables the obtained correlation is attenuated. The first step in any of these procedures is to correct the obtained correlational coefficient for unreliability in the variables. To do this the researcher must first obtain an estimate of the reliability of each variable. A review of the literature indicated that a coefficient of stability is at present the only method of obtaining the reliability of indices of discrimination. As the original test sheets were not available for the subjects in this study these analyses could not be performed. Thus, the question of whether the indices were identical could not be directly answered within the scope of this study.

Suggestions for Future Research

The following suggestions are offered for further investigation into comparisons of D, pre-post and expert-novice indices of discrimination with different item formats:

1. The results of this study were based on expertly constructed items. For practical purposes it would be appropriate for future research investigations to utilize item pools that consist of items which are developed by teachers or other educators who have differing levels of expertise in item writing skills. This would allow for generalizations to be more applicable to the vast number of test constructors who possess varying degress of competency in item development.

This might also provide additional insights as to how reliable tests are when the indices are used as the sole criterion for item selection.

2. This study revealed that in most cases the indices are related. There was no available method for determining the reliability of each index. It would be beneficial for future research investigations to obtain the test sheets of all subjects so that this measure can be calculated. This then would allow for a more thorough analysis regardless of whether the researcher is interested in testing whether the coefficients are different than zero or different than unity.

3. The results of this study suggested that there were differences between true-false and multiple choice item formats. However, because there was a large discrepancy in the number of items in the respective pools, these results were suspect. Future research might include a large and equal number of items in each of the item pools.

4. This study estimated that pre-post and expert-novice indices displayed a preference for keyed falsed responses. Future research might be designed to investigate if the number of alternatives or the type of alternatives (all the above or none of the above) used with multiple choice items are reacted to differently by the indices.

5. As many criterion-referenced tests are used with lower grade students, research is necessary to discover the

similarities and differences of the indices with younger students. This is especially important as younger students may not be as test taking oriented or as test wise as their older counterparts in college.

6. To more adequately represent pre-post index, future research should involve subjects who will be administered both versions of the achievement test. This would more closely address Carver's concern about growth within an individual over time.

7. Research is needed to investigate the stability of the pre-post index over time. To address this concern various time spans could be incorporated with the pre-post index. This would allow the researcher to see if items obtain similar values if the time between the pre test and post test are altered.

TRUE-FALSE ITEM POOL

APPENDIX A

FORM B

T 1. In any set of scores only two scores, the highest and the lowest, determine the range.

T 2. Each of the class intervals in a frequency distribution includes the same number of raw score units.

F 3. The class intervals in a frequency distribution should be broader (i.e., include more raw score units) at the extremes of the distribution, where scores are fewer, than in the center of the distribution, where scores are concentrated.

F 4. The median of the scores 87, 16, 72, 89, 3, 96 and 78 is 63.

F 5. In order to determine the raw score deviation of a single test score from the mean of all test scores one must first calculate the standard deviation.

T 6. In a perfectly normal distribution of scores, the mean score has the same value as the median score.

T 7. The variance, in grouping units, of a set of scores is 30.05. If each of the grouping units includes five raw score units, the variance in raw score units is 751.25.

T 8. In a set of 100 test scores the ratio of the range of scores to their standard deviation is likely to be larger than it would be in a set of 20 test scores.

T 9. In a set of five test scores; 2, 4, 6, 8, and 10; the percentile rank of the score 8 is 70.

T 10. If the same 100 item test is given to two classes, one composed of good students and the other of poor students, and if percentile ranks are figured separately for scores from the two classes, a student in the good class who answers 75 items correctly is likely to get a lower percentile ranking for his score than a student from the poor class who also answers 75 items correctly.

T 11. When raw scores are converted to percentile rank scores, the shape of the distribution of scores is changed.

F 12. The formula $1 - \frac{6\Sigma d^2}{N^2 (N^2-1)}$ is used to estimate the variance of a set of test scores.

F 13. For a given student who makes high scores on three tests in a given class, the average of the three percentile rank of the average of his three raw scores.

F 14. The distribution of stanine scores is approximately rectangular.

F 15. In order to make two different tests carry equal weight in a composite, one needs information on the number of items in each test.

T 16. A stanine score of 7 and a stanine score of 9 are 1 S.D. apart.

T 17. One estimate of correlation can be obtained from difference in the total scores of high and low scoring groups of students.

F 18. In calculating a correlation coefficient, the quantity ΣX^2 is found by squaring the sum of all the X-s.

T 19. If the scores 12, 10, 10, 9, 9, 9, 8, 7, 7, 4 are converted to ranks (for purposes of calculating a coefficient of correlation) the score 9 should be assigned a rank of 5.

T 20. The probable error of sampling is smaller for a coefficient of correlation of .90 than for one of .10 if both are based on samples of the same size.

F 21. It is possible to correlate the scores of different students on the same test when only a single score is available for each student.

T 22. A step in one process for estimating a coefficient of correlation is to subtract the rank of one member of each pair of scores from the rank of the other member of that pair.

T 23. In a table showing the relation between raw scores and percentile ranks on a test the following pairs of numbers appear:

Upper row	90	70	50	30	10
Lower row	80	60	50	40	20

The raw scores are probably shown in the upper row, and the percentile ranks in the lower row.

T 24. If a scatter diagram showing how scores on two tests are related has one and only one tally mark in each cell of the diagram, then the correlation between scores on the two tests must be zero.

T 25. If the scores on two tests, X and Y, are each determined by two equally potent factors, so that scores on Test X depend on factors A and B, while scores on Test Y depend on factors A and C, then one should expect a correlation of .50 between the scores on Test X and on Test Y for the same group of students.

Items 26 - 29 are based on these scores.

Student	Scores X	Y
A	10	4
B	5	8
C	3	1
D	2	3

T 26. The variance of scores on Form X is greater than 9.

T 27. The standard deviation of scores on Form Y is less than 3.

T 28. The XY equals 89.

T 29. The quantity $N(N^2-1)$ (used in calculating the rank difference coefficient of correlation) is 60.

F 30. One essential step in the best method of calculating the variance of a set of 100 test scores is to find the differences between each of the scores and the mean score.

T 31. The correlation between scores on Aptitude Test A and subsequent achievement is -.60. The correlation between scores on Aptitude Test B and subsequent achievement is .20. More accurate predictions of achievement can be made on the basis of scores from Test A than on the basis of scores from Test B.

F 32. A test is said to possess balance if the items are moderate in difficulty, and highly discriminating.

F 33. A test meets the criterion of objectivity if test-wise novices get scores at or near the chance level.

F 34. Different standards from those used with objective tests must be used in judging the quality of essay tests.

F 35. The best evidence of item relevance is statistical.

T 36. To build a test showing adequate balance one should classify items on the basis of their verbal form instead of on the basis of the mental abilities they require.

T 37. A question which asks how a mental age is determined is a factual information question.

F 38. A good test of educational achievement will include more items on facts than on generalizations.

T 39. Factual true-false items tend to be more efficient than multiple-choice items involving applications.

T 40. When most of the students in the upper 25% of a class answer a test question incorrectly the teacher should consider rewriting the item.

F 41. A passing score of 70% on a multiple choice test will generally result in approximately the right proportion of passing scores.

T 42. A speed test is more wasteful of examination questions than is a power test.

T 43. A test composed of highly discriminating items (i.e., .40 and higher) will yield scores that differ greatly from student to student.

T 44. Indices of discrimination should be based on the responses of all students tested.

T 45. The instructor is better qualified than are the students to judge the fairness of a test.

T 46. The more students guess on an objective test, the lower its reliability will be.

T 47. Objective test scores corrected for guessing by adding a fraction of the number of questions omitted correlate perfectly with scores on the same test corrected for guessing by subtracting a fraction of the number of questions answered incorrectly.

T 48. The best true-false test items are those that more than 70% and less than 90% of the students answer correctly.

F 49. Kuder-Richardson reliability coefficients provide a composite estimate of reader, examinee and test reliability.

F 50. When reliability is estimated by correlating scores from equivalent forms of a test, the Spearman-Brown correction must be applied.

F 51. If the correlation between two 50 item tests is .50, the correlation between two similar 100 item tests should be .75.

F 52. One step in determining the value of KR-21 is to determine, for each item, the proportion who passed it and failed it.

T 53. If all items in a test are equally difficult the reliability coefficient obtained by using Kuder-Richardson Formulas 20 and 21 will be identical.

F 54. A biology test item that is difficult for students who have studied biology is low in specificity.

F 55. Kuder-Richardson Formula 20 can be used to estimate the reliability of a test regardless of whether or not the test scores have been corrected for guessing.

F 56. The correlation between scores on equivalent halves of a test is equally likely to be either an overestimate or an underestimate of the true test reliability.

F 57. Test reliability coefficients of .50 to .70 are usually regarded as quite satisfactory by expert test constructors.

T 58. Error-free test scores would be perfectly reliable.

F 59. A test on which the scores range from 75 to 100 is likely to be more reliable than one on which the scores range from 25 to 100.

T 60. The theoretical true scores on a test are less variable than the actual obtained scores.

T 61. If the reliability coefficient is .70, then 70% of the score variance is true score variance.

T 62. It is easier to obtain high reliability in scores from a test if the items that compose it are homogeneous (like each other) than if they are heterogeneous (different from each other).

F 63. If the reliability of a 50-item test is .60, and we desire to estimate the reliability of a similar test, twice as long, by the Spearman-Brown formula, then the denominator in the formula will be 1.20.

T 64. If a 100 item test yields scores having a standard deviation of 10 in group A, and a standard deviation of 5 in group B, the scores are likely to be more reliable for group A than for group B.

T 65. The correlation between scores based on the number of items answered correctly and scores based on the number of minutes required to complete the test can be expected to be positive but low.

T 66. To estimate the standard error of measurement from the reliability coefficient one much also know the standard deviation of the test scores.

T 67. If extreme groups of 33% instead of 27% are used for item analysis, the groups will be more alike in average ability.

F 68. If six of ten students who score high (upper 27%) on a test answer a particularl item correctly, while four of ten who score low (lower 27%) on the test answer the same item correctly, the index of discrimination D is .67.

F 69. In ordinary item analysis a distinction is made within the upper 27% group between the 9% receiving highest scores and the others.

F 70. The primary goal of item analysis is to improve test validity.

F 71. Indices of item discrimination derived from the responses of fewer than 30 students are subject to such large sampling errors that they are practically worthless.

T 73. If Tests A and B are composed of equal numbers of items, but those in Test A are more highly discriminating than those in Test B, then scores on Test A will be more variable than those on Test B.

T 74. The mean index of item discrimination must be zero in a test of zero reliability.

T 75. In general, the more variable the distribution of item difficulties in a test, the less variable the distribution of student scores on the test.

T 76. Multiple choice test items can sometimes be made more discriminating by making some of the distracters more obviously incorrect.

T 77. Item analysis is more useful to a teacher who reuses items than to one who does not.

F 78. Good classroom test items should have indices of discrimination of .50 or more.

F 79. A test can be valid even though it does not yield valid scores.

F 80. If scores from a classroom test correlate highly with grades in the same course, the test is a valid test.

F 81. If a test appears to measure what it claims to measure, it is said to possess construct validity.

F 82. To determine predictive validity properly, criterion measures must be obtained for each student before the test is given.

F 83. Educational measurement is handicapped more by lack of suitable techniques of measurement than by uncertainty about what to measure.

T 84. It is possible to construct a relevant test which is not valid.

F 85. If handled properly, the problem of marking student achievement can be made relatively simple.

T 86. Percent marks are more dependent upon the teacher's standards than are relative marks.

F 87. Empirical studies of marking standards of experienced teachers show more uniformity than was previously supposed.

F 88. The reliability of the composite semester marks issued by typical classroom teachers rarely exceeds .60.

F 89. Absolute marking is more likely than relative marking to stimulate student efforts to achieve.

F 90. Many teachers who like absolute grading find a percentile scale useful for this purpose.

F 91. In one form of relative marking, the teacher sets standards relative to what he thinks the class should accomplish.

F 92. The studies of Starch and Elliott were mainly concerned with discovery of the factors considered by Teachers in assigning marks.

F 93. One serious drawback of relative marking is that it permits and encourages the students in a class to slow down so that no student has to work too hard to earn a passing mark.

F 94. Flexibility is more essential than uniformity in an institution's marking system.

F 95. The unit on the stanine scale of marks is one ninth of the standard deviation of the scores used as the basis for marking.

T 96. If the test has been constructed properly, accurate measurements of educational achievement are more likely to be obtained from students who are skilled in test taking than from those who are not.

T 97. School marks ought to be regarded as impersonal measurements, rather than as personal evaluations.

T 98. The reliability of a status measure tends to be greater than that of a growth measure.

F 99. The first step toward improved marking in most schools and colleges should be to replace a single mark reflecting a student's over-all achievement in a course with a series of several marks to indicate his achievement of several different objectives of the course.

F 100. If the information on which marks are based in quite unreliable, the use of few categories in marking will report that information more accurately than the use of many categories.

T 101. Numbers have significant advantages over letters as symbols for reporting marks.

F 102. If the difference between the sums of scores in the upper and lower one-sixths of the distribution is divided by one half the number of scores in the distribution one obtains an estimate of the reliability of the scores.

T 103. The median is a good point of reference for establishing the score intervals that correspond to each mark to be issued.

T 104. In any distribution of scores, the sum of the deviations from the mean always equals zero.

T 105. The expected chance score on a 90 item test composed of three-alternative multiple choice items is 30.

F 106. How much weight one particularl component of a final mark carries in determining the final mark depends on the mean value of scores for that component.

F 107. When marks are based on total scores obtained by adding several component scores, each component will carry equal weight in determining the mark unless special differential weighting is used.

F 108. The scoring formula is designed to transform a distribution of raw scores into standard scores having a normal distribution and a specified mean.

F 109. The Spearman-Brown formula is used to obtain the correlation between the odd numbered items and the even numbered items in a test.

F 110. A split halves reliability coefficient is obtained by subtracting the mean score on the first half of the test from the mean score on the second half of the test, and dividing this difference by the mean score on the total test.

F 111. In the method of grade assignment described in the text, the median score determines the mid-point of the C interval.

T 112. In the method of stanine assignment described in the text, the lower limit of the interval for stanine scores of nine is determined by three numerical quantities.

T 113. Teachers should motivate students to make the best scores they possibly can on standardized aptitude tests.

F 114. Although some newer IQ tests have experimented with deviation scores, most test experts have not accepted them as superior to ratio IQ scores.

FORM A

F 115. Externally produced standardized or program tests have more influence on what students study and learn than do teacher-made classroom tests.

T 116. The more students know in advance about the kinds of knowledge and ability a test will require, the more wholesome will be its influence on learning.

F 117. The principal criterion of quality in an achievement examination is its direct contribution to the student's learning.

F 118. Most published collections of test items have been designed to be useful to elementary school teachers.

F 119. It is possible for teachers to improve greatly the quality of their classroom tests while at the same time reducing greatly the time required to prepare and score them.

F 120. With proper guidance from a test specialist, even a mediocre teacher should be able to develop an excellent achievement test.

F 121. Teachers tend to rely too much on relative judgments of student achievement, paying too little attention to absolute standards of achievement.

T 122. If a student gets a score of 20 on the odd numbered words in a 50 word spelling test, and a score of 15 on the even numbered words, the difference between 20 and 15 illustrates the effect of sampling errors.

T 123. The purpose of giving achievement tests under specially devised and carefully controlled conditions is more to improve test reliability than to improve test validity.

F 124. Classroom tests should be used primarily to discover particular errors and omissions in student learning.

F 125. The more a test constructor succeeds in asking the questions that he ought to ask, the more reliable his test will be.

F 126. The more natural the situation in which teachers attempt to measure student achievement, the more precise those measurements are likely to be.

T 127. It is logically contradictory to assert that an important outcome of education is intangible.

T 128. Any intelligence test is one operational definitive of intelligence.

F 129. Herbert Spencer believed that a study of the great books ought to occupy most of a student's time in secondary school and college.

F 130. The most authoritative and useful statements of educational goals are those that have been derived by logical deduction from a single basic statement of the meaning and purpose of life.

T 131. The Taxonomy of Educational Objectives includes items taken from actual tests which are intended to illustrate how attainment of the objective can be measured.

F 132. Items which describe specific situations and call for the examinee to choose the most appropriate behavior tend to be clear, definite and efficient.

T 133. Everything that a person has experienced becomes a part of his knowledge.

F 134. Understanding an idea involves much more than a knowledge of relationships.

F 135. It is usually more difficult to learn something (i.e. get it into the mind) then to recall it when needed (i.e. get it out again).

F 136. The problem of measuring noncognitive educational achievements has been solved as satisfactorily as the problem of measuring cognitive educational achievements.

F 137. The meaning of a quantitative concept can not be defined arbitrarily, it must be discovered by investigation.

F 138. If left to their own devices in study, most students tend to rely heavily on rote learning.

F 139. A student can achieve command of knowledge to a high degree without acquiring the ability to use it.

T 140. Test scores corrected for guessing tend to correlate highly with uncorrected scores on the same test.

F 141. Experts agree that cheating can be eliminated by the use of open-book examinations.

F 142. Most achievement tests should be planned to include items testing a variety of different mental processes.

T 143. A test composed entirely of items of moderate difficulty (neither very easy nor very hard) can nevertheless discriminate well among the very best students, and among the very poorest students.

F 144. Substantial improvement in validity can usually be obtained (though at a cost of greater difficulty in scoring) by correcting the scores on objective achievement tests for guessing.

F 145. One of the most difficult, but most essential, steps in the process of testing educational achievement is to determine a reasonable passing score on the test.

F 146. A more reliable measure of achievement can be obtained from a single long end-of-course test than from three shorter unit tests given during the course. (Assume that the time spent and the number of items used are the same for the single long test as for the total of the three short tests.)

F 147. One should choose among essay, true-false, ultiple-choice and other item forms depending on the particular mental ability that is to be tested.

T 148. One unique advantage of the essay test is that it is easier to prepare than most objective test forms.

T 149. It is possible and desirable to improve the objectivity of scores from an essay test.

F 150. An achievement test should include enough test items to keep nearly every student busy during the entire test period.

T 151. The larger the number of items in a test the smaller the chance that a poorer student will get a higher score than a better student.

T 152. Essay tests ought to be analyzed and evaluated systematically against objective standards of test quality.

T 153. Test specialists recommend avoidance of optional questions on essary examinations.

F 154. The quality of an essay test, like that of an objective test, depends almost entirely on how skillfully the questions are written.

F 155. A social studies teacher grades his students not on the conclusions they reach, but only on the soundness of their reasoning in defense of those conclusions, whatever they are. The teacher's policy is a good one for those who use essay tests to follow.

F 156. The first step in the process of scoring essay test answers analytically is to sort them into piles on the basis of the over-all quality of the answer given.

F 157. Although some have attempted to improve essay test reliability by averaging scores of independent readers, in actual practice this has proved to be of doubtful value.

F 158. Though it is less obvious, guessing on an essay test is just as much of a problem as guessing on an objective test.

T 159. If a test is expected to yield raw scores that conform to some predetermined distribution, then an essay test is likely to be easier to use than an objective test.

T 160. Most of the abilities that can be measured by using essay tests can also be measured by using objective tests.

F 161. Objective test items should not be used to test a student's ability to solve problems.

F 162. If an objective test is constructed properly, few if any subjective judgments will be involved in the process.

F 163. Scores from essay tests can be checked for accuracy of scoring as easily and definitely as scores from objective tests.

T 164. If a small class (fewer than 20 students) is to be tested with a new test, and if the teacher is short on time for getting the job done, he should probably use an essay test.

F 165. Essay questions are poorly suited to testing a student's command of essential knowledge.

F 166. One of the useful characteristics of an essay test is that equally good answers can be almost totally different.

T 167. In general, an essay test composed of 20 questions, each of which can be answered in a single paragraph, will yield more reliable scores than an essay test composed of only 5 questions, each of which requires an answer of several paragraphs.

T 168. True-false test items were more popular in 1930 than they were in 1960.

F 169. Triviality and ambiguity are inherent weaknesses of true-false test items.

T 170. The wrong learning which results from presenting false statements to students in a true-false test is negligible.

T 171. All important aspects of verbal knowledge can be dealt with effectively in true-false test items.

T 172. It is more convenient to present some test problems in multiple choice than in true-false form.

F 173. Well constructed true-false items are no more subject to guessing than are well constructed four-choice multiple choice items.

T 174. True-false items are simpler to write than multiple choice items.

F 175. If half the students find a true-false item ambiguous, it certainly needs revision.

T 176. It is good practice, in writing false statements for a true-false test, to include words like "often," "usually," or "sometimes."

T 177. The difference in achievement of good and poor students shows up more clearly on the false than on the true statements in a true-false test.

T 178. The idea behind every multiple choice test item should be expressible as a single, independently meaningful proposition.

F 179. The popularity of multiple choice test items has declined sharply in recent years.

T 180. Good writers of multiple choice test items sometimes lead students who lack understanding to choose a wrong answer by including a familiar textbook phrase in it.

F 181. A multiple choice item may be a good and useful item even though more students of low achievement than of high achievement succeed in consistantly answering it correctly.

T 182. A test constructor should suspect the validity of a physics or geography test item that can be answered correctly by many students who have made no special study of physics or geography.

F 183. The most efficient way to wrote multiple choice test items is to type each item on a 3 x 5 card in as nearly its final form as possible.

F 184. Incomplete sentences usually make better stems for multiple choice test items than do direct questions.

F 185. Multiple choice items whose stems are stated negatively (with the word not playing a crucial role) tend to be more discriminating than those whose stems are stated positively.

F 186. The response "none of the above" makes a good fourth or fifth response to almost any multiple choice test item.

F 187. The more diverse the responses to a multiple choice test item are in structure, content and appearance, the more discriminating the items is likely to be.

T 188. Good distracters in multiple choice test items possess two essential qualities--plausibility and incorrectness.

F 189. An item writer should make the incorrectness of distracters harder to discern than the correctness of the answer, so as to prevent a student from reaching the correct answer by a process of elimination.

F 190. A writer who is particular and exacting in his choice of words is likely to write ambiguous multiple choice test items.

F 191. Any multiple choice test item that less than 50% of the examinees answer correctly is a poor item.

T 192. General questions tend to be easier to answer than specific questions, if the answers offered for each type are equally good.

F 193. If an item writer works carefully, he has little need for an independent (i.e. by someone else) review of the items he has written.

F 194. Multiple choice test items that call for only a "best" answer, instead of a perfectly correct answer, tend to be less discriminating and more ambiguous.

F 195. Indirect measures of achievement, based on the examinees' ability to recognize details from the materials used in instruction, or from the process of instruction, are often highly effective and desirable.

F 196. Distracters in multiple choice test items should

F 196. Distracters in multiple choice test items should never themselves be true statements.

F 197. If no entirely correct and perfectly adequate answer can be given to the stem question of a multiple choice test item, the question should be discarded.

F 198. In order to discriminate properly, a multiple choice test item must provide at least four alternative responses (possible answers).

F 199. Unless special precautions against it are taken, objectives test items may give students a false notion that a complex question can be answered adequately by one and only one brief, simple statement.

F 200. The best vocabulary test items use the statement of a definition as the stem, and words that conceivably might fit the definition as the responses.

F 200. The best vocabulary test items use the statement of a definition as the stem, and words that conceivably might fit the definition as the responses.

F 201. The stem of a multiple choice item should be limited to a single sentence or sentence fragment. Multiple sentence item stems should be avoided.

F 202. It is often difficult and seldom advantageous to make all of the responses to a multiple choice item parallel in point of view, grammatical structure or general appearance.

F 203. To maximize the reliability of the test scores, item difficulty values should be normally distributed.

F 204. The discrimination index is used mainly to distinguish between easy and hard items.

T 205. Most classroom tests will differentiate various levels of achievement better if they include no very difficult and no very easy items.

F 206. If an ordinary guessing correction is applied, the wise examinee will answer only those questions that he is reasonably sure of answering correctly.

F 207. A well balanced program for testing student achievement will include some tests that come as a complete surprise to the students.

F 208. Time spent attempting to teach students how to do well on an objective test is usually time wasted.

T 209. If a test is properly constructed, the advantage of the test-wise student over the test-naive student is greatly reduced.

F 210. Test anxiety is a major cause of the low validity that plagues most paper and pencil tests of educational achievement.

F 211. To obtain the most valid scores from an achievement test, the test administrator should refrain from attempting to influence the student's natural rate of work on the test.

F 212. Honor systems provide the best general answer to problems of cheating on achievement tests.

T 213. Test scoring machines have been developed that are economical and accurate enough to replace hand scoring for most classroom tests.

F 214. If 10 students guess blindly at the answers to 25 true-false questions, and if their scores are corrected for guessing, all 10 will get scores of zero.

T 215. If a student attempts 80 of 100 questons on a true-false test and answers 55 of the 80 correctly, his score corrected for guessing might be 65.

F 216. The correlation between scores corrected for guessing and the uncorrected scores is ordinarily positive but low.

T 217. If 100 students all guess blindly at all 64 items in a true-false test, most should be expected to get scores below 40.

T 218. Correction for guessing is more useful on a speed test (i.e. one with a short time limit) than on a power test (i.e. one with a generous time allowance).

F 219. Students who attempt all items regardless of certainty tend to get much lower scores than their equally able but more cautious colleagues when a typical guessing correction is applied to the scores.

F 220. Differential weighting of items, or of distracters in multiple choice test items, almost always makes a substantial improvement in the reliability and validity of the test scores.

F 221. It is easier to correct test scores for guessing when the tests are scored by ahdn than when the tests are scored by machine.

T 223. Studies of guessing corrections suggest that scores of highest reliability are obtained when students think of guessing correction will be applied, but when in fact no correction is made.

T 223. For scores 1,3,5 the median is the same as the mean.

F 224. For scores 6,8,9 the median is the same as the mean.

T 225. If the scores in a set are all the same, the variance of the scores is zero.

T 226. If the variance of a set of scores is 15, the standard deviation of those scores is less than 4.

T 227. If the variance of a set of scores is 1, the standard deviation is also 1.

T 228. The variance of the set of scores 5,10,15 is the same as the variance of the set of scores 25,30,35.

F 229. If two sets of scores have different means they must have different variances.

T 230. If two sets of scores have different variances they must have different standard deviations.

F 231. If two sets of scores have the same mean they must have the same median.

F 232. The larger the number of scores in a set the larger the standard deviation of those scores must be.

F 233. If a set of scores includes only two scores, the percentile rank of the higher score is 67.

T 234. If a set of scores includes twenty scores, and if the two highest scores in the set are both 28, the percentile rank of 28 is 95.

T 235. Under the preferred definition of percentile ranks, no student receives a score whose percentile rank is 100.

T 236. In a set of 25 test scores, the higher one of any two different scores in the set always has the higher percentile rank.

T 237. When 25 test scores are converted to stanines, some different scores will always be assigned to the same stanine equivalent.

F 238. When 25 test scores are converted to stanines, some of the same test scores will always be assigned to different stanine equivalents.

T 239. If a set of 25 test scores consists of all the whole numbers beginning with 1 and ending with 25, a score of 14 will be assigned as stanine equivalent of 5.

F 240. If in a set of scores more scores fall above than below the mean there will be more stanine equivalents above than below 5.

MULTIPLE CHOICE ITEM POOL

APPENDIX B

1. Making judgments about the worth of an educational program objectives information is called

   *a. evaluation
   b. measurement
   c. standardization
   d. testing
   e. validation

2. Which of the following represents an "evaluative" statement?

   a. John's score is three standard deviations above the mean of his class.
   b. Eighty percent of this class scored above the national median.
   *c. Jean had a good score on this test.
   d. Mary had a higher score than Bill.

3. Historically, educational measurement has been least concerned with

   *a. goal priority determination.
   b. instructional organization.
   c. objective measurement.
   d. goal assessment.

4. Ultimate educational goals as opposed to immediate instructional objectives

   a. are not sufficient guidelines for educational audiences to determine instructional direction.
   b. require explicit clarification to establish their meaningfulness.
   c. are not amenable to adequate evaluation requirements.
   *d. all of the above.

5. In theory, which of the following is the <u>least</u> important criterion in selecting specific objectives?

   a. Close relation to student behavior
   b. Acceptance of a common meaning
   *c. Relation to textbook material
   d. Agreement with broad goals

6. Problems arise in attempting to develop measures of ultimate objectives mainly because

   *a. such goals concern behavior not usually observable under classroom conditions.
   b. teachers have been reluctant to depart from traditional testing methods.
   c. measurement methods have not given proper weight to all objectives.
   d. it is difficult to construct tests in broadly defined areas.
   e. teachers do not have time to develop adequate test instruments.

7. The mastery model suggests that the degree of learning is a function of

   a. unit time spent by instructional period.
   b. learning facility by unit time.
   *c. time spent by time required.
   d. interest by time unit.
   e. aptitude by interest.

8. Which one of the following exemplifies norm referenced measurement?

   a. Alice got 63 out of 100 points on the anthropology final.
   b. Joe was admitted to honors college this year.
   c. Ed earned a 4.0 in his measurement class.
   *d. Susan's score was average for her class.

9. Classical measurement theory developed primarily from a concern for measuring individual's levels of

   a. achievement
   *b. aptitude
   c. characteristics
   d. interest

10. Criterion referenced tests are useful in

   a. applying instructional decision making to analysis of student achievement.
   b. new curricula such as I.P.I. (Individually Prescribed Instruction).
   c. testing with the mastery model of learning.
   *d. all of the above.

11. We are assuming at least which level of measurement when we compute the standard deviation of a distribution?

*a. Interval
b. Nominal
c. Ordinal
d. Ratio

12. If one point is added to each score in a set of scores, which one of the following measures would need to be corrected?

a. The reliability coefficients
b. The validity coefficient
c. The standard deviation
d. The mean
e. All of the above

13. In symmetrical distributions the mean is

a. sometimes larger and sometimes smaller than the median, depending on the range.
*b. always identical to the median.
c. always smaller than the median.
d. always larger than the median.

14. Which of the following statements about the mean and median is true for all distributions?

a. 50 percent of the scores will fall below the mean and the percent falling below the median may be less than, equal to, or more than 50 percent.
*b. 50 percent of the scores will fall below the median and the percent falling below the mean may be less than, equal to, or more than 50 percent.
c. 50 percent of the scores will fall below the mean and less than 50 percent will fall below the median.
d. 50 percent of the scores will fall below the median and less than 50 percent will fall below the mean.

15. In a frequency distribution of 270 scores, the mean is 66 and the median is 79. One would expect this distribution to be

*a. negatively skewed
b. normal
c. positively skewed
d. rectangular
e. symmetrical

16. A cumulative frequency graph is closest to a

    a. frequency polygon
    *b. percent curve
    c. bar graph
    d. histogram

17. The degree of homogeneity of a class on a particular measure is indicated by the

    a. correlation coefficient
    b. covariance
    c. mean
    d. mode
    *e. semi-interquartile range

18. What percent of students in a distribution fall between the first and third quartiles?

    a. 25
    *b. 50
    c. 68
    d. 75
    e. Indeterminate without having the distribution

19. The middle 2/3 of the scores on a test with a normal distribution fall in the range 85 to 125. The standard deviation of the test might be estimated to be

    a. 8
    b. 10
    c. 15
    *d. 20
    e. 30

20. To compute a correlation coefficient between traits A and B, one must have

    a. one group of subjects, some of whom possess characteristics of trait A, the remainder possessing those of trait B.
    b. one group of subjects, some who have both A and B, some with neither, and some with one but not the other.
    c. two groups of subjects, one which could be classified as A or not A, the other as B or not B.
    d. measures of trait A on one group of subjects, and of trait B on another.
    *e. measures of traits A and B on each subject in one group.

21. Which of the following is theoretically handled as part of a person's true score?

*a. the traits measured are generally unstable
b. measurements cannot be made directly
c. instruments are not precise
*d. all the above

22. Which of the following is theoretically handled as part of a person's true score?

*a. constant error
b. random error
c. both a and b
d. neither a nor b

23. An individual's score on an achievement test is 73. The standard error of measurement for the test is reported to be 30 points. What are the chances that the individual's true score is between 70 and 76.

a. About 1 chance in 3.
b. About 1 chance in 6.
*c. About 2 chances in 3.
d. About 9 chances in 10.
e. About 19 chances in 20

24. Which of the four sets of data below will product the highest reliability of difference scores?

a. $r_{xx} = .80$, $r_{yy} = .80$, $r_{xy} = .80$

b. $r_{xx} = .50$, $r_{yy} = .50$, $r_{xy} = .40$

c. $r_{xx} = .80$, $r_{yy} = .80$, $r_{xy} = .40$

d. $r_{xx} = .70$, $r_{yy} = .90$, $r_{xy} = .00$

25. The ratio of the true score variance to the total test score variance is a measure of

a. standard measure for variability
b. standard error of measurement
*c. consistency in measurement
d. relevance in measurement

26. The correlation between a predictor and a criterion was found to be .7. This means that the predictor technically accounts for ____ of the variance in criterion measure.

a. 70 percent
*b. 49 percent
c. 14 percent
d. 7 percent

26. The correlation between a predictor and a criterion was found to be .7. This means that the predictor technically accounts for ___ of the variance in criterion measure.

a. 70 percent
*b. 49 percent
c. 14 percent
d. 7 percent

27. If scores are distributed normally, the percentile rank of a score one standard deviation below the mean is about

a. 5
*b. 16
c. 34
d. 68
e. 84

28. A distribution of linear z-scores is always

*a. similar to the raw score distribution
b. a rectangular distribution
c. a normal distribution
d. a peaked distribution
e. a skewed distribution

29. A person whose z-score was zero would have a raw score equal to

*a. the mean
b. the median
c. zero
d. none of the above

30. Assuming a reasonably normal distribution of scores, what would be our best guess for the percent of scores falling between the T values of 50 and 70?

a. 20%
b. 34%
*c. 47%
d. 50%
e. 68%

31. Given a group mean of 40 and a standard deviation of 4, a raw score of 46 would convert to a stanine of approximately

    a. four or less
    b. five
    c. six
    *d. seven or more

32. On test XYZ Pupil A scores at the 80th percentile and Pupil B scores at the 40th percentile. What statement may be made relative to their abilities?

    a. We need to know more about the distribution of scores before we can compare the two percentiles.
    *b. If pupil A's score is superior to those of 70 pupils, Pupil B's score is superior to those of 35.
    c. Percentile ranks are meaningless in comparing the performance of A and B.
    d. Pupil A's test score will be twice that of Pupil B.
    e. Pupil A has twice the measured ability of Pupil B.

33. The mean score on a 70 item test is 49, and the standard deviation of the scores is 6. What z-score should be assigned to a raw score of 40%?

    a. -9
    b. -1.5
    c. 5
    d. 35
    *e. none of the above

34. In a normally distributed set of score, which of the following measures represents the highest degree of relative performance?

    a. a percentile rank of 76.
    b. a stanine score of 7.
    *c. a T-score of 74.
    d. a z-score of 1.5.

35. Regardless of the data involved, the act of making a proper interpretation requires one to

    *a. evaluate
    b. objectify
    c. postulate
    d. qualify

36. The more useful statements of objectives are ones which

a. emphasize attitudes and appreciations instead of knowledge.
b. apply to the general instructional needs of all pupils.
*c. relate skills and knowledge to observable behavior.
d. can be restated as items in objective tests.

37. Which of the following list of behaviors reflect an institutional decision.

*a. applying to college
b. hiring job applicants
c. joining the armed forces
d. selecting a course

38. Which one of the following verbs could be used to correctly state a behavioral objective?

a. appreciate
b. believe
*c. derive
d. understand

39. A criterion-referenced tests should be utilized when making which of the following educational decisions?

a. To award a limited number of scholarships.
*b. To certify a student for secretarial proficiency.
c. To compare students in different school district.
d. To select a student for an honor program.

40. Criterion-referenced tests are useful in

a. applying instructional decision making to analysis of student achievement.
b. new curricula such as IPI (Individually Prescribed Instruction).
c. testing with the mastery model of learning.
*d. all of the above.

41. The major objective of curriculum evaluation is to

a. eliminate unintended or expressive instructional outcomes.
b. make formative and summative evaluation of curricular programs.
*c. judge the effect of the cirriculum on students.
d. assess students' instructional goal attainment.

42. Which measure will best represent the central tendency of the scores 2, 21, 23, 20, 33, 23?

a. Mean
*b. Median
c. Mode
d. One is as good as the other

43. Measures of variability provide an index of the

a. achievement of a group in relation to the norms.
b. specific weaknesses of individual pupils.
c. shape of the distribution for a group.
*d. relative homogeneity of a group.
e. over-all ability of a group.

44. Mrs. Carter correlates her classes' scores on a science and spelling test. She obtains an r - .95. This means that

a. her class did well on both measures.
*b. the top spellers are generally the top in science.
c. the test discriminates well for her class.
d. the test is quite reliable.

45. With objective tests which source of error typically causes the least problem

a. trait instability.
b. administrative errors.
c. sampling errors.
*d. scoring errors.
e. personal errors.

46. The standard error of measurement is closely related to

a. central tendency.
b. difficulty.
c. objectivity.
*d. reliability.
e. validity.

47. The standard error of measurement is useful for

*a. reporting an individual's score within a band of the score range.
b. converting individual raw scores to percentile ranks.
c. reporting a group's average score.
d. comparing group differences.

48. A ninth grade class was administered Form A of a test followed immediately by Form B. The test scores from the two forms placed the students in nearly the same order. This would be evidence concerning the test's

*a. equivalence reliability.
b. stability reliability.
c. concurrent validity.
d. predictive validity.
e. content validity.

49. The Kuder-Richardson Formula #20 for calculating reliability uses

a. Form A at one time and Form A at a second testing time.
b. Form A and Form B scores.
*c. every item on the test.
d. odd and even subscores.

50. The advantage of Cronbach's Coefficient Alpha over the KR_{20} formula that

a. the sum of p and q products need not be computed.
b. item variances need not be assumed equal.
*c. items need not be scored right or wrong.
d. mean item variance is omitted.

51. The correlation between test scores and a criterion is a measure of

a. causation.
b. objectivity.
c. reliability.
*d. validity.
e. variability.

52. Mr. Smith compares the items on the Cooperative Mathematics Test with his course objectives. With what kind of test validity is he concerned?

a. concurrent
b. construct
*c. content
d. predictive

53. Which statement <u>best</u> represents the definition of <u>test norms</u>?

a. A relative performance standard for a desired group.
b. A representative performance for a selected group.
*c. Typical performance based on specified criteria.
d. An estimated performance by an inclusive group.

54. If the percentile equivalents of test scores are plotted as a frequency distribution, what will be its shape?

a. Negaitvely skewed because the mean on most tests is above 50 of the total possible score.
b. You must know the shape of the original distribution to answer the question.
c. The same as the test score distribution.
*d. Rectangular.
e. Normal.

55. On a record form someone failed to record Harold's percentile, but wrote his raw score, 39. The test manual was missing, but the counselor found two other records where both raw score and percentile were reported as follows:

Edwin:	raw 33	percentile 70
Mike:	raw 36	percentile 80

Assuming a normal distribution of raw scores, our "best bet" for Harold's percentile is:

*a. less than 90
b. 90
c. more than 90
d. Not enough information given to answer

56. Criterion referenced tests used for decision making should generally be longer for

a. groups.
*b. individuals.
c. equal for groups and individuals.

57. Which of the following methods of item construction for criterion-referenced tests begins with a definition of the objectives to be measured?

a. Panel of experts
b. Systematic snapling
c. Systematic item generation
*d. All of the above

58. A major advantage of criterion referenced tests rests on the fact that they

a. are easier to construct.
b. are easier to score.
c. provide higher internal validity coefficients.
*d. provide absolute measures of performance.

59. Which of the following forms of evaluation can be carried out at any stage of the total evaluation?

*a. diagnostic
b. formative
c. placement
d. summative

60. Which of the following type of evaluation is <u>least</u> likely to be judged against criterion-referenced standards?

a. diagnostic
b. formative
c. placement
*d. summative

BIBLIOGRAPHY

Aiken, Jr., Lewis R. Psychological Testing and Assessment, Allyn and Bacon, Inc., Boston, Mass. 1976.

Anastasi, Anne. Psychological Testing, The MacMillan Company, Collier-MacMillan Limited, London 1968.

Anderson, Richard C. "How to Construct Achievement Tests to Assess Comprehension," Review of Educational Research, Vol. 42 No. 2, 1972, p. 145-170.

Brennan, Robert L. "A Generalized Upper-Lower Item Discrimination Index," Educational and Psychological Measurement, Vol. 32, 1972, p. 289-303.

Campbell, Donald T. and Stanley, Julian C. Experimental and Quasi-Experimental Designs for Research, Rand McNally College Publishing Company, Chicago, 1973.

Cardinet, Jean; Tourneur, Yvan; and Allal, Linda. "The Symmetry of Generalizability Theory: Applications to Educational Measurement," Journal of Educational Measurement, Vol. 13, No. 2, Summer 1976, p. 119-136.

Carver, Ronald P. "Two Dimensions of Tests Psychometric and Edumetric," American Psychologist, July 1974, p. 512-518.

Carver, Ronald P. "Reading Tests in 1970 Versus 1980: Psychometric Versus Edumetric," The Reading Teacher, Vol. 26, December 1972, p. 299-302.

Carver, Ronald P., and Darby, Jr., Charles A. "Analysis of the Chunked Reading Test and Reading Comprehension," Journal of Reading Behavior, Vol. 5, No. 4, Fall 1972, p. 282-296.

Chase, Clinton I. Measurement for Educational Evaluation, Addison-Wesley Publishing Company, Reading, Mass., 1974.

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

Costin, Frank. "Three-Choice Versus Four-Choice Items: Implications for Reliability and Validity of Objective Achievement Tests," Educational and Psychological Measurement, Vol. 32, 1972, p. 1035-1038.

Cox, Richard C. "An Empirical Investigation of the Effect of Item Selection Techniques on Achievement Test Construction," Ph.D. Thesis, Michigan State University, 1964.

Cox, Richard C. "Item Selection Techniques and Evaluation of Instrumental Objectives," Journal of Educational Measurement, 1965, p. 181-185.

Crehan, Devin D. "Item Analysis for Teacher-Made Mastery Tests," Journal of Educational Measurement, Vol. 11, No. 4, Winter 1974, p. 255-262.

Davis, Frederick B. "Item Analysis in Relation to Educational and Psychological Testing," Psychological Bulletin, Vol. 49, No. 2, March 1952, p. 97-121.

Diederich, Paul B. "Short-cut Statistics for Teacher-Made Tests," Eric Document No. ED081785, 1973.

Ebel, Robert L. Essentials of Educational Measurement, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1972.

Ebel, Robert L. "Expected Reliability as a Function of Choices per Item." Educational and Psychological Measurement, Vol. 29, 1969, p. 565-570.

Ebel, Robert L. Measuring Educational Achievement, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1965.

Ebel, Robert L., et al. "Multiple Choice Items for a Test of Teacher Competence in Educational Measurement," Document prepared for National Council on Measurement in Education, Iowa State University, May 1962.

Ebel, Robert L. "The Relation of Item Discrimination to Test Reliability," Journal of Educational Measurement, Vol. 4, Fall 1967, p. 125-128.

Ebel, Robert L. "The Reliability of an Index of Item Discrimination," Educational and Psychological Measurement, Vol. 11, 1951, p. 403-408.

Findley, Warren G. "A Rationale for Evaluation of Item Discrimination Statistics," Educational and Psychological Measurement, Vol. 16, 1956, p. 175-180.

Flanagan, John C. "Calculating Correlation Coefficients," American Institute for Research, Pittsburgh, Pa. 1962.

Flanagan, John C. "The Effectiveness of Short Methods for Calculating Correlation Coefficients," Psychological Bulletin, Vol. 49, July 1952, p. 342-348.

Forsyth, Robert A., and Feldt, Leonard S. "An Investigation of Empirical Sampling Distributions of Correlational Coefficients Corrected For Attenuation," Educational and Psychological Measurement, Vol. 29, 1969, p. 61-71.

Forsyth, Robert A., and Feldt, Leonard S. "Some Theoretical and Empirical Results Related to McNemar's Test that the Population Correlation Coefficient Corrected for Attenuation Equals 1.0," American Educational Research Journal, Vol. 7, No. 2, March 1970, p. 197-207.

Frisbie, David A. "Comparative Reliabilities and Validities of True-False and Multiple Choice Tests," Ph.D. Thesis, Michigan State University, 1971.

Frisbie, David A. "Multiple Choice Versus True-False: A Comparison of Reliabilities and Concurrent Validities," Journal of Educational Measurement, Vol. 10, No. 4, Winter 1973, p. 297-304.

Frisbie, David A. "The Effect of Item Format on Reliability and Validity: A Study of Multiple Choice and True-False Achievement Tests," Educational and Psychological Measurement, Vol. 34, 1974, p. 885-892.

Glass, Gene V., and Stanley, Julian C. Statistical Methods in Education and Psychology, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970.

Green, Robert L.; Nyquist, Julie G.; and Griffore, Robert J. "Standardized Achievement Testing: Some Implications for the Lives of Children," Paper presented at National Institute of Education's Test Bias Conference, December 2 through 5, 1975, Washington, D.C.

Gruber, Howard E., and Weitman, Morris. "Item Analysis and the Measurement of Change," Journal of Educational Research, Vol. 55, No. 6, March 1962, p. 285-289.

Guilford, J. P. Psychometric Methods, McGraw-Hill Book Co., New York, 1954.

Hales, Loyde W. "Method of Obtaining the Index of Discrimination for Item Selection and Selected Test Characteristics: A Comparative Study," Educational and Psychological Measurement, Vol. 32, 1972, p. 929-937.

Helmstadter, G. C. "A Comparison of Traditional Item Analysis Selection Procedures With Those Recommended for Tests Designed to Measure Achievement Following Perofrmance Oriented Instruction" Paper presented at American Psychological Association Convention, Honolulu, Hawaii, September 1972.

Henryssen, Sten. Gathering, Analyzing, and Using Data on Test Items. In Robert L. Thorndike Educational Measurement 2nd edition, American Council on Education, Washington, 1971.

Hills, John R. Measurement and Evaluation in the Classroom, Charles E. Merrill Publishing Company, Columbus, Ohio, 1976.

Kelley, Truman. "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, Vol. 30, 1939, p. 17-24.

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

Kwansa, K. B. "Content Validity and Reliability of Domain Referenced Tests," African Journal of Education Research, Vol. 1, 1974, p. 73-79.

Lange, Allan; Lehmann, Irvin Jr.; and Mehrens, William A. "Using Item Analysis to Improve Tests," Journal of Educational Measurement, Vol. 4, 1967, p. 65-68.

Lennon, R. T. "A Glossary of 100 Measurement Terms," Test Service Notebook, No. 13, Harcourt Brace Jovanovich, New York.

Lentz, Jr., Theo F.; Hirshstein, Bertha; and Finch, F. H. "Evaluation of Methods of Evaluating Test Items," Journal of Educational Psychology, Vol. 23, 1932, p. 344-350.

Levine, Murray. "The Academic Achievement Test: Its Historical Context and Social Functions," American Psychologist, Vol. 31, No. 3, March 1976, p. 228-238.

Levine, Richard, and Lorel, Frederic M. "An Index of the Discriminating Power of a Test at Different Parts of the Score Range," Educational and Psychological Measurement, Vol. 19, No. 4, 1959, p. 497-503.

Lord, Frederic M. "A Significance Test for the Hypothesis that Two Variables Measure the Same Trait Except for Errors of Measurement," Psychometrica, Vol. 2, No. 3, September 1957, p. 207-220.

Magnusson, David, Test Theory, Addison-Wesley Publishing Co., Reading, Mass., 1966.

Marshall, Jon; Clark and Hales; and Loyde Wesley. Classroom Charles E. Merrill Publishing Co. 1971

McNemar, Quinn. "Attenuation and Interaction," Psychometrika, Vol. 23, No. 2, September 1958, p. 259-265.

Mehrens, William A., and Lehmann, Irvin J. Measurement and Evaluation in Education and Psychology, Holt, Rinehart and Winston, New York, 1972.

Millman, Jason, and Popham, W. James. "The Issue of Item and Test Variance for Criterion-Referenced Tests: A Clarification," Journal of Educational Measurement, Vol. 11, No. 2, Summer 1974, p. 137-138.

Oosterhof, Albert C. "Similarities of Various Item Discrimination Indices," Journal of Educational Measurement, Vol. 13, No. 2, Summer 1976, p. 145-150.

Oosterhof, Albert C. "Stability of Various Item Discrimination Indices," Paper presented at annual meeting of American Educational Research Association, New Orleans, Louisiana, February 25-March 1, 1973.

Popham, W. James, and Husek, T. R. "Implications of Criterion-Referenced Measurement," Journal of Educational Measurement, Vol. 6, No. 1, Spring 1969, p. 1-9.

Pyrczak, Fred. "Validity of the Discrimination Index as a Measure of Item Quality," Journal of Educational Measurement, Vol. 10, No. 3, Fall 1973, p. 227-231.

Ramos, Robert A., and Stern, June. "Item Behavior Associated with Changes in the Number of Alternatives in Multiple Choice Items," Journal of Educational Measurement, Vol. 10, No. 4, Winter 1973, p. 305-310.

Reynolds, Carl, and Cobean, Nancy. "Constructing an Edumetric Test," Paper presented at annual meeting of American Educational Research Association, San Francisco, Calif., 1976.

Roudabush, Glenn E. "Item Selection for Criterion-Referenced Tests," A paper presented at the American Educational Research Association meeting in New Orleans, February, 1973.

Saupe, Joe L. "Selecting Items to Measure Change," Journal of Educational Measurement, Vol. 3, No. 3, Fall 1966, p. 223-228.

Slakter, Malcolm J. Statistical Inference for Educational Researchers, Addison-Wesley Publishing Co., Reading, Mass. 1972.

Thomas, David R. "A Comparison of Edumetric and Psychometric Item Analysis Techniques," Unpublished Study not yet in print, 1975.

Woodson, M.I. Charles E. "Classical Test Theory and Criterion-Referenced Scales," ERIC Document No. 083298, Not Dated.

Woodson, M.I. Charles E. "The Issue of Item and Test Variance for Criterion-Referenced Tests," Journal of Educational Measurement, Vol. 11, No. 1, Spring 1974, p. 139-140.

Woodson, M.I. Charles E. "The Issue of Item and Test Variance for Criterion-Referenced Tests: A Reply Journal of Educational Measurement, Vol. 11, No. 2, Summer 1974, p. 139-140.