AN EVALUATION OF THE SEQUENTIAL METHOD OF PSYCHOLOGICAL TESTING

Thesis for the Degree of Ed. D.
MICHIGAN STATE UNIVERSITY
John James Paterson
1962

This is to certify that the

thesis entitled

AN EVALUATION OF THE SEQUENTIAL METHOD
OF
PSYCHOLOGICAL TESTING

presented by

John James Paterson

has been accepted towards fulfillment
of the requirements for

Ed.D. degree in Education

David R. Krathwohl
Major professor

Date June 15, 1962

O-169

ABSTRACT

AN EVALUATION OF THE SEQUENTIAL METHOD OF PSYCHOLOGICAL TESTING

by John J. Paterson

In the sequential method of psychological testing the examinees are directed to subsequent items on the basis of their responses to prior items. No examinee responds to all the items of a sequential test, and any given examinee might complete the test by responding to any of several combinations of items. Scores on the sequential test reflect the difficulty of items correctly answered not the number correct.

The evaluation did not involve an actual population of individuals, but used probability models and hypothetical populations. The probability of passing a given item in a test was calculated from the ability level of the individual, the difficulty of the item, and the precision of the item. (Precision may be computed from the item-total biserial correlation.) The probability of passing a sequence of items was determined for each of fifteen ability categories by multiplying together the probabilities of passing or failing a sequence of six items. Sixty-four different sequences were calculated for each ability category.

The problem involved was the comparison of the sequential model with the traditional cumulative model (in which

all items we
determine ho
were classif
rial test (d
errors in es
relation to

One se
item-total b
such that th
wal's abili
into which t
Even though
separate ex
in different
the calculat
deviation un
ability was

Both m
were used ac
test models
individuals
that regardl
input; the l
significantly
at middle ab
lower variance

all items were at the 50 per cent level of difficulty) to determine how well individuals at different ability levels were classified by the tests. The parameters of the sequential test (difficulty and precision) and the effects of errors in estimating these parameters were examined in relation to the resulting classification of individuals.

One sequential test model was constructed with an item-total biserial correlation of .75 and item difficulties such that the sum of the squared deviations of the individual's ability level from the mean ability level of the group into which the individual was classified would be a minimum. Even though individuals in each ability category were kept separate from individuals in other categories, individuals in different categories took the same difficulty item if the calculated difficulties were less than .20 standard deviation units apart. A rectangular distribution of ability was assumed in these calculations.

Both normal and U-shaped distributions of ability were used as input for the above sequential and cumulative test models to determine how well the results classified individuals of different ability levels. It was concluded that regardless of the distribution of ability used as input; the individuals in the extreme ability categories had significantly less variance of scores in the sequential test. At middle ability levels the sequential test did have slightly lower variance of test scores than the cumulative. For the

top scores t

level than t

The se

each separat

resulting nu

level of ind

for top and

of ability i

individuals

test was not

and distribut

When pr

varied, tests

troubles app

variance of s

variance of a

scoring indiv

It was

to distinguish

items among t

variance of ab

should be regr

which give the

similar abilit

top scores the sequential test had less variance of ability level than the cumulative.

The second and fifth items in the sequential test were each separately changed in difficulty and precision. The resulting number of people at each score, the mean ability level of individuals at each score, the variance of scores for top and middle ability level individuals and the variance of ability level scores for the top and middle scoring individuals were all insignificantly changed. The sequential test was not sensitive to errors in estimating the precision and difficulty of the items.

When precision of items in the sequential tests was varied, tests consisted of higher precision items (with difficulties appropriate for that precision level) had less variance of scores for ability level categories and less variance of ability level categories for top and middle scoring individuals.

It was concluded that more difficult items are needed to distinguish among more able students; less difficult items among the less able. If extreme scores having low variance of ability level are desired, the item difficulties should be regressed toward the mean from those difficulties which give the best discrimination between individuals of similar ability level.

AN EVALUATION OF THE SEQUENTIAL METHOD OF PSYCHOLOGICAL TESTING

by

John James Paterson

A THESIS

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

DOCTOR OF EDUCATION

College of Education

1962

ACKNOWLEDGMENTS

The writer wishes to express his appreciation for the guidance given by Dr. David R. Krathwohl in the preparation of this thesis and to the Bureau of Educational Research for arranging the time necessary for completion of the research.

TABLE OF CONTENTS

TABLE

1. Analy
Score
Distr
and C

2. Analy
Score
Normal
Sequen

3. Analy
Level
Score
Abilit
Tests

4. Level
Analy
Scale

5. Level
Analy
Scale

6. Level
Analy
Scale

7. Analy
Scores
Scale
Sequen

8. Analy
Scores
Scale
Sequen

9. Analy
Level
Scores
Scale

LIST OF TABLES

TABLE

10. Anal
uals
of D

11. Anal
Scor
Leve

12. The
to Ex
Diff

13. The
Atti
cult
Pred

14. Dist
Test
of Po

15. Dist
Test
of Po

16. Anal
Score
Level
in D

17. Anal
[illegible]

18. Dist
[illegible]

19. Dist
[illegible]

20. Dist
[illegible]

TABLE PAGE

TABLE

21. Distri
Top S
Prefis
tion o

22. Distri
Top S
Items
Distri

23. Per Ce
Sequen

24. Mean
Level
Sequen

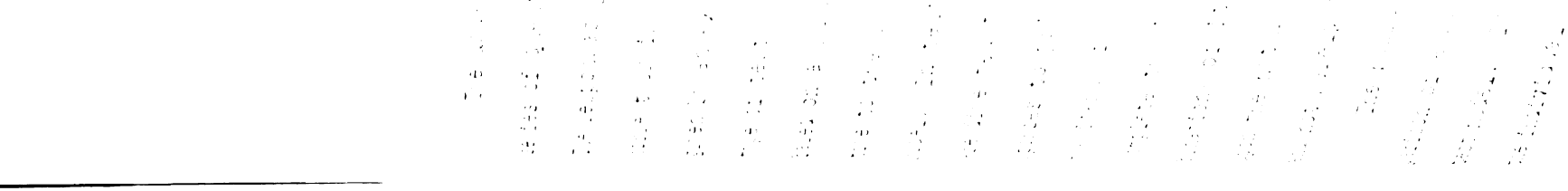
25. Distri
Sw
Distri

26. Distri
Stores
the I

27. Distri
Stores
in a S
Distri

28. Distri
Stores
Sequen
butio

LIST OF FIGURES



CHAPTER I

DESCRIPTION OF THE PROBLEM

The usual objective-type mental test consists of a series of questions or items to which examinees respond. The responses given to each item are scored as either correct or incorrect and the number of correct responses given by an examinee is taken as his score. This traditional type of test and the scoring procedure used with it are based on a cumulative model of test behavior. One alternative to the cumulative model may be called the sequential model. In tests based upon the sequential model, examinees are directed to subsequent items on the basis of their responses to prior ones. Thus no examinee responds to all of the items of a sequential test and any given examinee might complete the test by responding to any of a variety of combinations of items. Scores on sequential tests are based upon the nature of items to which correct responses are given and not merely the number of correct responses.

The basic problem considered in this paper is the comparison of a sequential model with the traditional cumulative test model. In the sequential model developed in this paper the individual who passes an item is automatically directed

to a more di

to an easier

al who pars

if the iter

item closer

The opposite

directly rel

individual i

In add

sequential t

methods of i

the results

better than

may be impro

The pr

psychologica

features of

cumulative t

encountered

problems are

for the sequ

hypotheses a

test models

hypotheses a

overview of

to a more difficult item; if he fails an item he is directed to an easier item. If the item is very precise, the individual who passes it is given a much more difficult item; and if the item is not very precise, the individual is given an item closer to the difficulty level of the item just answered. The opposite is true for failing an item. The score is directly related to the difficulty of the item to which the individual is directed at the final stage of testing.

In addition to the comparison of testing methods, the sequential test is examined for its strengths and weaknesses. Methods of improving the sequential model are suggested from the results so that even if the present procedure is not better than the cumulative test, future sequential procedures may be improved.

The present evaluation of the sequential method of psychological testing consists of (1) a description of the features of the sequential method as compared with the usual cumulative test; (2) a description of some of the problems encountered in the use of the cumulative test and how these problems are handled by the sequential model; (3) a rationale for the sequential solution; and (4) the formulation of hypotheses as to the behavior of the cumulative and sequential test models in regard to specific problems. Following the hypotheses are (5) the limitations of the study and (6) an overview of the remainder of the dissertation. To aid the

reader a few

sentation are

I. I

In any

(1) the indi

decision must

score must b

succeeding i

be informed

cumulative

Because they

decision p

Starting Poi

Depend

therefore w

any level of

easy item t

while the i

with an ite

consideratio

test. The

the individ

covered at

50% of whi

reader a few of the more frequently used terms in this dissertation are explained in Appendix B.

I. DESCRIPTION OF THE SEQUENTIAL TEST MODEL

In any testing situation certain decisions must be made: (1) the individual must be told where to start, (2) the decision must be made when to stop testing, (3) the final score must be determined, (4) the characteristics of each succeeding item must be stipulated, and (5) the testee must be informed as to where and how he should proceed. In the cumulative test the character of these decisions is obvious. Because they are unusual in the sequential item test, these decision points will be described in some detail.

Starting Point

Depending on the purpose of the test and what one therefore wishes to emphasize, the starting point may be at any level of difficulty. For instance, one may start with an easy item that most individuals will be able to pass and with which the individual would feel comfortable, or one may start with an item at the middle of the score distribution with no consideration as to the individuals who may be taking the test. The sequential test model developed in this paper has the individual take as his first item one that would be considered at the fifty per cent level of difficulty for the group of which he is a part. The reason for this choice is

explaine

must, of

need to b

Choosing [illegible]

Crit

by the p

in the ti

the same m

better cla

that the [illegible]

but could [illegible]

sertation

If time is

accuracy f

used for th

ability le

ability le

testing whe

probably no

model in th

Scoring

Reason

depend upon

group from

a group, or

explained in Chapter III, Section 1. The present discussion must, of necessity, ignore the psychological effects which need to be empirically determined.

Stopping Point

Criteria for deciding when to stop are also determined by the purpose of the test. If doing the best job possible in the time allowed is paramount, then everyone is given the same number of items knowing that the extremes will be better classified than the middle ability levels. (Note that the criterion measure need not be a measure of ability but could be an attitude or interest. However, in this dissertation the criterion will be referred to as an "ability.") If time is flexible and there is a prescribed degree of accuracy for each score, then a fewer number of items is used for the extreme and more items used for the middle ability levels. If the rapid classification of extreme ability level individuals is desired, then one may stop testing when it can be determined that the individual is probably not at some middle ability level. In the sequential model in this paper all people will take six items.

Scoring

Reasons for choosing one system of scoring over another depend upon whether the score is to discriminate one ability group from another, to discriminate among the individuals in a group, or to describe the response pattern of the individual.

If one wishes to discriminate one ability group from another, one would probably assign a score reflecting the difficulty of the final item. If one wishes to discriminate among individuals, then the score may represent the number of people in, for example, one hundred that the individual would rank above in the population. If the score is to represent a response pattern, it may be an estimate of the number of items the testee could have answered correctly if he does answer an item of a given difficulty, or it may identify the precise pattern of correctly and incorrectly answered items. The sequential test model in this paper assigns the individual a score which is the difficulty of the item to which he is directed at the final stage of testing as his score.

Pattern of Items

The problem in the sequential test is to select that sequence of items which will yield the information needed to assign the individual a score. At any stage in the test the decision as to the succeeding item to be taken may depend upon (1) the number of preceding items one has answered correctly, (2) the pattern of preceding items, or (3) the difficulty and precision of the immediately preceding item. This sequential model uses the difficulty and precision of all preceding items to determine the next item.

Difficulty of the item for this model is measured in terms of standard score units for a theoretically normal

group. An

would pass

item is

The measure

deviation

lated to the

$s = 1$

the "t_i",

<u>Directions to</u>

The

given item,

formable, or

combination

the test.

Individua

of those who

an item

response may

and, or each

and sequence.

but only

than one possib

the item.

Frederic

Monograph No.

group. An item that fifty per cent of the theoretical group would pass is designated as "0.00." The precision of the item is essentially a measure of the validity of the item. The measure of precision, σ_d, may be defined as the standard deviation of the item characteristic curve. (It is also related to the measure of precision "h" used in psychophysics: $h = 1/(2\sigma_d)^{1/2}$; and, as Lord indicates, σ_d is identical with his "b_i".)[1]

Directions to Testee

The testee may be told how well he performed on any given item, may be told what is right or wrong with his performance, or may be simply directed to another item. Any combination of the above may be used at different stages in the test.

Individuals may be directed to items which are taken by those who perform differently, or they may be directed to an item unique to their pattern of response. Pattern of response may be determined from correctness or incorrectness only, or each alternative to any item may designate a different sequence. In this sequential test, pattern was determined from only correctness or incorrectness of items, and more than one possible sequence of responses could lead to the same item.

---

[1]Frederic M. Lord, A Theory of Test Scores, Psychometric Monograph No. 7 (Chicago: University of Chicago Press, 1952), p. 7.

Many methods of giving the necessary information to the testee are available. In the empirical tests that have been built by Krathwohl and Paterson, the succeeding item that the individual should attempt is disclosed to the individual when he erases the opaque covering under the letter that has been selected as the answer to the question at hand. The final erasure disclosed a letter used to indicate a score rather than the number of the next item.[2] The testee must answer each item as he comes to it as he receives no directions if he does not answer. Other response techniques which could be used are tabs, envelopes within envelopes, sliding masks, and scrambled books.

A Diagram of a Sequential Test Used in This Study

Figure 1 is a diagram of one of the sequential tests used in this study. It is the one constructed by the "least squares" method which is described later. The pattern shown is only one of many possible sequential patterns.

Difficulty of items.--Items are represented by circles, the ordinate position of which represents the difficulty of the item. The closer the item is to the top of the page, the more difficult it is. Difficulty is expressed in standard score units, i.e., an item that fifty per cent of the normative

---

[2]Unpublished material developed in the Bureau of Educational Research, Michigan State University, East Lansing, Michigan, 1956-1959.

Difficulty of Items
(in standard scores)

1.00
.90

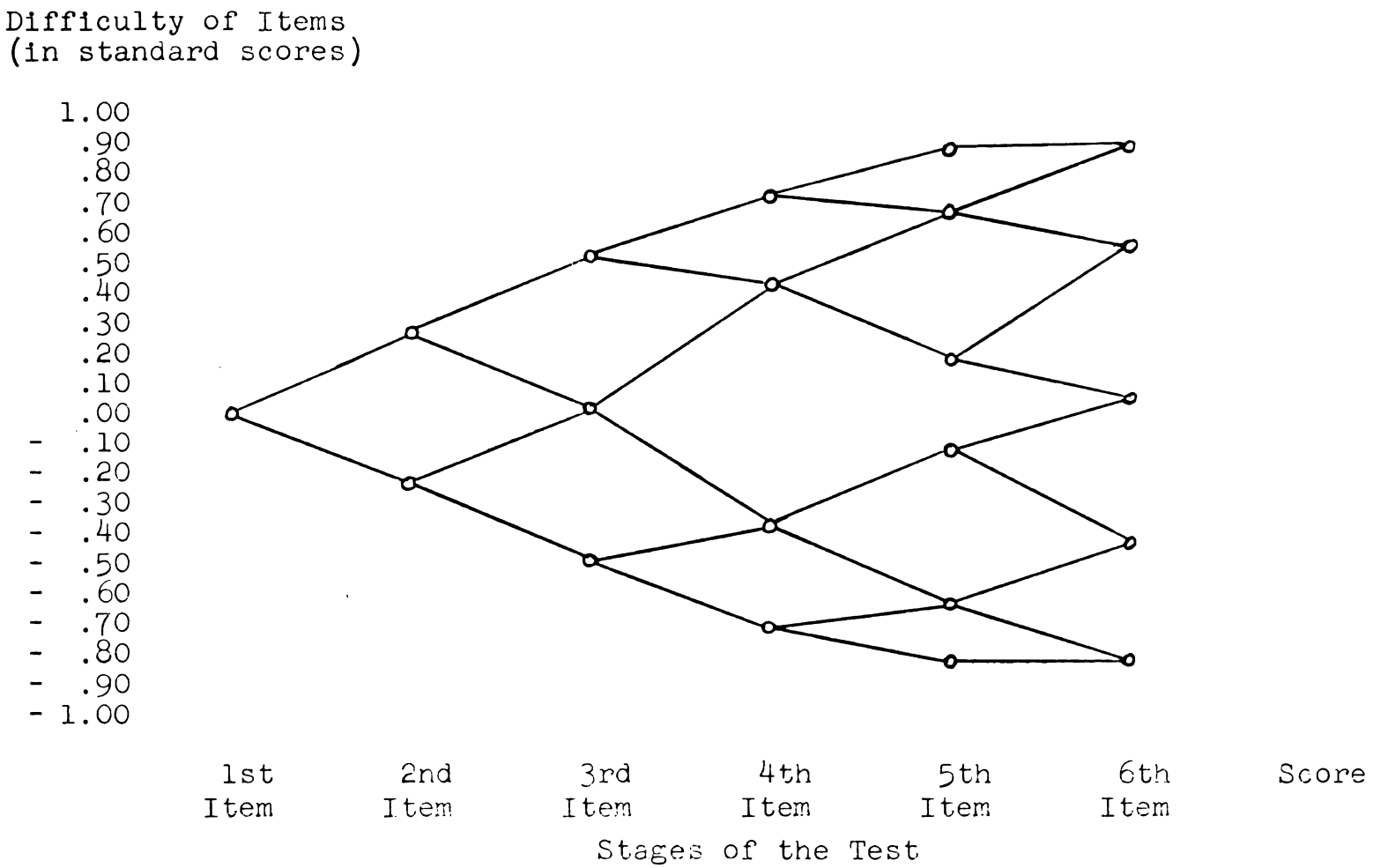


Fig. 1.--Graphic Representation of the "Least Squares" Sequential Test Model

group would answer correctly is labelled "0.00". An item that 84 per cent of the normative group would answer correctly is labelled "-1.00".

Sequence of items.--The sequence is represented by the abcissa value for the item. The first item of the test is at the left-hand side; the sixth item at the right of the diagram. The individual confronts one item at each "stage" of the test.

Size of step.--The size of the step or the increase or decrease in difficulty from the item at one stage to the item at the next stage is represented by the difference in ordinate positions of the items as can be seen in Figure 1. There would be a large increase in the difficulty of the second item if one were to correctly answer the first item. There would be less difference between the easiest item at stage four and the easiest item at stage five.

Route taken.--Lines slanting upward designate that those who are considered to have passed an item at the preceding stage should proceed to a more difficult item for the next stage. Lines slanting downward designate that the individuals are considered to have failed the item at the previous stage and should proceed to a less difficult item at the next stage. It may occur that passing a less difficult item will lead the individual to a more difficult

item for the next stage than he would have encountered by failing a more difficult item. In this case the lines between items will cross. (This case is not illustrated in Figure 1.) The other alternative not yet mentioned is that individuals passing a less difficult item or failing a more difficult item may be lead to the same difficulty of item at the succeeding stage.

II. NEED FOR TEST IMPROVEMENT

In order to lay the background as to why the sequential test is worth considering, one should examine what problems have been encountered in the use of the cumulative test. Present test procedures seem to have encountered three important problems related to: (1) utilization of items to operate most efficiently with the group taking the test, (2) controlling the score distribution to arrive at a useful scale, and (3) production of a score with a precise meaning.

Maximally Efficient Use of the Items Selected

Once one has decided upon a purpose, then one can solve the problem of the most efficient selection of items either completely empirically, or theoretically in terms of the effect of varying certain item characteristics. The approach in this paper is the theoretical one. If one uses this theoretical approach, one of the problems is that of utilizing the most precise items available in a pool. The

cumulative t

items.

In the

correct resp

ficulty, the

better meas

precise item

The at

paradox." T

following ab

the attenuati

in ability

scores on ab

interval sca

homoscedasti

that violati

the lack of

precision,

relation wi

One re

that validity

noted by Er

cumulative test cannot always use all of the more precise items.

In the cumulative test, if the score is the number of correct responses and if all of the items are of equal difficulty, then a test with less precise items would give a better measure of the scale of ability than a test with more precise items.[3]

The above phenomenon has been called the "attenuation paradox." Violation of any one or a combination of the following assumptions has been given as an explanation for the attenuation paradox: (1) scores are normally distributed, (2) ability is normally distributed, (3) the regression of scores on ability is linear, (4) measurement produces an interval scale of ability, and (5) response distribution is homoscedastic. There is evidence to support the contention that violation of any one of these could be the reason for the lack of a monotonic relationship between item reliability (precision) and the validity of scores in the usual testing situation with the cumulative test.

One method of using the most precise items and increasing test validity is to use a spread of item difficulties as suggested by Brogden.[4] However, this does not seem to be a

---

[3]Ledyard R. Tucker, "Maximum Validity of a Test with Equivalent Items," Psychometrika, 11:1-14; March, 1946.

[4]Hubert E. Brogden, "Variation in Test Validity with Variation in the Distribution of Item Difficulties, Number of Items, and Degree of their Intercorrelation," Psychometrika, 11:197-214; December, 1946.

completely satisfactory solution because (1) there is no scheme to determine the appropriate spread and (2) the most extreme difficulties cannot be efficiently used any time the majority of the individuals taking the item guess at the answer.[5] There should be some procedure which would allow use of precise items no matter what their difficulty level. If items are to be efficiently used in the discrimination of a group into two parts, the items should be at the 50 per cent level of difficulty for the hypothetical group the median ability level of which is at the point where the discrimination is desired.[6] This means that if discriminations are desired among a few high ability individuals then difficult items should be used. The usual cumulative test cannot efficiently use such items.

---

[5]Paul E. Meehl and Albert Rosen, "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores," Psychological Bulletin, 52:194-216; May, 1955.

[6]Brogden, op. cit.; Lee J. Cronbach and Willard G. Warrington, "Efficiency of Multiple-Choice Tests as a Function of Spread of Item Difficulties," Psychometrika, 17:127-147, June, 1952;Frederick B. Davis, "The Selection of Test Items According to Difficulty Level," American Psychologist, 4:243, July, 1949; Harold Gulliksen, "The Relation of Item Difficulty and Inter-item Correlation to Test Variance and Reliability," Psychometrika, 10:79-91, June, 1945; Lloyd G. Humphreys, "The Normal Curve and the Attenuation Paradox in Test Theory," Psychological Bulletin, 53:472-476, November, 1956; D. N. Lawley, "On Problems Connected with Item Selection and Test Construction," Proceedings of the Royal Society of Edinburgh 61 (Section A, Part III):273-287, 1942-1943; Jane Loevinger, "The Attenuation Paradox in Test Theory," Psychological Bulletin, 51:493-504, September, 1954; Frederic M. Lord, "Some Perspectives on 'The Attenuation Paradox in Test Theory'," Psychological Bulletin, 52:505-510, November, 1955; Frederic

Control of the Score Distribution

The problem of score distribution is not only to assign a certain number of individuals to a given score, but to assign only like individuals to that score. The particular type of distribution which is desired depends upon the purpose for which the test is designed. A normal distribution is assumed in most statistical computations and interpretations. A rectangular distribution would give the best set of rankings in that people are spread evenly over all the scores. A bimodal distribution may be desired to classify individuals into accept or reject categories. Other than differences in the use of scores, factors which influence the score distribution are the distribution of ability levels of those taking the test, the item precision, and the difficulty of the items. A test able to produce any type of score distribution desired irrespective of the distribution of ability level of those taking the test and irrespective of the precision or difficulty of items available would have considerable utility.

M. Lord, A Theory of Test Scores; M. W. Richardson, "The Relation Between the Difficulty and the Differential Validity of a Test," Psychometrika, 1:33-49, June, 1936; Thelma G. Thurstone, "The Difficulty of a Test and Its Diagnostic Value," Journal of Educational Psychology, 23:335-343, May, 1932; Ledyard R. Tucker, op. cit.; and David A. Walker, "Answer-Pattern and Score-Scatter in Tests and Examinations," British Journal of Psychology, 30:248-260, January, 1940.

Meaning of a Score

The problem in assigning a meaning to a score is that the conventional cumulative score is typically a conglomeration which may represent the ability level of the individual, the rank of the individual, the pattern of response, or any combination of these. It is not possible to clearly represent the ability level of the individual with the usual cumulative test. While it is possible to just rank individuals or to just indicate the pattern of response with the cumulative test, this is not usually done. (In indicating the pattern of response the score is assigned to the sequence of items passed not to the number of items passed.) It may be useful to examine each of these possible elements in turn.

The ability level of the individual cannot be determined by knowing that he passed a difficult item in a cumulative test, because all people must take each item and difficult items are often passed by chance as the majority of the group must guess at these items. This clouds any interpretation of the number of correctly answered items as a measure of performance. To get a better measure of the ability level of the individual from the score, White and Saltz have argued that the items should be scaled as to difficulty so that one knows which set of items a person has answered correctly if he knows the total number answered

correctly.[7] The usual cumulative test score does not permit one to infer which items the individual has passed. The score in the type of test suggested by White and Saltz would probably be used to represent the level of subject matter learned rather than how the individual ranked with others. In addition to the infrequent use of the above solution, the suggestion does not solve the problem of the majority of individuals guessing the answer to difficult items.

To rank individuals in a normal distribution of ability so they are spread evenly throughout the score range, the test must make finer discriminations of ability at the middle ability range than it does for the extremes. Thus the test designed to rank individuals does not have a score scale which has the same relationship to the ability scale at the middle as at the extremes. Rarely is this relationship of scores to ability level reported. The cumulative test often compromises between using scores which rank individuals best and scores which tend to be normally distributed (as assumed in many statistical computations). The cumulative test may do either of the above alternatives well, but the decision made should be explicit and communicated to the test user. The decision should be to use the test score which permits one to infer rank (if this is what is desired), not

---

[7]Benjamin W. White and Eli Saltz, "Measurement of Reproducibility," Psychological Bulletin, 54:81-99, March, 1957.

to contaminate

into a distrib

efficient wit

Anothe

response.

concerned wit

Once the "i

been discusse

content" mea

of informatio

the mean in

the individu

ability. The

content) t

the second

we would

respect t

respect t

that to

of the

there are

correct

measure

with per

individ

specifi

to contaminate the meaning of a score by forcing the scores into a distribution just to create a higher correlation coefficient with normally distributed measures.

Another use of a score is to indicate the pattern of response. Cronbach has concluded that one should be as concerned with heterogeneity in content as in difficulty. Since the "level of difficulty" meaning for a score has been discussed above, the "heterogeneity with respect to content" meaning is considered here. For example, one bit of information is given when an individual is placed above the mean in pitch discrimination. With another set of items, the individual might be placed relative to the mean in visual acuity. The two items (with heterogeneity with respect to content) together place him in one of four categories. (If the second item had been a further measure of pitch, then he would have been placed in one of three categories with respect to pitch). The use of items with heterogeneity in respect to content thus seems useful, but one must remember that to recover all four categories the test cannot be scored by the number correct. Too often the items in cumulative tests are heterogeneous with respect to content and the number correct is used for the score. This cumulative scoring procedure permits the precise meaning of a score from a test with perfectly precise items to be inferred only when the individual possesses all of the characteristics above the specified levels or possesses none of the characteristics at

or above the spe

even more diffi

perfectly precis

Rarely is a

correct used, an

which is desired i

teristics, the p

ability are adde

teristics the in

III. PA

The sequen

this model is exp

of items, (2) co

a score with a p

for using one of

presented.

Maximally Effici

The sequen

of all items, i

model provided t

of difficulty fi

beginning stage in

progressively mo

abilities of iter

or above the specified level. These cumulative scores are even more difficult to interpret when the items are not perfectly precise.

Rarely is any method of scoring other than the number correct used, and, if the level of ability in any characteristic is desired in conjunction with the pattern of characteristics, the problems discussed above for reflection of ability are added to lack of knowledge about which characteristics the individual possesses.

III. RATIONALE FOR THE SEQUENTIAL ITEM MODEL

The sequential item model is now examined to show why this model is expected to (1) give maximally efficient use of items, (2) control the score distribution, and (3) yield a score with a precise meaning. In addition, the rationale for using one of the several sequential procedures is presented.

Maximally Efficient Use of Items

The sequential test is expected to make optimal use of all items, irrespective of difficulty, because this test model provides that each item be at the fifty per cent level of difficulty for the group taking the item. At each succeeding stage in testing the original group is divided into progressively more homogeneous ability groups and the difficulties of items are matched to the average abilities of

each group th

by the lowest

highest abili

This p

ham, Davis,

Richardson,

that if the

two groups,

of difficult

is at the po

means that

within high

within low a

the difficul

of the group

The se

will operate

ferent abili

the number

to chance w

reducing the

out by Meen

discriminating

ating power

5See

each group taking the item. Thus the easiest items are taken by the lowest ability groups and the hardest items by the highest ability ones.

This procedure accords with the works of Brogden, Cronbach, Davis, Gulliksen, Humphreys, Lawley, Loevinger, Lord, Richardson, Thurstone, Tucker, and Walker which indicate that if one wishes maximum discrimination of a group into two groups, then all items should be at the 50 per cent level of difficulty for a hypothetical group the median of which is at the point where the discrimination is desired.[8] This means that one needs difficult items to best discriminate within high ability groups and easy items to discriminate within low ability groups. The sequential procedure allows the difficulty of the item to be suited to the ability level of the group answering the item.

The second reason for assuming that the sequential test will operate better than a cumulative test is that since different ability level individuals do not take the same items, the number of low ability people passing a difficult item by chance will not exceed the number of high ability people passing the item due to their ability. As has been pointed out by Meehl, in the cumulative test an item with poor discriminating power is better than one with greater discriminating power if fifty per cent of the people are expected to

---

[8]See footnote 6.

pass the fi

<u>Control of</u>

The p

assign like

distributi

Since the di

which of a

difficulty a

have each st

and with it

cumulative p

Humphreys m

item diffic

a spread of

be sequenti

assumed that

until the

all for only

individuals

are used to

may be

this additio

of subgroups

[illegible]

Lord, [illegible]
Snedden, Ch. [illegible]

[illegible]

pass the first, and only 10 per cent to pass the second.[9]

Control of the Score Distribution

The problem of control of score distribution is to assign like people the same score, and to yield a score distribution which will best serve the purpose of the test. Since the distribution of scores depends upon the distribution of ability of those taking the test and upon the difficulty and precision of the items, Lord and Brogden have each stated that for a normal distribution of ability and with items of equal difficulty and usual precision, the cumulative test cannot produce normally distributed scores.[10] Humphreys has suggested that the answer is to spread the item difficulties.[11] He gives no method to show how such a spread of difficulties is determined. Another answer is the sequential process developed in this paper. It is assumed that the sequential procedure will more adequately control the score distribution because the items must operate well for only a small group of people not for all of the individuals taking the examination. After precise items are used to validly split a given group, the resulting groups may be further divided into whatever size is desired by using additional items of appropriate difficulty. Any number of subgroups may be combined if desired to produce appropriate

---

[9]Meehl and Rosen, op. cit.

[10]Lord, A Theory of Test Scores, op. cit., p. 11; and Brogden, op. cit., p. 207.

[11]Humphreys, op. cit.

distributi

of control s

tation.

Ranking of

A seq

of the indiv

of response,

these at th

is represent

of the fina

by the rank

is an equal

films are

difficulty

ability lev

at differen

different

individual

sequence o

every indiv

sence of

assigned a

noted the

different

this pract

ual.

distributions or to combine like individuals. These methods of control should allow maximum control of the score distribution.

Meaning of a Score

A sequential test score may represent the ability level of the individual, the rank of the individual, or the pattern of response, but it does not represent more than one of these at the same time. The ability level of the individual is represented by the score when the score is the difficulty of the final item. The rank of the individual is represented by the rank of difficulty of the final item. (The rank scale is an equal interval scale on ability when equal discriminations are made at all ability levels--in this case rank of difficulty and difficulty represent the same factor--the ability level of the individual. If unequal discriminations at different ability levels are made the scales represent different information.) The pattern of response of the individual would be represented by a score assigned to the sequence of items taken in the sequential test. Even though every individual may pass the same number of items, the sequence of items taken by an individual may be specified and assigned a score different from that of an individual who passed the same number of items but via a different route. Different routes (sequences) will represent different items being passed even though the number of items passed is identical.

Since

each yieldin

sequential

test score

scoring pro

the differen

the use of

<u>Selection o</u>

The ty

purpose of

ability indi

accuracy fo

in the time

made to do

for acceptin

other decisi

The

it as eithe

object, and

categories

the more th

into the th

developed b

and the nu

random samp

Since the sequential test has several scoring procedures each yielding a different but precise score meaning, the sequential score is more interpretable than the cumulative test score which is typically a conglomerate of all of these scoring procedures. In addition to the precision of meaning, the different scoring procedures allow great versatility in the use of the test.

Selection of the Sequential Procedure

The type of sequential procedure used depends upon the purpose of the test: (1) rapid classification of extreme ability individuals, (2) reaching a prescribed degree of accuracy for each score, or (3) doing the best job possible in the time allowed. In the present case the decision was made to do the best possible job with six items. The reasons for accepting this decision and the reasons for rejecting the other decisions are outlined briefly.

The rapid classification of individuals may be thought of as either classification into such categories as accept, reject, and continue testing--or classification into score categories which would more closely represent the results of the more traditional scoring procedures. The classification into the three categories closely resembles the procedure developed by Wald for industry where the concern was to predict the number of faulty objects in the population. A random sample of the population was used at each stage.

In the

me the set

her (e.g.,

often, the

rejected

that

qualified

individual,

Fiske

ature as

ered the

which can be

grounds.[13]

To clas

ified the

proceeded to

attempt

from the

ded with

ately with

The

classify

--that

In the Wald procedure two sets of values are computed: the one set is such that after each sample if results are lower (e.g., in number of correct items) than a specified number, then one may classify the population (or individual) as rejected with probability; and the other set of values such that after each sample if results are higher than a specified number, then one may classify the population (or individual) as accepted with probability.[12]

Fiske and Jones have advocated that the sequential procedure as outlined by Wald be used only when the problem involves the choice between two possible parameter values which can be specified on a priori, but not arbitrary grounds.[13]

To classify people into additional categories, Cowden modified the Wald procedure. He assumed that the fewer items one needed to meet the criteria for classification into either the accept or reject categories, the farther the individual was from the specified level. He thus created five categories with the extreme categories being classified very rapidly with few items.

The second sequential procedure suggested above--that is, classifying until a specific degree of accuracy has been reached--has not yet been investigated. Exploration of this

---

[12]Abraham Wald, Sequential Analysis (New York: John Wiley and Sons, 1947).

[13]Donald W. Fiske and Lyle V. Jones, "Sequential Analysis in Psychological Research," Psychological Bulletin, 51:264-275, May, 1954.

procedure w

might be no

plate under

variables in

this paper.

Where

the model a

this is not

about the a

tion situati

ability the

individual

sequential p

of the indi

attempt to

gories as

of mathemati

the apparent

fruitful at

not seem fr

this,the pro

then can be

to be that

which would

situation.

procedure was rejected because it was felt that this procedure might be more fruitfully explored after there was more adequate understanding of the interrelationships of the variables involved in the sequential procedure developed in this paper.

Whereas in the industrial system of sequential testing the model assumes a random sample of ability at each level, this is not the best procedure for obtaining information about the ability level of an individual. Except in selection situations, the purpose is to determine the level of ability the individual possesses rather than whether the individual is above or below a given ability level. In the sequential procedure developed in this paper, a random sample of the individual's behavior is not used; there is rather an attempt to classify individuals into as many ability categories as can be adequately differentiated. There has been no mathematical model developed for the above procedure and the apparent alternative of developing one did not seem fruitful at this time. An empirical study of the problem did not seem fruitful because neither the ability level of individuals,the precision of the items nor the difficulty of the item can be determined exactly. The best alternative seemed to be that of creating exact data and then creating a model which would use this data in a manner resembling the actual situation.

Preliminary

used a probability

with actual data

electronic comput

of the computer p

six items and per

possible where it

The problem

to the type of de

the investigation

according to the

any variable, a

those taking the

these problems.

From the ra

one can deduce th

efficiency, contr

produced. The ra

variables when us

items at the fifty

was used with the

this statement is

1...Unpublished
National Research
Michigan, 1956-195

Preliminary work with the sequential procedure had used a probability model that had been empirically checked with actual data and which had been programmed for the electronic computer.[14] It was thus decided to take advantage of the computer program for this study. The program used six items and permitted calculation for any sequence possible where items were used to make dichotomous decisions.

IV. HYPOTHESES

The problems of testing are best described according to the type of decisions that need to be made; however, the investigation of these problems is best classified according to the variables that are changed. Changes in any variable, such as the type of ability distribution of those taking the examination, may affect one or more of these problems.

From the rationale developed in the previous section, one can deduce the effects these variables should have on efficiency, control of score distribution, and type of score produced. The rationale will explain the effect of the variables when used with the six-item cumulative with all items at the fifty per cent level of difficulty as well as when used with the sequential model. The one exception to this statement is that Lawley's work would indicate that

---

[14]Unpublished material developed in the Bureau of Educational Research, Michigan State University, East Lansing, Michigan, 1956-1959.

precise scores (scores which have small variance of ability level for individuals assigned the score) are created for only a single group by using items quite removed from the ability level of those individuals whom one wishes to precisely classify. For example, if we wished to have the extreme scores precisely defined then we would use items at the fifty per cent level of difficulty. The hypotheses on precision of score are derived from the above conclusion of Lawley. The score distribution examined in this study is the one actually produced although it is clear that scores could be combined to yield shapes of distributions different from the one initially produced. The score meaning that is examined here is that of reflection of the criterion ability scale.

The general hypotheses arising out of the rationale will be described here. The operational hypotheses that are tested are stated in Chapter III. There are (1) a set of hypotheses concerned with the effect of the type of ability distribution on both the six-item cumulative model and the six-item sequential test model; (2) a set of hypotheses concerned with the effect of precision and difficulty on the output distribution of the sequential test model; and (3) a set of hypotheses concerned with the effect of the errors in estimating the parameter values on the output.

Effect of

The ef

efficient

variance of

level, or b

to a score.

used to des

assigned di

be used to

ate of appr

of determin

to determin

involves de

and meaning

people is a

individuals

tion is not

As the

signed to di

the effici

separation o

actual abili

will be show

the cumulati

the extreme

the ability

Effect of the Type of Ability Distribution

The effect of type of ability distribution on maximally efficient use of items may be examined by determining the variance of scores which are assigned to a given ability level, or by examining the variance of ability levels assigned to a score. "Discrimination among ability levels" shall be used to designate whether different ability levels are assigned different scores, and "precision of scores" shall be used to indicate whether all individuals at that score are of approximately the same ability level. Another method of determining the effect of type of ability distribution is to determine discrimination among people. (This procedure involves decisions as to both control of score distribution and meaning of the score produced.) Discrimination among people is a measure of the ability of the test to rank individuals according to ability. This type of discrimination is not considered in the following hypotheses.

As the sequential test being considered here is one designed to discriminate among ability levels, it should work quite efficiently for all distributions with respect to the separation of the ability levels and the reflection of the actual ability distribution in the score distribution. As will be shown in Chapter II in the review of Lawley's work, the cumulative test should have a greater precision of scores for extreme scores, but should be equal to the sequential in its ability to accurately discriminate among the ability

levels of individuals only at the middle ability levels. These expectations are examined under conditions where two different distributions are input--normal and U-shaped.

Normal distribution.--(1) The cumulative and sequential test models should have equal ability to classify individuals of mean ability level. This hypothesis follows from the fact that middle ability people will take 50 per cent level of difficulty items in the cumulative test, and should take items near the 50 per cent level of difficulty in the sequential test. If the sequential does not operate efficiently, the cumulative test will have the more discriminating scores.

(2) The sequential test model should more accurately classify the individuals at the extremes of the ability scale than should the cumulative test model. This is based upon the rationale that the sequential test can use difficult items because it discriminates among high ability individuals (as these items are at the 50 per cent level of difficulty for these high ability individuals). The test item does not have to discriminate between low and high ability individuals as only high ability individuals will take the item.

(3) The cumulative test model should have more precise scores at the extremes of ability than the sequential test model. This follows from the work of Lawley which showed that the variance of ability levels for individuals assigned to high scores would be low if the items were easy for these individuals.

(4)

represent t

extremes wi

reflect the

among abili

ent level

median abil

ation is de

ications on

whereas, in

quite equal

U-sha

should more

II (see "Is

ent model.

ication be

the mean an

individuals

extreme. Da

individuals

The reason t

these people

level of abi

used in the

(2) In

classify the

(4) The scores for the cumulative test model should represent finer ability units in the middle than at the extremes while the sequential test model scores should reflect the ability level scale. The best discriminations among ability levels should be made by using items at 50 per cent level of difficulty for the hypothetical group the median ability of which is at the point where the discrimination is desired. For the cumulative test the best discriminations should be at the 50 per cent level of ability; whereas, in the sequential test items should discriminate quite equally over the entire range of ability.

U-shaped distribution.--(1) The sequential test model should more accurately classify the individuals of catetory 13 (see "Ideal T Score" in Table 24) than the cumulative test model. Category 13 individuals are the focus of consideration because in a U-shaped distribution few people are at the mean and the question becomes how well one can classify individuals who exist in larger number and are not at the extreme. Category 13 represents this mean value for those individuals in the upper half of the distribution of ability. The reason that the sequential should more accurately classify these people is that the items are more appropriate for their level of ability than 50 per cent level of difficulty items used in the cumulative.

(2) The sequential test model should more accurately classify the individuals at the extremes of the ability

distribution than the cumulative test model. The reason for these expected results is again that items are more appropriate for the individuals, and individuals taking the items have a smaller variance in ability than those taking the cumulative items.

(3) The cumulative test model should have more precise scores at the extremes than the sequential test model. Again this follows from Lawley's work.

(4) The sequential test model should have equal score discriminations for all groups including the mean group, whereas the cumulative test model should have finer score discriminations for middle ability levels than for the extreme ability group. This follows from the wide distribution of item difficulties used in the sequential as compared to the cumulative tests. Items discriminate best only at once ability level and should be used only with individuals close to that ability level.

Effect of Item Precision and Difficulty for the Sequential Test

The relationship of item precision and difficulty to output characteristics must be examined together as change in precision results in change of the appropriate difficulty levels in the manner described in Chapter III. There are five levels of precision used: r_{bis} = .79, .75, .71, .60, and .45. Since the ability distribution also effects score distribution, a normal distribution of ability is used as this is the type of distribution most likely to occur in the

practical situation.

(1) The variance of scores for a given ability level should be less with the test using the most precise items. The value for the precision of an item indicates how effectively the item differentiates individuals of one ability from those in the next closest ability level. If the item is precise then each item can make a different distinction in ability rather than more accurately making the distinction that should have been made by a prior item.

(2) The test consisting of the most precise items should have more equal discrimination between adjacent ability levels than will the less precise test. If the ability of an item to discriminate among ability levels is dependent upon the difficulty level of the item, then the more precise test which has a wider range of difficulties should discriminate at all levels while the less precise test which has a smaller range of difficulties should discriminate well among middle ability individuals where difficulties are appropriate. The less precise test should not discriminate as well among extreme ability individuals where difficulties are not as appropriate.

Effect of Errors in Estimating Parameters

The usefulness of the model for practical purposes depends upon the sensitivity of the test design to the use of an item which only approximates the precision and difficulty level which would be called for by the "ideal" model.

If the values need not be very accurately determined before use can be made of the sequential test model, one is more likely to use the model. Preliminary studies have indicated that the sequential test will probably be more sensitive to precision estimates than to difficulty estimates. The effect of errors of parameter estimates is the same effect as is involved in the use of items which have parameter values other than those required by the test.

As is noted in Chapter III, Section 1, each succeeding item in a sequential test is selected in such a way as to maximize discrimination based on data from the effects of previous items. The effect of using a more precise item than called for should be that the next item would not be difficult enough or easy enough for maximum discrimination. The effect of using an item too easy should be to increase the precision of score for the upper group, but to decrease the discrimination among ability levels.

Since the effect of errors made in early stages is either corrected or magnified by the effect of later items, and since the effect of errors made in later stages has no chance to be corrected or magnified, one would expect differences in the effect of errors at early and late stages. The hypotheses made as to effect of errors at these different stages are as follows:

(1) Errors in difficulty at an early stage should not have any serious effects as there would be a wide range of

ability and the item would operate well for some of that range.

(2) Errors in difficulty at the final stages should increase the variance of ability levels assigned to one of the two subgroups into which the total group would be separated, but should not lower the variance of scores assigned to the ability levels.

(3) Errors in estimates of the precision of the item should be more serious in the initial stages where wide separations in difficulty level of the next item would be used.

(4) Errors in the estimates of the precision of the items should make little difference at the final stages as the next item would be appropriate.

If the sequential testing procedure is robust in that errors in estimating parameters do not seem to greatly effect type of output, then it would be possible to design the test with parameter values determined from one sample of a population and use this same test in different situations. (The value used for the precision of the item is dependent upon the spread of ability in the sample used to determine the precision value. If the spread of ability is great in contrast to item sensitivity, one has a precise item. If the spread of ability is narrowed, the same item would be considered a less precise item.)

V. LIMITATIONS OF THE STUDY

The three major contributions of this study are that it: (1) discusses the problems of the cumulative test and shows how the sequential model attempts a solution to each of these; (2) provides a model that may be used in construction of any sequential test; and (3) presents a rationale for the sequential test model which, when tested, should allow the construction of additional sequential tests. There are, however, many problems that are not examined. Six of these are listed and discussed because the background material gives suggestions as to the probable answers to these problems also. These are: (1) the best possible cumulative test, (2) the score distributions desired for the cumulative and sequential models, (3) the types of ability distributions that may be present in the usual situation, (4) likely test parameters for usual test items, (5) commercial test construction procedures, and (6) test presentation procedures and the psychological effects of the sequential model.

Best Cumulative Test

The work of Brogden and Humphreys indicates that the best cumulative test with precise items is one with a spread of difficulties.[15] The exact relationship between spread of

---

[15]Brogden, op. cit.; and Humphreys, op. cit.

difficulties and precision to yield maximum validity (measured by correlation with ability distribution) is not known, but Cronbach and Warrington indicate that for a cumulative test of a given length, $\sigma_y^2 + \sigma_d^2$ will have a preferred value.[16] (The term σ_y is the standard deviation of the spread of item difficulties and σ_d is the measure of precision which is the same as the one used in this paper.)

The sequential test models are not compared to the best possible cumulative model, but the use of items all at the 50 per cent level of difficulty creates a test that is more than sufficient for most uses for most levels of precision.[17] The purpose of the cumulative test model in this dissertation is to put the sequential test model material into perspective.

Distribution of Scores

If the purpose of testing is selection, then a test need only produce two scores, one for the individual who is selected and the other for the one rejected. In this situation the sequential model developed here would require modification both in method of scoring and in number of items taken by individuals. The previously discussed sequential model developed by Wald, involving a variable number of items taken by

---

[16]Cronbach and Warrington, op. cit.

[17]Ibid.

individuals, is probably the optimal solution. The problem of test construction thus is no longer that of determining the difficulty of the item, but rather the number of items needed to make the most rapid classification. There is no score distribution as such, only accept, reject and continue testing categories of individuals.

The cumulative test used to differentiate two groups would be one with all the items at the level of difficulty appropriate for the ability level at which one wishes to make the decision. A test of this nature would have a score distribution which would be platykurtic, rectangular, or bimodal depending upon the precision of the items in the test. The test with most precise items would have a bimodal score distribution.

If one desired to rank individuals by the scores from the test, one would make fine discriminations in ability for those ability levels where there were many people. In this way the individuals would be assigned scores which would be rectangularly distributed. This can be accomplished by use of a cumulative test which has either fairly precise items at the 50 per cent level of difficulty or a spread of item difficulties for less precise items. For the sequential test, there would be more items included at the difficulty level appropriate for the discriminations that are desired.

The construction of either a sequential or a cumulative test which has the score distribution discussed above is

outside the scope of this dissertation. Further research is needed to determine the items for a sequential test which would have a rectangular distribution with the input of a normal ability distribution.

Ability Distributions

Lord has stated that perhaps test constructors should not consider ability as normally distributed.[19] It is possible that a bimodal distribution of ability is common in that there are many individuals who perform adequately and many individuals who perform inadequately with a large gap between these two performance groups. If this is true, the sequential test model should operate well for these distributions, as it should operate well with any type of distribution. Abberations in its operation would show up most clearly when the test model is tested against a U-shaped distribution of ability. In Chapter IV the results are reported for testing the model against the U-shaped and normal distributions. These results indicate how the sequential test scores may be interpreted when used with different ability distributions. However, no rationale is developed to indicate what the results should be and, therefore, the interpretation of scores across ability levels depends upon a rationale developed post facto, not upon the rationale tested in the study.

---

[19]Lord, A Theory of Test Scores, op. cit.

Test Parameters

The effect of the number of items has not been examined. The six-item test was used because the probability model for the test had been programmed for the electronic computer and six-items were the maximum for this program. Further research is needed to determine how rapidly the output characteristic changes (if at all) when the test consists of more items.

Test Construction Procedures

The computational model described in Chapter III for the construction of a sequential test has a method of selecting items with the best possible parameter values. This method could be used in the construction of a sequential test with the data in terms of difficulty and precision taken from actual items. The criterion may be a measure of the number of individuals desired to pass the item or a measure of the variance of ability levels of individuals assigned to the pass and fail categories.

It would seem reasonable that one should use the most precise items to differentiate the individuals as to ability level and then the difficulty of a less precise item could be used to control the number of individuals assigned to any one score category. The second differentiation would not be as valid as the one made with the more precise item, but the shape of the distribution could be well controlled.

In addition to lack of a complete evaluation of the score distribution control procedure, there has been no attempt

to follow the standard criteria such as that published by the Committee on Test Standards of the American Educational Research Association.[20] These criteria include content validity, concurrent validity, predictive validity, construct validity, error of measurement at different score levels, equivalence of forms reliability, internal consistency reliability, stability reliability, and information on norms and scales.

Since this dissertation uses hypothetical data, content validity is not considered. It is assumed that the test items are homogeneous and thus measure only one content or ability which may or may not be a composite of several abilities.

The six-item sequential is compared with the six-item cumulative but no correlation is computed between the two sets of scores, as is common in concurrent validity studies. In this type of a model one can probably obtain more information from the correlation with a known criterion score than from correlation between sequential and cumulative test scores.

The predictive validity of the test is not determined as it made no sense to use hypothetical data to predict hypothetical performance. Predictive validity needs to be

---

[20]American Educational Research Association, Committee on Test Standards, and National Council on Measurements Used in Education, Committee on Test Standards, Technical Recommendations for Achievement Tests, 1955.

studied through the construction of a sequential test with actual items, testing of a group, and then the prediction of future performance. This would be a logical next step if the model data studied here show that the sequential item test is a better test than a six-item cumulative under the conditions of this study. If sequential test does not have results which may be considered better than the results from the cumulative test, then there is no need to study the sequential under less favorable conditions.

In construct validity it is assumed that the characteristics measured and related are not affected by the type of items used in the test. Results from this study may be used to indicate that these assumptions are not met in most situations. A study of the attenuation paradox literature should make one aware of the problems involved in the measurement of characteristics and their relationships. There is no attempt to evaluate the construct validity of the sequential test. Neither is there any attempt made to correlate test scores with other abilities that should be related to the particular hypothetical ability being measured. That which is measured is any homogeneous ability measured by the items with the given level of precision--all of the items in the sequential model have the same precision.

Error of measurement at different score levels is examined in detail as suggested by the criteria for evaluation of a test. The discriminating power of the test at a given

level of test score is to be distinguished from the discriminating power at a given level of ability. Both the variance of the test scores of each ability level, and the variance of the ability levels at each score are examined.

The equivalence of forms reliability is not determined as there is only one form. It would be quite simple to build two tests in a computer and determine how well the scores on the one test could be predicted from the scores on the other. It is possible that quite equivalent tests could be built from quite different items. This possibility is not examined in this dissertation.

Due to the hypothetical nature of the data the internal consistency reliability is not examined. Stability reliability is not determined as it would be necessary to administer a test twice to a group to determine this, and no test is actually used in this paper. This is another area that needs to be examined.

There is a fairly complete discussion of the score distribution of the sequential item test. It is hoped that the rationale which predicted the type of score distribution would be proved correct and thus a tested rationale would be presented rather than a rationale derived from the results.

Norms (like many of the criteria used to evaluate a test rather than a test procedure) are irrelevant to the test procedure.

Another limitation to the study is that no attempt is made to examine the effects of errors of estimating the parameter values when the level of precision is low. However, one would suppose that the effect of errors will be less at lower precision levels. If the effects produced at high levels of item precision are within the error range for practical significance, then there is little need to examine the effects at low level of item precision. If the effects at high levels of item precision are beyond the error allowed for practical significance, then one must determine the effects of lower item precisions or develop methods of obtaining better estimates. This decision can be made later.

Test Presentation Procedures and Effects

In the area of sequential test presentation to the testee little is known as to how to proceed in actual practice. For example, it may be psychologically advantageous to give the easiest items first, allowing some individuals to subsequently try more difficult items, rather than to have everyone start at an item of 50 per cent difficulty. Since the test is not given to an actual group this procedure cannot be examined in this dissertation.

The greater the number of score categories that one wishes to use, the more cumbersome is the presentation of the items. Some of the teaching machine methods ofpresentation may prove to be useful if one wishes to use a large number of score

categories and a large number of items for each individual. A complete exploration of methods of presenting the test should be considered once the advantages prove to be great enough to warrant such an exploration.

VI. OVERVIEW OF THE REMAINDER OF THE DISSERTATION

Up to this point an attempt has been made: (1) to describe the physical characteristics of the sequential test model illustrated in Figure 1; (2) to present the problems which suggest a sequential test model; (3) to outline the decisions made in regard to which problems were to be investigated, and to delineate them from alternatives. The second chapter, "Review of Literature," will report material used in arriving at the decisions made and reported in Chapter I. Chapter III describes the actual procedures used in the construction of the sequential test, the operational hypotheses tested in this study, and procedures used to test each of the three major hypotheses. Chapter IV gives the analyses and results of the procedures used to test the three major hypotheses. Chapter V offers the conclusions reached by the author as to the questions raised in the three major hypotheses and in relation to the general problems of testing raised in Chapter I. Chapter VI gives the summary and recommendations for further study.

CHAPTER II

REVIEW OF LITERATURE

The literature relevant to the study of sequential testing is reported in four sections related to: (1) maximally efficient use of items, (2) control of score distribution, (3) meaning and use of scores produced, and (4) sequential testing procedures. This is the same organization as that used for the Rationale in Chapter I. The decision as to organization is made with cognizance of the fact that research is used to study the effects of variables as well as which variables are related with certain effects. In the review of literature, the data from studies involving research about effects of variables are placed in the section where the major effect was noted and mention is made of other related effects even though these effects may be more closely tied to problems considered in another section.

The interrelationships among the test item parameters of difficulty and precision, validity, reliability, and score distributions had not been extensively explored for the sequential test. However, many of these relationships have been studied for the cumulative test and yield data which are relevant to sequential testing procedures. This lack of exploration can not be due to the length of time that the procedure has been

available for study, for L. L. Thurstone advocated some of the notions of the sequential testing procedure as early as 1926, and Binet, before this time.

However, whatever the explanation for the lack of study of the sequential procedure, the problem at hand is the evaluation of the cumulative test literature which is relevant to the sequential test procedure.

I. MAXIMALLY EFFICIENT USE OF ITEMS SELECTED

As has been noted in Chapter I, one of the problems of efficiency is that of using the most precise set of items selected from a pool of items which has a range of precision and a range of difficulty. The usual cumulative test cannot efficiently use a difficult item even if it were precise.

Tucker stated that a test with imperfect items gives a better measurement of the scale of ability than a test with perfect items if the score is the number correct. He reported the amazing fact that low-value item intercorrelations yielded the best measurement under cumulative test procedures.[1] These item reliabilities for maximum test scores vs. ability correlations were within the range of practical experience. In fact, for an n of 10 the maximum validity came from intercorrelations of .50.

---

[1]Ledyard R. Tucker, "Maximum Validity of a Test with Equivalent Items," Psychometrika, 11:1-14, No. 1, March, 1946, p. 11.

It has usually been believed that increasing the reliability or item precision of the test always increases its validity. However, as Gulliksen has pointed out, increasing the reliability of a test beyond a certain point will, under certain conditions, decrease the validity.[2] The literature in regard to this "attenuation paradox" is reviewed here, as these articles indicate reasons why reliability and validity have not shown a monotonic relationship.

To explain this paradox, one may question any of the assumptions that are conventionally made in test construction and analysis. One may question the measurement of validity, the scales produced for the criterion and their comparison with the criterion, the assumption of normally distributed populations, the approximations used to measure data, and the basis upon which the test was scored.

Gulliksen--who seems to have been the first person to point out the paradox--argued that his formulas indicated that if all the items were concentrated at one difficulty level then the test reliability could be higher than is possible if the items cover a rather wide difficulty range.[3] But instead of arguing that items should be at one level of difficulty he argued that items of graded difficulty should

---

[2]Harold Gulliksen, "The Relation of Item Difficulty and Inter-item Correlation to Test Variance and Reliability," Psychometrika, 10:79-91, No. 2, June, 1945.

[3]Ibid.

be used, and that the error was in the scoring procedure of counting one point for each item correct. According to Gulliksen, the score assigned should be the best estimate of the difficulty level reached. It is true that a change in scoring procedure may be a solution, but before any general solution can be reached about the relationship between item reliability and test validity many assumptions must be checked.

Humphreys stated that the supposed lack of monotonic relationship between reliability and validity was due to normal curve and interval data assumptions.[4] To explore the effect of the assumption of normally distributed abilities, Humphreys determined the r_{pbis} in two ways: one computation involved the assumption of normality; the other computation did not involve this assumption. He found that $r_{pbis} = \sqrt{\phi}$ did not result in values for validity that were the same as those obtained from $r_{pbis} = [\sqrt{r_{tet}}] \cdot [z/(pq)^{1/2}]$. Using the usual point biserial correlation, Humphreys also found a monotonic relationship between reliability and validity for all difficulty levels. He concluded that the "assumption of a normal distribution of the criterion is not compatible with the mechanics of adding items together."[5]

---

[4]Lloyd G. Humphreys, "The Normal Curve and the Attenuation Paradox in Test Theory," Psychological Bulletin, 53:472-476, No. 6, November, 1956, p. 473.

[5]Ibid., p. 474.

Humphreys thus questions both the assumptions of scoring and the normal distribution of the criterion. The problem of determining the maximally efficient use of an item thus becomes more difficult because there is no agreement as to how efficiency can be measured.

The problem is further confounded by the knowledge that even if the criterion were normally distributed one has to measure the relationship between test scores and the criterion. This relationship is usually specified from the slope of a straight line, but both Brogden and Lord stated that since the regression curve must be the sum of the item characteristic curves, it is inevitably curvilinear and, in particular, will be strongly curved if the items are all of equal difficulty and have high intercorrelations.[6] Thus if ability is normally distributed, the test scores cannot be normally distributed. Lord stated that with progressive increase in the item intercorrelations, the progressive decrease in the product-moment correlation between test score and ability is due in part to the fact that as the item intercorrelations increase, the regression becomes more and more curvilinear.[7]

---

[6]Hubert E. Brogden, "Variation in Test Validity with Variation in the Distribution of Item Difficulties, Number of Items, and Degree of Their Intercorrelation," Psychometrika, 11:197-214, No. 4, December, 1946, p. 207; and Frederic M. Lord, A Theory of Test Scores, op. cit., p. 11.

[7]Lord, A Theory of Test Scores, op. cit., p. 19.

However, in a later article, Lord decided that the attenuation paradox was not due entirely to the violation of linear regression as he had assumed earlier.[8] His later work showed that the problem was even more complex in that it was not intuitively valid to demand a nonparadoxical relationship due to the hodgepodge we now have in reliability. Lord concluded that there is a serious lack of homoscedasticity, for when item intercorrelations are high the standard error of measurement is very different for test scores at different ability levels.[9]

Thus the efficiency of an item may depend upon the scoring procedure which changes the distribution of scores, or the type of measures used to describe the relationship and the assumptions used in the computation of these relationships. However, instead of explaining the paradox by the above characteristics, the possibility exists for explaining the paradox by the effects of each individual item.

Cronbach and Warrington gave the following explanation of the paradox:[10]

---

[8]Frederic M. Lord, "Some Perspectives on 'The Attenuation Paradox in Test Theory'," Psychological Bulletin, 52:505-510, No. 6, November, 1955, p. 506.

[9]Ibid., p. 507.

[10]Lee J. Cronbach and Willard G. Warrington, "Efficiency of Multiple-Choice Tests as a Function of Spread of Item Difficulties," Psychometrika, 17:127-147, No. 2, June, 1952, p. 139.

> If an item has perfect precision, it gives no information about which of the men whose criterion score is below y_i are best. All of these men will have the same score (zero) on a group of perfectly precise free-response items, if guessing is impossible. If each item allows two or more choices, the scores will vary but the differences will not be related to ability. Since the obtained scores are equal or differ only by chance, the test does not discriminate among low-ability men having different criterion scores. Likewise the peaked test gives no information about individual differences within the high-ability group, whose thresholds are above the scale position of the items. In a less precise item, the proportion passing is a sloping function of criterion score, and a man whose ability falls slightly below the scale position of the item will tend to earn a higher score than the man who is far below the scale position. Each item contributes information along the whole scale.

The "attenuation paradox" has many possible explanations but the examination of the solutions that are derived from these explanations may give more valuable information.

The solution advanced by Gulliksen--the first person to point out the paradox--of changing the scoring procedure does not seem to have been widely adopted.[11]

The impetus for changing the scoring procedure seems to come from those concerned with the meaning of the score, rather than from individuals interested in solving the attenuation paradox.

The most common solution to the attenuation paradox seems to be that of changing the idea of how to measure the "best test." This material is particularly relevant to this dissertation because some of the test evaluation ideas from this literature are used in the present paper in lieu of the more traditional techniques of test evaluation.

---

[11]Gulliksen, op. cit.

Lord has argued that a more basic concept than validity is that of the discriminating power of the test at various ability levels. Lord felt that the test constructor's goal should be to achieve a desired degree of discriminating power rather than to maximize any single composite validity coefficient. The conventional reliability and validity coefficients are indices of discrimination for the test as a whole; however, except under certain limited conditions, these overall indices do not apply at all points along the score scale.[12]

Levine and Lord examined the discriminating power of a test at different parts of the score range. They used as the measure of discrimination the ratio of the slope at a given x value (score measured in sigma units) to the standard deviation of the y scores (criterion measure) at that x value.[13] The lower the deviation of y scores and the greater the slope, the higher the discrimination index.

Levine and Lord stated that there is no precise standard error of the discrimination index known, but the expected value of the discrimination index for a homoscedastic linear scatterplot serves as a standard against which we may judge the computed values at various score points.[14] This value of the

---

[12]Lord, "Some Perspectives on 'The Attenuation Paradox in Test Theory'," op. cit., p. 506.

[13]Richard Levine and Frederic M. Lord, "An Index of the Discriminating Power of a Test at Different Parts of the Score Range," Educational and Psychological Measurement, 19:497-503, No. 4, Winter, 1959.

[14]Ibid., p. 502.

discrimination index for a homoscedastic linear scatterplot is computed from the following formula: $\frac{r_{xy}}{\sqrt{1 - r_{xy}^2}}$

Using a 25-item test as the x variable, and a 107 - item criterion test as the y variable, Lord determined the value of the discrimination index between 27 adjacent ability levels on the criterion test.[15] Values for the discrimination index did not indicate a lack of discrimination at the extremes or in either half of the ability distribution. For both extreme quartiles and for the upper and lower halves, Lord found nine discrimination indices below the expected value and 17 above the expected discrimination index value.

This literature only indicates a method of measuring the efficiency of a test and some methods of interpreting the data obtained. Since there are no data presented as to the spread of item difficulties, no conclusion can be drawn about the relationship between difficulty of item and the expected discrimination index.

Levine examined test validity in terms of the discrimination at different parts of the score range rather than by some measure of correlation between test scores and a criterion measure.[16] Again, the relationship between item reliability and this measure of validity was not studied. However, Levine suggested that if the reader is interested in discriminations at both extremes of the score scale, then two criterion tests

---

[15] Ibid., p. 501.

[16] Levine and Lord, op. cit.

of different difficulty would probably be needed. This indicates that if one wishes to discriminate among the high ability individuals of a group, one would use more items which would be considered difficult, while if one wishes to discriminate among the low ability individuals of a group, one would use more items which would be considered easy. Item difficulty can thus be related to discrimination as well as to the more traditional measures of validity. (This conclusion is further supported by the work of Lord reported in a later part of this section dealing with the decision to make discrimination by using items of the appropriate level of difficulty.)

It should be noted that the discriminating power of a test may have more than one aspect. Loevinger has suggested that it has three aspects--fineness, probability, and range.[17] Lord has made the following distinction between "discriminating power" and "effective discrimination":[18]

> A test may have low discriminating power for examinees in a certain range of ability. If in any given group of examinees there are only a few individuals spread out thinly over this range of ability, however, the rank order of these individuals on ability may be more accurately determined by the test scores than is the rank order of examinees in some other range of ability where the discriminating power of the test is greater, but where there are many examinees of almost identical ability. The effective discrimination is greatest where the rank order of the examinees is most accurately determined.

---

[17]Jane Loevinger, "The Attenuation Paradox in Test Theory," Psychological Bulletin, 51:493-504, No. 5, September, 1954.

[18]Lord, A Theory of Test Scores, op. cit., pp. 24-25.

Both discriminating power and effective discrimination are examined in this dissertation. (However, it should be remembered that the rationale for the test construction is based upon discriminating power.) The specific literature relevant to the distinction between discriminating power and effective discrimination is presented, respectively.

Davis contributed to knowledge of discriminating power by pointing out that tests may be constructed so that items can be selected according to difficulty in such a way as to control the standard error of measurement at different points on the ability scale.[19] However, Davis pointed out that little systematic work has been done to show analytically the relationship between the shape of the distribution of item difficulty indices and the size of the standard error of measurement at various levels of ability.

The summarizing statements made by Davis about controlling the magnitude of the standard error of measurement at various ability levels are as follows:[20]

> 1. To minimize the aggregate of errors of measurement of a test (thus perhaps sacrificing over-all validity or differential validity), all items should be of 50 percent difficulty. This would maximize the over-all test reliability coefficient and minimize the standard error of measurement at the center of the range of ability measured.
>
> 2. To minimize the standard error of measurement at any one point on a scale of ability, all items should be concentrated at that level of difficulty.

---

[19]Frederick B. Davis, "Item Analysis in Relation to Educational and Psychological Testing," Psychological Bulletin, 49:97-121, No. 2, March, 1952, p. 106.

[20]Ibid., p. 107.

3. To minimize the standard error of measurement at two or more points on a scale of ability, items should be apportioned to each of the levels of difficulty specified.

4. To minimize the standard error of measurement throughout a certain part of the range of ability, items should be distributed within the corresponding range of difficulty in accordance with the procedure suggested for obtaining maximum over-all test validity.

5. To equalize as nearly as possible the standard error of measurement throughout the range of ability measured, items should be distributed over the entire range of difficulty in accordance with procedure suggested for obtaining maximum over-all test validity.

These suggestions are useful, but the scale upon which the scores are based should be considered in the evaluation of item efficiency. Davis mentioned the scale upon which the scores were based when he stated that if the Mollenkopf data were reworked to express scores derived from his various tests as approximations to interval scores on a single scale, these date might show that the size of the standard error of measurement in a given range of ability decreases with an increase in the number of discriminations among examinees obtaining scores in that range of ability.[21] (The Mollenkopf data are presented in Section II, Control of Score Distribution.)

An even more dramatic consequence of the changing of scales was noted by Symonds. In his examination of the standard error of measurement, Symonds gave six groups on

---

[21] Ibid.

three tests with 24, 14, and 2 difficulty levels.[22] Results from these tests yielded means of 20.48 and 19.54 for the first test, 24.39 and 21.70 for the second test, and 26.74 and 32.42 for the third test, respectively. The variances of scores were 5.86 and 5.88, 9.10 and 9.22, and 11.12 and 11.82. The standard error of measurement was thus greatest for the third test which had a narrow range of difficulty--the standard errors of measurement were 2.33, 3.38, and 4.58. However, Symonds pointed out that as far as assigning the difficulty as a score was concerned, on the first test one score unit equalled .25 units on the Ayres scale of difficulty, for the second test one unit equalled .125 Ayers units, and for the third test one score unit equalled .02 units on the Ayers scale of difficulty. In terms of difficulty levels on the Ayres scale the standard errors of measurement were then .58, .42, and .09. Thus changing the scale reversed the order of the size of the standard errors of measurement.

When one evaluates a test using "effective discrimination" rather than "discriminating power," one examines the type of individuals assigned to a score rather than the distribution of scores assigned to an ability level.

---

[22]Percival M. Symonds, "Factors Influencing Test Reliability," Journal of Educational Psychology, 19:73-87, No. 2, February, 1928.

Lawley assumed that all items composing a given test were measuring the same ability x; and that the scale in which this ability was measured was so chosen that x was normally distributed over the whole population of individuals for whom the test was designed--with zero mean and unit variance.[23] With these assumptions Lawley described the error variance of a score.[24] When the mean difficulty of the items is not at the 50 per cent level of difficulty for the individual, the error variance of the score is defined as below:

$$\sigma^2_{E(X)} = n\,(t_o - t_o^2 - t_1^2\rho_1 - t_2^2\rho_1^2 - t_3^2\rho^3 - \ldots)$$

When the mean difficulty of items is at the 50 per cent level of difficulties for the individual then the error variance of the score is defined as below:

$$\sigma^2_{E(X)} = \frac{n}{2\pi}\cos^{-1}\rho_1$$

The terms are defined as follows:

$\sigma^2_{E(X)}$ = error variance of score

n = number of items

X = score value

---

[23]D. N. Lawley, "On Problems Connected with Item Selection and Test Construction," Proceedings of the Royal Society of Edinburgh, 61 (Section A, Part III):273-287, 1942-1943, p. 273.

[24]Ibid., p. 279.

t_o, t_1, t_2, etc. = values from Table 29 of Pearson's Tables for Statisticians and Biometricians (ordinarily used to calculate r_{tet})

$$\rho_1 = \frac{\sigma_1^2}{\sigma_o^2 + \sigma_1^2}$$

σ_1^2 = variance of item difficulties (standard score form)

$\frac{1}{\sigma_o^2}$ = precision of item

From these equations, and the assumptions mentioned above, one can determine that large ρ_1 would reduce the error term whether the ability level is equal to the mean difficulty of the items or not. The size of ρ_1 can be increased by decreasing σ_o^2 (using more precise items), by decreasing σ_1^2 in the denominator (or using all items at one difficulty), or by increasing σ_1^2 in the numerator (using items at more than one difficulty level). This immediately suggests that the best procedure is to use more precise items if one wishes to reduce error variance in the score, as σ_1 appears in both the numerator and denominator. This is in contrast with the most valid test results reported by Tucker, who empirically found that the most valid test was the test with imperfect items.[25]

Another way of reducing error variance would be to use the small t_o values. (The value, t_o, is necessary to enter

---

[25]Tucker, op. cit.

Pearson's tables.) Lawley gives the following formula for t_o:[26]

$$t_o = \frac{1}{\sqrt{2\pi}} \int_{\frac{x - \bar{\alpha}}{\sigma}}^{\infty} e^{-\frac{u^2}{2}} du = 1/2(1 - \bar{\alpha})$$

where

x = ability level (standard score form)

$\bar{\alpha}$ = mean difficulty level of cumulative test

$\sigma^2 = \sigma_o^2 + \sigma_1^2$ (as defined above)

To aid in the understanding of the interpretation of the formula given above the following summary data is reported for a test with the mean difficulty level of items nearly equal to mean ability level ($\bar{\alpha}$ = .045) and with a σ (a combination of the spread of item difficulty and precision of items) of 1.30 for a 100-item test.[27] The values of $\sigma^2_{E(X)}$ for given values of $\frac{x-\bar{\alpha}}{\sigma}$ are as follows:

$\frac{x - \bar{\alpha}}{\sigma}$	$\sigma^2_{E(X)}$
0.0	20.8
0.1	20.7
0.2	20.4
0.3	19.8
0.4	19.0
0.5	18.0
0.6	16.9
0.7	15.6
0.8	14.3
0.9	13.0
1.0	11.6

---

[26] Lawley, op. cit., p. 279.

[27] Ibid.

As can be seen from the preceding data, for given $\bar{\alpha}$ and σ values, the higher the ability level (x), the lower the error variance for the score ($\sigma^2_{E(X)}$) for a cumulative test. If the items had a large value (fixed) for the mean difficulty level (i.e., the value of $\bar{\alpha}$ increased) then the value of $\frac{x - \bar{\alpha}}{\sigma}$ would be smaller and thus the error variance ($\sigma^2_{E(X)}$) would be larger.

Lawley also pointed out that the effective discriminating power of a test may be computed as follows:[28]

$$\lambda = \frac{F(x) - F(x')}{\sigma^2_{E(X)} + \sigma^2_{E(X')}}$$

If $x = \bar{\alpha}$ then the above formula becomes:

$$\lambda = \frac{\sqrt{N} \cdot e^{\frac{\bar{\alpha}^2}{2}}}{\sqrt{\sigma_0^2 + \sigma_1^2 \cos^{-1}\left(\frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2}\right)}}$$

x and x' are two different ability levels

X and X' are two different score values

Other terms are defined as before.

As Lawley pointed out, in order to increase the effective discriminating power the numerator must be increased which means obtaining large values for $\bar{\alpha}^2$, or the denominator may be decreased, and, assuming σ_0^2 is constant (as one cannot change precision) then one must change σ_1^2 which is the spread of difficulty.[29] The smaller the spread of difficulties the

---

[28] Ibid., p. 280.

[29] Ibid., p. 281.

lower the value.

The effective discriminating power for a test would thus be greatest when the mean difficulty of items was equal to the ability level for the extremes in ability, and, when there was no spread of item difficulties. This type of test would be used to create scores which would be assigned only to individuals that are the same. It is not used to differentiate between the ability levels of individuals. The same logic which states that middle scores will be more precise (i.e., representing only one type of ability level individual) when difficulties are extreme would indicate that extreme scores will be more precise when 0.00 level of difficulty items are used in the test. (Remember the formula uses α^2 so it would operate for either extreme of difficulty.

Support for this position is given by Lord who stated that the standard error of measurement would be practically zero for extreme positive or negative values of ability.[30] He argued that there would exist individuals whose ability would be so low that the test would not be discriminating for them, and other individuals whose ability would be too high to be discriminated. The standard error of measurement is low for these zero or perfect scores and is necessarily smallest for those examinees for whom the test is least discriminating.

The above solutions to changing the criteria of test validity still do not exhaust the solutions to the attenuation

---

[30]Lord, A Theory of Test Scores, op. cit., p. 14.

paradox. Brogden offers yet another solution. He found that the correlation continued upward when a spread of item difficulties instead of one level of difficulty was used.[31] He concluded that the problem was that of determining the distribution of item difficulties to yield a more valid score. Brogden showed that by using items with $r_{tet} = .60$ or higher, a distribution of item difficulties will produce (for an 18-item test) a higher validity than will be obtained with all items at the .50 difficulty level.[32] The spread of difficulty seemed to be important when items were of this reliability.

Brogden's solution of determining the spread of items for a test such that the results would correlate highest with a criterion seems to be inadequate since there remain the problems of measuring the relationship and the meaning of the coefficients that are computed. It is impossible to solve all of these problems at this time, but assuming that the difficulty of the item is an adequate score, and assuming that discrimination among ability levels (with an examination of the effective discriminating power) is the important question, a rationale can be built for the sequential test developed in this dissertation.

Two areas of literature will now be examined to build the rationale for effective use of items in the sequential

---

[31]Brogden, op. cit., p. 240. [32]Ibid.

test. They are (1) literature on Bayes' Theorem, and (2) literature on the use of items at the 50 per cent level of difficulty for the hypothetical group with a median ability level equal to the value at which the discrimination is desired.

Meehl and Rosen, through the use of Bayes' Theorem, point out that the practical value of a psychometric sign, pattern, or cutting score depends jointly upon its intrinsic validity (in the usual sense of its discriminating power) and the distribution of the criterion variable (base rates) in the clinical population.[33] They note that if the base rates of the criterion classification deviate greatly from a 50-50 split, the use of a test sign having only moderate validity will result in an increase of erroneous clinical decisions.

One reason that the sequential test is assumed to have maximally efficient use of items is that the base rate does not have to deviate from the 50-50 split. The other reason is that the sequential test uses items at the 50 per cent level of difficulty for the group taking the item. These items have been found to be efficient with various criteria for efficiency.

Lord concluded from maximizing the ratio of difference in means to standard error of difference, that if one desires

---

[33]Meehl, Paul E. and Rosen, Albert. "Antecedent Probability and the Efficiency of Psychometric Signs Patterns or Cutting Scores," Psychological Bulletin, 52:194-216, No. 3, 1955.

to construct a test that will have the greatest possible discriminating power for examinees of a given level of ability, $c = c_o$, then all items should be of equal difficulty (no spread) and of such difficulty that half of those examinees whose ability score is c_o would answer each item correctly and half would answer it incorrectly.[34] This measure of discriminating power is completely independent of the distribution of ability in the group tested.

However, when item precision is such that item-total biserial correlations are .447, Lord empirically showed that a test composed solely of items at the 50 per cent difficulty is more discriminating (as measured above) than any other test for examinees at any level of ability between -2.5 and +2.5.[35] Lord does not show results of more highly correlated items which will be investigated in the present study.

Lord's empirical study above is supported by Cronbach and Warrington's theoretical study. They stated that for items of the type ordinarily used in psychological tests, the test with uniform item difficulty gives greater over-all validity and superior validity for most cutting scores, as compared with a test with a range of item difficulties.[36] It is the cutting score validity which is new here and of some relevance to the sequential test constructor. For

---

[34]Lord, A Theory of Test Scores, op. cit., p. 26.

[35]Ibid., p. 29.

[36]Cronbach and Warrington, op. cit., p. 127.

example, Cronbach and Warrington found that if $\sigma_d = .2$ (i.e., $r_{tet} = .94$ or $\phi = .80$), if no guessing is possible (or $r_{tet} = .55$ or $\phi = .37$ if the probability of chance success by guessing is one-third), and if all items are at the 50 per cent level of difficulty, better results are obtained for separating out from 40 to 62 per cent below the cutting score than if there were a normal distribution of item difficulties.[37]

The empirical determination of the best difficulties for discrimination has not always been as nonsupportive of the present rationale as the work of Lord. Lord used discriminating power (as defined by him) as his criterion. Richardson's empirical study had more supportive results. He created five subtests of different difficulty levels: 78-95, 60-77, 41-54, 23-40, and 5-22.[38] He then calculated the biserial correlations for 23 different divisions of the criterion starting at 4.17 per cent of the people in the lower category, and, by percentage units of 4.1667, continuing to 95.83 per cent in the lower category. He graphed these results and noted that the test consisting of items from 78-95 per cent passing produced the highest biserial correlation for those divisions where 4.17 to 25.00 per cent of the people were in the lower category. Likewise the 60-70 per cent pass test

---

[37] Ibid., p. 135.

[38] M. W. Richardson, "The Relation Between the Difficulty and the Differential Validity of a Test," Psychometrika, 1:33-49, No. 2, June, 1936.

was best for the 25.00 to 35.00 divisions; the 41-49 per cent pass test for the 35.00 to 61.50 divisions; the 23-40 per cent pass test for the 61.50 to 82.00 divisions; and the 5-22 per cent pass test produced the highest biserials where 82.00 to 95.83 per cent of the people were in the lower category. Although these results are from 50-item tests, the results indicate that different difficulty tests for different discriminations should be useful.

Other results from studies which would support the position that items at the 50 per cent level of difficulty for the group are the best items, are those which indicate differentiation of a group by items of different difficulty. In these studies the ability level of the individuals are not known and differentiation for each ability level is not reported separately. The reader must assume that the individuals were normally distributed around an ability level equal to the difficulty of the items. If this assumption is made then low differentiation by difficult items support the conclusion that items appropriate for the ability level are the best items.

Such a study as described above is reported by Cleeton. Cleeton used four well selected ability groups--one superior group and three inferior groups.[39] He then constructed two measures of the differential or predictive value of the test.

---

[39]Glen U. Cleeton, "Optimum Difficulty of Group Test Items," Journal of Applied Psychology, 10:327-340, No. 3, September, 1926.

One of these was $(R_1 - R_4)$ in which R stands for the number of items answered correctly by group 1, 2, 3, or 4. The other measure was $(R_1 - R_2) + (R_1 - R_3) + (R_1 - R_4) + (R_2 - R_3) + (R_2 - R_4) + (R_3 - R_4)$. (Terms having the same meaning as above). These are criterion II and criterion I in the following results, respectively. Cleeton examined difficulty by grouping 1/10 of the items in each interval and by grouping 1/10 of the range of difficulty in each interval. For present purposes it is most informative to look at the actual difficulty divided into 10 parts even though the number of items in each interval is different. The following data show the results of 240, 240, and 480 individuals each taking three tests of 400, 236, and 109 items. (For the computation of criterion indices, Cleeton assumed that he had only 720 individuals.)

Interval for % Passing Item	Rank Criterion I	Rank Criterion II	Value Criterion I	Value Criterion II
91 - 100	8	8	44.4	14.7
81 - 90	6	6	104.9	28.9
71 - 80	5	5	125.9	40.8
61 - 70	4	4	152.6	46.8
51 - 60	3	3	158.9	47.6
41 - 50	1	1	175.2	51.3
31 - 40	2	2	163.9	51.1
21 - 30	7	7	85.8	26.1
11 - 20	10	10	35.9	11.1
0 - 10	9	9	37.3	11.9

From the above data one may determine that the slightly more difficult items seem to have the greatest predictive value as measured by both these estimates of predictive value. This would support the decision to use items at the 50 per cent level of difficulty for the group which is to be discriminated among.

Logical analysis also supports the above decision. Flanagan pointed out the extremes of this difficulty and item validity argument. He stated that if one wanted the maximum amount of discrimination between the individuals in a particular group, a test should be composed of items all of which are at 50 per cent difficulty for that group--provided the intercorrelations of all the items are zero.[40] If intercorrelations were other than zero, the decision would not be this clear.

Lord studied theoretical test models which had either high or low item reliabilities with easy, difficult, or easy and difficult test items. After examining the relationship of the true score distribution to the distribution of ability, he reached the following conclusion:[41]

> A test composed of items of equal discriminating power but of varying difficulty will not be as discriminating in the neighborhood of any single

---

[40]John C. Flanagan, "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from Data at the Tails of the Distribution," Journal of Educational Psychology, 30:674-680, No. 9, December, 1939.

[41]Lord, "The Relation of Test Score to the Trait Underlying the Test," op. cit., p. 543.

ability level as would a test composed of similar items all of appropriate difficulty for that level.

Thus, most of the literature supports (1) the use of items at the appropriate difficulty for each level and (2) the separation of individuals into groups that would have a base rate of 50 per cent.

Because the base rate is near a 50 per cent split each time, the sequential model should permit the use of only moderately discriminating items. In the cumulative test, there will be only 5 or 10 per cent of the individuals who should pass a difficult item, as all people take the item. In the sequential method 50 per cent should pass this difficult item, as only those with high ability will take the item. According to Bayes' Theorem the probability of high ability people passing the item must be much higher than the probability of low ability people passing the item if 90 per cent of those taking the item have low ability. Once the group taking the item has a base rate of 50 per cent (as is the case in the sequential method), then the item should work better--i.e., increase the number of correct clinical decisions.

In the sequential test, those groups which are different in ability would use items at the 50 per cent level of difficulty for that group. This would allow the use of difficult items which are precise. Such items could not be efficiently used in a cumulative test.

II. CONTROL OF THE SCORE DISTRIBUTION

The problem of score distribution is not only to assign a specified number of individuals to each score value, but also to assign like individuals to each score value. The score distribution is not only related to the item parameters, but should also be related to the use. The score distribution problem may be studied through the use of a theoretical model or empirically.

Lord attempted to study the problem of control of score distribution through the use of a theoretical model. He made the following assumptions: (1) the item characteristic curves have the general shape typical of cognitive items that are not answered correctly by guessing; (2) the items are homogeneous in a certain specified sense; (3) the items are scored 0 or 1; and (4) the raw test score is the number of items answered correctly.[42] (A homogeneous test is, for Lord's purpose, defined as a test composed of items such that, within any group of examinees all of whom are at the same ability level, the response given to any item is statistically independent of the response given to the remaining items.)

The generalizations reached by Lord were as follows:[43]

> 1. Since the test characteristic curve is in general nonlinear, the test score distribution will not in general have the same shape as the distribution of

---

[42] Ibid., p. 546.

[43] Ibid., pp. 541-542.

ability; in particular, if the ability distribution is normal, the score distribution in general will not be strictly normal.

2. U-shaped and roughly rectangular score distributions can be produced provided sufficiently discriminating test items can be found. (All appropriate individuals pass or all appropriate individuals fail an item if they are perfect items at the 50 per cent level of difficulty.)

3. Typically, if a test is at the appropriate difficulty level for the group tested, the more discriminating the test, the more platykurtic the score distribution.

4. The skewness of the test score distribution typically tends in a positive direction as the test difficulty is increased above the level appropriate for the group tested; in a negative direction as the test difficulty is decreased below that level.

These generalizations aid in interpreting the empirical results of a study made by Mollenkopf.[44] He selected 1000 answer sheets chosen on the bases that: (a) every person must have attempted every item, and (b) a wide range of scores should exist in the sample chosen. Items were then chosen to make up nine synthetic tests. These nine tests contained score distributions with three types of kurtosis and three types of skewness. A study of the literature revealed that the total test score distributions were believed to be controlled for skewness by item difficulty. However, since easy items tended to have higher correlations with the total score than did difficult items, control on mean difficulty alone was found not to be sufficient. When building a test with a symmetrical score distribution, Mollenkopf found that a set

---

[44]William G. Mollenkopf, "Variation of the Standard Error of Measurement," Psychometrika, 14:189-229, No. 3, September, 1949.

of items of the same type (all of difficulty close to .50) yielded scores with a definitely flat distribution. (From Lord's work, it looks as though the item precision must have been very good.) To secure a leptokurtic score distribution Mollenkopf tried sets of items with .40 and .60 difficulties, but found that homogeneous sets of items of .20 and .80 difficulties were needed.

If one uses Lord's work to translate back from score distribution (by assumed highly precise items) to ability level, one can determine that the distribution of ability must have been near normal. Also of interest in the Mollenkopf article is the fact that the standard error of measurement for a nonskewed platykurtic distribution of scores is greatest in the middle sections and lowest at the extremes. This may be accounted for by what Mollenkopf has labelled the "end effect."[45] This effect means that at the ends large differences in parallel forms cannot occur. A perfect score is perfect in each half. Small empirically observed errors of measurement are inevitable in the tail where the pile-up occurs on skewed distributions but not for normal distributions.

This explanation would suggest that the variance of ability levels for a given test score may be small, but it does not indicate, as Mollenkopf also pointed out, that there

---

[45] Ibid., p. 212.

is a small variance of scores for a given ability level. Both points are of interest if reflection of the ability distribution is desired in the score distribution.

The cumulative test can be used to yield the type of score distribution that one wishes. The important parameters are item difficulty and item precision, but only general statements are available as to the relationship between these parameters and the score distribution. Empirical studies are used to determine exact parameter values for given score distributions.

Hymphreys stated that the variance of item difficulties forces scores toward the center of the distribution and thus counters the effect of high item intercorrelations.[46] It is thus necessary to have a spread of difficulties, only if the items are very precise. Whereas very highly intercorrelated items of one difficulty level would produce two scores, if one were to use a spread, one could force people into a distribution that would be expected to have some validity. Humphreys advocated that the shape of the score distribution be controlled by the difficulty level of the test items.[47] The type of distribution favored by Humphreys was a rectangular distribution--a distribution that would allow individuals to be ranked.

---

[46] Humphreys, op. cit., p. 474.

[47] Ibid., p. 475.

If the items were perfect, the procedure to produce the rectangular distribution desired by Humphreys would be as reported by Davis. Davis reported that if the tetrachoric item intercorrelations are all unity, a rectangular distribution of raw scores is most likely to be obtained by selecting items with difficulty levels of $1/(n + 1)$, $2/(n + 1)$, $3/(n + 1)$, . . . $n/(n + 1)$. However, if the tetrachoric intercorrelations are all .50, a rectangular distribution of raw scores is most likely to be obtained by selecting all items at the 50 per cent level of difficulty.[48] He argued that for any level of tetrachoric item intercorrelations from zero to .50, the maximum number of discriminations that could be made by the total score would be insured by selecting all items at the 50 per cent level of difficulty.

Davis went on to say that this simple mathematical procedure employed to specify the exact difficulty levels of items for two- and three-item tests cannot be applied to specifying the exact difficulty levels of items for tests containing larger numbers of items except in the limiting case when the item intercorrelations are all unity. The reason one cannot generalize is that when intercorrelations are not unity, errors in classification will be made, and the spread of ability represented by those who pass or fail will be greater but undetermined. Thus, the appropriate difficulty

---

[48]Davis., op. cit., p. 103.

for the resulting group cannot be easily determined. The effect of errors is difficult to determine, but as pointed out by Davis, there is need for a general solution.

Whereas the general rules about control of score distribution are known, there is no general solution in the sense that the actual score distributions are known. The actual score distributions must be empirically determined for each test. The literature indicates that if the sequential method of testing could more easily and predictably control the score distribution, a real contribution would be made to the solution of a difficult measurement problem.

III. MEANING AND USE OF SCORE PRODUCED

Both the score distribution and the meaning of a score are related to the use of the test. Ferguson has pointed out that for discrimination between two groups one would need a bimodal distribution of scores; the discrimination between two groups and among the members of one group would require an asymmetrical distribution of scores; and, if one were establishing the order of ability of individuals, one would use a rectangular distribution. Ferguson concluded that the construction of tests to yield distributions approximating the normal form results in a loss of discriminatory capacity.[49]

---

[49]George A. Ferguson, "On the Theory of Test Discrimination," Psychometrika, 14:61-68, No. 1, March, 1949, p. 68.

Not all scoures have the same meaning. A score resulting from the discrimination between two groups is more a probability statement that the individual should be classified into a given category than it is a statement that the individual's ability is at a certain level. The score from a test designed to rank individuals compares any individual in relation to others.

In addition to the meanings necessary for the above uses, Gulliksen (as stated in the first section) would have the score be the best estimate of the difficulty level reached.[50] This type of score represents the "true ability" level of the individual. This type of score is also advocated by those who argue for reproducibility as a measure of the best test. However, it should be noted that it has been the practice to determine how well a pattern of responses from an instrument will reproduce original results, not hypothesized "true" results. As reported by White and Saltz, these indices will reflect without equivocation the amount of information thrown away by representing the subject's performance on the test by a total score based on the number of items passed. "They indicate, in other words, how adequately a unidimensional model fits the obtained data."[51]

---

[50]Gulliksen, op. cit.

[51]Benjamin W. White and Eli Saltz, "Measurement of Reproducibility," Psychological Bulletin, 54:81-99, No. 2, March, 1957, p. 95.

However, a reproducibility score from a unidimensional test does not insure either an interval scale or a known behavior domain being sampled. Individuals may be ranked by the test scores (compared to other individuals) or be assigned an ability level (compared to a standard). The behavior domain may be related to the test label or it may not--the only assurance one has is that the domain is unidimensional.

The question as to domain samples (which seems like a validity question) has actually been studied as a part of reliability. Tryon in theory related reliability to the behavior domain sampled.[52] He reviewed the two theories of test reliability: (1) the Spearman-Yule theory that tests are unreliable because of an error factor and reliable because of a true factor which may be a composite of more than one common factor; and (2) the Brown-Kelley theory that reliability may be explained by equivalent test-samples in which all items in the total score have equal standard deviations and equal intercorrelations. (To obtain equivalent test-samples the content and difficulty of items must be considered, but all items do not have to be equally difficult.)

Tryon defined reliability as the value of "correlation, r_{tt}, between the observed X_t scores and a second set of composite scores, X_t', earned on a 'comparable form' of the X_t

---

[52]Robert C. Tryon, "Reliability and Behavior Domain Validity: Reformulation and Historical Critique," Psychological Bulletin, 54:229-249, No. 3, May, 1957.

composite."[53] (A comparable X_t composite is one in which the n test-samples vary on the average as much in standard deviations and intercorrelations as do the n test-samples in the observed X_t composite.)

If this definition of reliability is used, a reliable test is one that indicates how well the individual knew the domain or how he ranked with others in his knowledge of the domain. At least the domain sampled by the score is known and can be made part of the meaning of the score.

The literature reviewed to this point would indicate that the score (1) may be a function of difficulty which probably reflects the ability level of the individual, (2) may represent a pattern as to content, or (3) may indicate how well the individual did on the samples of the domain that the test is hypothesized to sample. Reliability measures may be a factor in determining what meaning can be assigned to the score, but there are still contributions coming from content and from difficulty.

Swineford examined the importance of the difficulty of the item as a factor in the score assigned to the individual. Swineford has shown that only if the items are quite precise and intercorrelated is the difficulty of the item an important factor in the score of an individual. Swineford used present

---

[53] Ibid., p. 230.

day tests and attempted to measure the impact of variability of item difficulty and item-item correlation.[54] The variability of item difficulty was designated σ_{Δ}, Δ being the normal-curve deviate (for a distribution with mean of 13 and standard deviation of 4) above which lies the area under the curve equal to the proportion of successful examinees. For a measure of inter-item correlation Swineford used the reciprocal of the square of the mean of the item-total correlation.

The results of Swineford's study showed that when the score was the number correct that the best formula for predicting this score was as follows:

$$Z_1 = .1530 \; Z_3 + .8649 \; Z_4$$

Z_1 is the predicted standard score on the test

Z_3 is the measure of the spread of item difficulties in a standard score form

Z_4 is the inter-item correlation measure in standard score form

$R_{1.34} = .9648$ for this formula.

When the score was the number right minus k times the number wrong the results were as follows:

$$Z_1 = .2117 \; Z_3 + .9222 \; Z_4$$

and $R_{1.34}$ was .9642. The symbols are the same as above. As can be seen from these formulas, the contribution of spread

---

[54]Frances Swineford, "Some Relations Between Test Scores and Item Statistics," Journal of Educational Psychology, 50:26-30, No. 1, February, 1959.

of item difficulties in the usual cumulative test is not great.

Another way of looking at the contribution of item difficulty spread is to specify the spread and inter-item correlation, and then examine the standard deviation of test scores. Swineford used (n - chance)/σ_t as her measure of standard deviation because "although it does vary with test length, the variation is not great for reliable tests when the longest is no more than 8 or 10 times the length of the shortest."[55] (In this formula "n" equals the number of items in the test and "chance" equals the number of items assumed to be correctly answered by chance.) The smaller the number from this quotient, the better one would assume the score to be because a large variance in relation to the total possible score is considered best according to present test theory. For the highest inter-item correlations, r_{bis} = .50, the values of (n - chance)/σ_t range from 5.8 to 3.0, for the largest (3.5 σ_Δ) to the smallest (0.0 σ_Δ) spread of item difficulties. Thus, 0.0 σ_Δ has the lowest value or produces the best test. However, the mean value for (n - chance)/σ_t is 6.2891 and the standard deviation is 2.6847 for the entire battery of tests studied. It should be noted that the (n - chance)/σ_t values reported above are all below the mean and within approximately 1.3 standard deviation units from each other. Therefore, the conclusion should be that in this range, the smaller the spread of difficulties the better the score

---

[55] Ibid.

value tends to be. There is no conclusion reached about the entire range but Swineford's data would support using no spread of item difficulties.

The standard deviation from the entire battery of tests may not be the most appropriate value to use, but this is the only value available. The standard deviation of σ_Δ is .4391 and the standard deviation of $1/r^2$ is 5.4344. From these values one may note that there are about six standard deviation units of σ_Δ included in the one axis of her chart where values of σ_Δ range from 5.8 to 3.0 for the highest (.50)r_{bis}, and from 14.8 to 11.9 for the lowest (.20) r_{bis}. The mean r_{bis} is .36, the highest r_{bis} (.50) is .70 sigma units away from the mean, and the lowest r_{bis} (.20) is 3.15 sigma units away from the mean. Thus, while the values of σ_Δ may be considered to be close to normally distributed and likely to be encountered in the usual cumulative test, the values for r_{bis} are not normally distributed. We might conclude that if r_{bis} were normally distributed, then higher values of r_{bis} might appropriately be investigated. A standard deviation unit on σ_Δ would indicate that today most tests do use items centered around the mean difficulty level, but that the reliability of items has a larger range. If one examines $\pm$ .70 sigma units of r_{bis}, one has about a three point change in (n - chance)/σ_t values which is about the same change encountered from $\pm$ 3.0 sigma units of σ_Δ . This supports the conclusion that conventional cumulative tests do not use

difficulty as a major factor in the score; the score is a conglomerate of difficulties and other factors.

The literature indicates that the cumulative test may be constructed to measure a single factor but that the attention of the test constructors has not been directed toward reporting the decisions made as to the meaning of the score. If one remains concerned with traditional operational definitions of reliability and validity, one may forget the construct operationalized and not change the construct when it needs to be changed.

The sequential test procedure developed in this dissertation will use reflection of true ability as the meaning of a scores. The literature indicates that this is only one of the many meanings that could be assigned to a score.

IV. SEQUENTIAL TESTING PROCEDURES

The literature indicates that there are many choices as to the use of the sequential testing procedure. The sequential process may be used (1) to quickly determine score to be assigned to good and poor students; (2) to determine to which of two categories the individual should probably be assigned, if assigned at all; or (3) to classify each individual as well as possible in time allowed. The sequential analysis developed by Wald would be most applicable to the second purpose, but this method has been modified by Cowden to serve the first purpose.

Cowden has indicated that when an examination is given to a student it sometimes happens that not enough questions are asked to permit a fair evaluation of his knowledge and ability.[56] On the other hand the examination is sometimes drawn out longer than is necessary. If a student is very good or very poor, only a few questions may be needed to establish this fact beyond reasonable doubt; but borderline students need to be examined at considerable length before deciding whether they should be passed or failed. If sequential testing is used, the fate of good students and of poor students tends to be quickly determined, but mediocre students must continue with the examination until the results give adequate grounds for a decision. By use of the sequential method the number of questions answered by a student is reduced to a minimum, and at the same time the probability of passing a poor student or failing a good student is controlled.

Cowden graded his students in a small class in elementary statistics at the University of North Carolina. Using D_1 (decision number 1) to indicate the number of questions that could be missed and still permit a student to pass, D_2 (decision number 2) to indicate the number of questions that must be answered incorrectly before a student is failed, and N to indicate the cumulative number of questions answered; the two linear equations used to make the decision follow:[57]

---

[56]Dudley J. Cowden, "An Application of Sequential Sampling to Testing Students," Journal of the American Statistical Association, 41:547-556, No. 236, December, 1946, p. 548.

[57]Ibid., pp. 548-549.

$$D_1 = a_1 + bN \qquad D_2 = a_2 + bN$$

As can be seen, the straight lines representing these two equations are parallel and differ only as to the constants a_1 and a_2. These constants a_1 and a_2 are shown to depend on the values of p_1, p_2, α, and β when: "p_1" is defined as the maximum proportion of errors in all possible questions of a given type made by a student who is definitely good; "p_2" is defined as the minimum proportion of errors in all possible questions of a given type made by a student who is definitely poor; "α" is defined as the probability of failing a good student; and "β" is defined as the probability of passing a poor student. The more widely p_1 and p_2 differ the closer together the lines will be, and, therefore, the more quickly will a decision be reached. The larger the values of α and β the smaller will be the value of a_2 and the larger (algebraically) will be the value of a_1. Therefore to bring the two lines closer together one must increase α and/or β. The value of a_1 is always negative, since answering all questions correctly does not strongely indicate knowledge of the subject until a reasonable number of questions is answered (what is a reasonable number depends on the value adopted for β, becoming larger as β is made smaller). On the other hand, a_2 is always positive, but a decision to fail cannot be reached until $D_2 = N$, since a student cannot miss more questions than he answers. When $\alpha = \beta$, $a_2 = -a_1$. The

slope b is independent of α and β, but depends exclusively on p_1 and p_2. Cowden gives the following formulas:[58]

$$g_1 = \log \frac{p_2}{p_1} \qquad g_2 = \log \frac{1 - p_1}{1 - p_2}$$

$$-a_1 = h_1 = \frac{\log \frac{1 - \alpha}{\beta}}{g_1 + g_2} \qquad a_2 = h_2 = \frac{\log \frac{1 - \beta}{\alpha}}{g_1 + g_2}$$

$$b = \frac{g_2}{g_1 + g_2}$$

Cowden thus develops two lines for pass, fail, and indeterminate, but has grades for six categories based on the following decisions:[59]

> After 20 questions if a student made errors in less than 10 percent of the questions, the grade of "A" was assigned; if 55 per cent or more of the questions were answered incorrectly, the grade of "F" was assigned; if the percent of incorrect questions was between these percentage values then testing was continued. After 40 questions if a student (not classified before) made errors in less than 22.5 percent of the questions the grade of "B" was assigned; or if more than 45 percent of the 40 questions were incorrect, the grade of "F" was assigned. Similar decisions were made after 60, 80, 100, 200, and 1,000 questions. After 1,000 questions those students not already classified were assigned "D" or "E" grades. Those individuals having errors in less than 34.89 percent of the questions were assigned "D" and those students having errors in more than 35.3 percent of the questions were assigned a grade of "E."

Sequential testing is thus changed to allow using more than three categories by changing the number of items that

---

[58] Ibid., p. 551.

[59] Ibid., p. 552.

are used to make the decision. Estimates of the size of the number of items can be obtained by the following formulas:[60]

$$\bar{N}_o = \frac{h_1}{b} \qquad \bar{N}_1 = \frac{h_2}{1 - b}$$

$$\bar{N}_{p_1} = \frac{(1 - \alpha)\, h_1 - \alpha\, h_2}{b - p_1} \qquad \bar{N}_{p_2} = \frac{(1 - \beta)\, h_2 - \beta\, h_1}{p_2 - b}$$

Cowden found that it took 13.5 items before it was possible to decide that the student should pass. This is due to a random sample of items assumed in the sequential process. It therefore seems worthwhile to investigate a purposeful sample of items instead of random sample even though the mathematics has not been worked out for this type of test.

To use the model developed by Wald, one must first state the probability of type I and type II errors that one will accept (as to a given alternative) and then continue until one satisfies the conditions of the mathematical model with probabilities.[61] The procedure may be used to decide upon pass or fail categories as was done by Moonan; or modified by making assumptions about the number of items needed to make the decision as done by Cowden; or an individual may wait for the mathematics of the multiple decision (or other modification) to be completed and reported as Wald indicates might be done in his book on sequential analysis.[62]

---

[60] Ibid., p. 553.

[61] Wald, Sequential Analysis, op. cit.

[62] Ibid., pp. 138-150.

The sequential procedure developed by Wald for a "most powerful" test is built upon the assumption that one may continue to sample the same universe. The procedure determines what decision is best after every sample and states whether one has attained the desired degree of probability (of being correct). It is not necessary to follow the lead of Cowden and Moonan and, therefore, use a random sample of items. It is known that certain items of different difficulties will give more information about an individual than other items, and this information should be used: this means that one does not wish to sample from the same universe of items each time. While the aptitude or ability being tested must remain unidimensional, there may be great advantage in allowing the difficulty of items to change. The sequential model herein described thus departs from the Wald sequential model in that it uses different difficulty levels so that fewer items are needed for the decision.

Fiske and Jones in an article intended to introduce sequential analysis to psychologists, stated that the uncritical use of sequential analysis obviously is not recommended.[63] It is a design which can have advantages when one or more of the following conditions actually holds: (a) The problem involves the choice between <u>two</u> possible parameter values which can be specified on <u>a priori</u> but not arbitrary

---

[63]Donald W. Fiske and Lyle V. Jones, "Sequential Analysis in Psychological Research," <u>Psychological Bulletin</u>, 51: 264-275, No. 3, May, 1954, pp. 273-274.

grounds--the null hypothesis will usually be one of the two; (b) the data are such that the cost per datum is high and economy is desired; and (c) the total amount of data is not fixed.

Such criteria would lead one to believe that the sequential model developed by Wald may not be the appropriate model for the test situation, as the total amount of data is fixed and one cannot afford to have 1,000 items as indicated by Cowden. It may be no more expensive to acquire the data from all candidates than from a few, unless one wishes to select only rather than classify. The decision to accept or not accept--the selection question--seems to be the most appropriate decision which can be answered by the sequential method as described by Wald.

The literature also indicates methods of presenting the material to the testee. Some of these are noted here. Glaser, Damrin, and Gardner constructed a tab item test to aid in training of electronics specialists.[64] In this test, the performance on one test yields information which supplies a cue for the selection of the next test and subsequent procedures. One "tab item" test, for example, had the trainee read a description of the malfunction of a television set and then, rather than actually performing various checking

---

[64]Robert Glaser, Dora E. Damrin, and Floyd M. Gardner, "The Tab Item: A technique for the Measurement of Proficiency in Diagnostic Problem Solving Tasks," Educational and Psychological Measurement, 14:283-93, No. 2, Summer, 1954.

procedures, the trainee pulled the tabs of those checks he would make if he were actually trouble shooting a real television set. Whenever he pulled a tab he uncovered the information he would have obtained if he actually had performed that check on a real set.

Another method of presentation was used by Krathwohl and Paterson in preliminary studies of the sequential test model. They had directions printed on the page, covered these with a transparent hard finish ink so that directions could not be erased, then covered this in turn with strips of opaque ink. The testee erased the strip of opaque ink under the letter he considered to be related to the correct answers. (This is similar to an IBM answer sheet, but instead of marking a spot, the testee erases a spot.) The appropriate directions were thus made available to the student.

Teaching machine presentations are also obvious methods to present material to the testee. The material is similar to that presented by teaching machines, but in the sequential model being developed in this paper, the individual does not obtain information about the correctness or the reason for the correctness or incorrectness of the response. However, the individual is told to take a more difficult item if he correctly answered the preceding item, or a less difficult item if he incorrectly answered the preceding item.

The literature suggests that if the decision is to best classify the individual by a sequential procedure, the

present sequential model may be better than past models which have been developed from different assumptions and for different problems. The literature also suggests that traditional scores represent more than one meaning.

The present sequential model has used reflection of input in the output as the proper meaning for a score; the cumulative test should not perform this function as well as the sequential test. The decision as how to measure the efficiency of these tests (and indirectly the items) was then related to the reflection of input in the output. The two factors considered in the output were (1) the means and variances of ability levels assigned to a score (precision of score) and (2) the means and variances of scores assigned to an ability level category (discrimination of test).

It should be noted that the decisions as to the type of score distribution desired and the meaning that should be assigned to a score had to be made before one could determine the efficiency of the test (or items). The decisions made in the present study were those decisions which it was hoped would favor the sequential test procedure.

There should be maximally efficient use of items in the sequential method as (1) there is a separation of individuals into groups which have a base rate of 50 per cent for the items used, and (2) the use of items at the 50 per cent level of difficulty for the subgroups permits the use of more

difficult items and makes better separation of these individuals (as the item is at the 50 per cent level of difficulty for the subgroup).

CHAPTER III

PROCEDURES

There are six sections to this chapter. First, the actual construction of the six-item cumulative and the six-item sequential test model is considered. The second section outlines the method of evaluating the hypotheses stated in Chapter I which relate to the effect of input distributions. The third and fourth sections show the methods for testing the hypotheses about item precision and difficulty, and effect of errors of estimating a parameter, respectively--both for the sequential model. Fifth, some general comparisons between test score distribution and ability level distribution are examined. And finally, a summary of procedures and hypotheses is presented.

I. TEST MODEL CONSTRUCTION

This section deals with the construction of six-item sequential and cumulative test models. Later these test models are used with different inputs of ability and the type of score output is examined.

The test model for the sequential and cumulative tests assumed that the probability of passing an item was dependent

upon three factors: (1) the ability level of the individual, (2) the precision of the item, and (3) the difficulty level of the item. The assumption was made that no one passed by randomly guessing the correct answer to the item.

The ability level of the individual was specified in terms of standard score units for a normalized distribution of ability. The precision of the item was specified in terms of either r_{bis} or σ_d. These two terms are related by the following formulas:[1]

$$\sigma_d = \frac{\sqrt{1 - r_{bis}^2}}{r_{bis}} \qquad (1)$$

or by algebraic manipulation;

$$r_{bis} = \frac{1}{\sqrt{1 + \sigma_d^2}} \qquad (2)$$

As can be seen from the second formula, r_{bis} is equal to one if σ_d is equal to zero. The smaller the σ_d value the more precise the item, and if σ_d were equal to zero, the individuals who had ability levels above the difficulty level of the item would pass the item, and vice versa.

The difficulty of the item was expressed in terms of standard score units for a normal population. It need be remembered that 80 or 90 per cent of a select group could pass (or fail) a 50 per cent difficulty item.

---

[1]Frederic M. Lord, "Some Perspectives on 'The Attenuation Paradox in Test Theory'," Psychological Bulletin, 52: 505-10, No. 6, November, 1955, p. 506.

The probability of passing a single item for a given small segment of ability was computed by determining the area under the normal curve from $-\infty$ to the value $\frac{a - d}{\sigma_d}$; where"a"is equal to the ability level of the individual in standard score or sigma units,"d" is equal to difficulty level in standard score or sigma units, and"σ_d" is the measure of precision described above.

The probability of passing a sequence of items for both the sequential and the cumulative was determined by multiplying the probabilities of passing each item in that sequence. This assumed that for that small segment of ability (for which the probability of passing an item was determined), performance on any one item was experimentally independent of performance on any other item. Since the concern was with classifying people by ability, it was assumed that each of these items measured only one factor other than the error factor, i.e., the test was unidimensional. The error factor on any one item was assumed to be independent of error on any other item.

Using the above scheme, one six-item sequential test model was constructed for a hypothetical population of 1500 individuals with 100 people at each of 15 ability levels as shown in Table 24. The item precision for all items in this model was arbitrarily set at $\sigma_d = .882$. The appropriate difficulties were determined by the following procedure. First, the number of people at each of the 15 ability levels who

would pass or fail an item was computed. The value of the sum of the deviations from the mean squared for each of the ability scores was computed for the pass and fail groups. This value was computed and graphed for different trial values of difficulty until the difficulty level was found for which the sum of all sets of deviations of ability level about the mean ability level for the entire group was a minimum. Since ΣX^2 was a constant, the value for difficulty level was calculated by maximizing $(\Sigma X)^2/N$. The difficulty level of the item taken by each group was not the same. For example, in Figure 1, both the group who passed the first item and failed the second item and the group who failed the first item and passed the second item take the same item at stage 3-- a 0.00 item. If this had not been done, the six-item sequential test would require 63 different items.

It was decided to use the same item for those groups for which $\Sigma(\Sigma X)^2/N$ maximized at a difficulty level no more than .20 standard score units away in difficulty from each other. This allowed the test to be built with fewer items and thus any test built to correspond to the model could use only the most precise items in a pool of items. Also, this corresponds with reality for it is unlikely that items which are less than .20 standard units of difficulty apart can be adequately distinguished one from the other under usual conditions for determining difficulty. When more than one group used an item of a given difficulty then the $(\Sigma X)^2/N$ was

maximized across all the groups using that item. (If one should desire to construct other tests along similar lines, it would seem desirable to use an electronic computer as there were over 100 hours needed to build this one test on a hand calculator.)

The six-item cumulative test was constructed for the same hypothetical population as for the sequential test. The item precision was likewise arbitrarily set at $\sigma_d = .882$. However, all items were at the 50 per cent difficulty level.

Raw scores for the sequential test were the rank of the mean criterion level of the group with 64 being the highest possible value. There are 64 possible sequences when there are six items with dichotomous classification for each item. The raw scores for the cumulative test were the number of items (out of six) that the individual was computed to have answered correctly. Both of these raw score distributions were converted to normalized "T" scores so that the two score distributions might be compared on an equivalent interval scale basis.

II. EFFECT OF SHAPE OF DISTRIBUTION OF ABILITY

It was hypothesized that the sequential test model constructed as described above should work well for any type of input distribution and thus be better than the six-item cumulative test model. The six-item cumulative test constructed with all items at 50 per cent level of difficulty

was not expected to be effective for those distributions which had many high ability individuals. It was hypothesized from the literature that these individuals would need more difficult items to discriminate among them. To test this hypothesis different ability distributions were used as input. The difficulty levels of the items used in the sequential test model were determined according to the method described in the last section, and were for a precision level of an r_{bis} item total correlation of .75 (or σ_d = .882). A precision of .75 was used because differences between the six-item cumulative and sequential models should be greatest at high levels of precision--.75 would be considered very high by the standards in use. Few tests have an average item-total correlation of .75. A rectangular input distribution was used as the items selected were hoped to operate well for any distribution. Not only is the rectangular distribution a good compromise, but with the same number at each ability level, only the ability level should determine the selection of the difficulty of the item.

To determine if the sequential model would work well when compared with the cumulative test model the two test models--the sequential and the cumulative--were each used with a normal and a U-shaped distribution of ability to make the total of four tests. These four tests were constructed in an electronic computer. (For both the normal and the U-shaped distributions the individuals were assumed to be

distributed over 15 input categories. Since the values used in the computer program were proportions at each category, any number of individuals may be assumed. The most common assumption made in interpreting this data is that there were 1000 individuals distributed over these 15 input categories.) The item difficulties for the sequential models were the ones computed above. The item difficulties for the two cumulative models were all at the 50 per cent level. It was thus possible to compare not only the sequential with the cumulative models, but also the effect of an input of normal and U-shaped distributions.

Effect of Normal Distribution

The effect of an input of a normal distribution of ability on the output distribution was examined in several ways, but before the examination of hypotheses related to these effects a description of the particular distribution used here is given. The normal curve was divided into 15 sections from +2.5 to -2.5 sigma units. The middle 13 categories were assigned the proportion of individuals that would fall in that portion of a normal curve, but all those individuals more than 2.5 sigma units from the mean were considered to be in the end categories. This was done because spreading these individuals over the middle of the distribution would have underrepresented the number of people likely to be at extreme values in ability. Using the above procedure the end

categories extended from ±1.612 to ±1.736 sigma units. Since there were so few to consider at the levels beyond ±1.736 sigma units, these individuals were all considered to be at the mean ability level for all people beyond ±1.612: that is, at 1.942 sigma units (see Figure 2).

To test the hypothesis that the cumulative and sequential test models have equal ability to classify individuals of mean ability level, the means and variances of comparable normalized scores from the six-item cumulative and "least squares" sequential test models for those 100 individuals assumed to be in category eight of ability (the middle category) were tested for significance of difference. The means were tested by use of a "t" test and the variances by use of an F ratio.

To test the hypothesis that the "least squares" sequential test model should more accurately classify the few individuals at the extremes of the ability scale than the six-item cumulative model, the means and variances of comparable normalized test scores for the 84 individuals in ability categories 14 and 15 were tested. (Testing of the individuals in the lower categories, one and two, was unnecessary since the resulting scores are symmetrical about the mean. Scores resulting from actual test administrations might be skewed by individuals guessing at the correct answer.)

To test the hypothesis that the cumulative test model should have scores with smaller variance of ability levels at

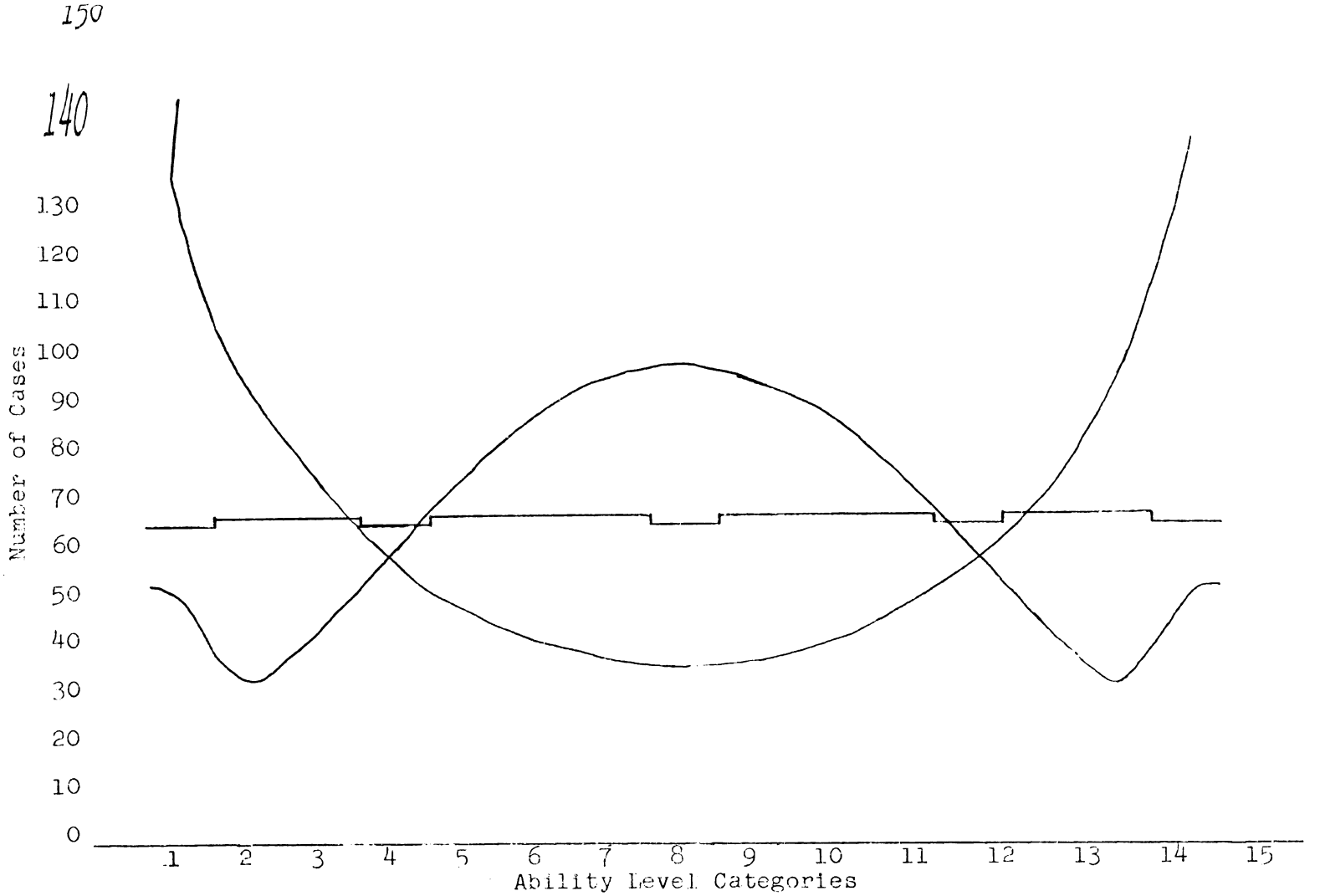


Fig. 2. Three Distributions of Ability

the extremes than the sequential test model, the means and variances of ability level scores for the individuals ranked in the top 8.4 per cent of the score distribution for each test model were tested. When it was necessary to take only a proportion of a score group to complete the top 8.4 per cent of scores, then the ability levels were proportionately sampled. The value of 8.4 per cent was selected because there were 84 individuals in the top two input ability levels of the hypothetical population of 1000 individuals.

It was hypothesized that the six-item cumulative test model would produce scores representing finer ability units in the middle than at the extreme score values, while the sequential test would more nearly reflect the ability scale. The differences between the mean ability levels of adjacent raw score categories for the cumulative test model were hypothesized to be smaller in the middle and greater as extremes were approached. These differences in mean ability values for the adjacent scores in one-half of the symmetrical score distribution are shown in Tables 5 and 6. In addition to this, the differences between mean normalized "T" scores for each adjacent ability level for both the sequential and cumulative tests are shown in Table 4.

Effect of U-shaped Distribution

The effect of the U-shaped distribution of ability was studied by the same procedures used with the normal distribution

of ability. The distribution used in these tests is the one shown in Figure 2. To determine if the "least squares" sequential test would more accurately classify individuals at the mean of the absolute ability levels than would the six-item cumulative test, the means and variances of normalized scores assigned to category thirteen were tested for significance of difference between scores assigned by the six-item sequential and the six-item cumulative models. Category 13 was selected as it included the mean value of ability for those individuals in the top half of the ability distribution.

To examine the hypothesis that the sequential test model would more accurately classify individuals at the extreme values of the ability distribution than would the six-item cumulative test, the means and variances of normalized scores assigned to category 15 individuals were compared for the sequential and cumulative test models.

To test the hypothesis that the cumulative test model would have more precise scores at the extremes of the ability distribution than would the sequential test model, the individuals ranked in the top 13.5 per cent for each model's score distribution were examined for differences in means and variances of ability level. These top-scoring individuals were proportionately selected as stated for the normal distribution. The top 13.5 per cent of the score distribution was used as there were 13.5 per cent of the individuals in the top ability category.

To determine if the classification of the middle ability level was more finely classified by the six-item cumulative, the mean normalized "T" score for each ability level was determined and shown in Table 24. The same was done for the sequential test model. The hypothesis was that the sequential model should have approximately equal distances between test score means for each of the ability categories, while the six-item cumulative model would have larger differences in mean test scores for the middle ability levels than for extreme values.

The differences in mean score values for adjacent ability levels are shown in Table 4. The mean ability levels for each score are likewise shown in Tables 25 and 26. It was hypothesized from Lawley's work that the extreme scores of the cumulative test should have lower variance of ability level than the extreme scores for the sequential test.[2] Since less variance of ability level means fewer lower ability individuals, it was assumed the extreme cumulative test scores would have higher mean values.

Effect of Ability Distributions for Additional Sequential Tests

In addition to the four tests described above, three other sequential tests were built with an electronic computer.

---

[2]D. N. Lawley, "On Problems Connected with Item Selection and Test Construction," Proceedings of the Royal Society of Edinburgh, 61 (Section A, Part III): 273-287, 1942-1943.

However, in these tests the difficulties of the items were not determined by a "least squares" procedure, but used difficulties determined by an adaptation of Lord's work.[3] The item difficulties used in these three tests were so selected that, it was hypothesized, depending on the particular selection, a normal, rectangular, and a U-shaped distribution of scores would be obtained. The number of individuals assigned to each score and mean ability level of these individuals are reported in Tables 18, 19, and 20.

It was assumed that a score from a test designed to output a rectangular score distribution should correlate highest with a rectangular input of ability. Scores with normal distribution should likewise correlate highest with the normal input of ability, and scores with U-shaped distribution should correlate highest with U-shaped input of ability. However, information was obtained as to the effect on both output distribution and the correlation values of changing the input distribution.

The rule stated by Lord was that if one wished to divide the group at a given point, then the item difficulty (expressed in standard score units) is represented by the item-total r_{bis} times the standard score unit which represents the proportion below the point where the split is desired. The procedure followed in constructing these three tests was

---

[3]Lord, "Some Perspectives on 'The Attentuation Paradox in Test Theory'," op. cit.

that if there were four different difficulties used at a given stage, then the abcissa should be divided into five equal ability segments. The difficulties necessary to produce these proportions were them computed from Lord's formula. One time the distribution of scores to be produced was considered normal; one time, rectangular; and one time, U-shaped. Since different proportions were to be selected for each distribution shape, different difficulties were needed for each. The rule used to determine the number of different difficulties at each stage was to add one more difficulty at each stage. It turned cut that this rule gave results approximating the results from the determination of difficulties by the rules developed in the past section on "Test Model Construction."

Lord has shown how to select item difficulties to yield a desired split of individuals by a cumulative test. These Lord difficulties assume an input of a normal distribution of ability; therefore, in the sequential test one should compute difficulties with a normal distribution of ability for each item of the test. This was not possible in the present sequential model. The differences in the difficulty levels of the items selected by Lord's technique and the above technique when an $r_{bis} = .75$ is used are noted, but no study of the effect at other values of r_{bis} was made.

III. ITEM PRECISION AND DIFFICULTY FOR THE SEQUENTIAL TEST

To determine the interrelationships among item precision, difficulty level, and output characteristics, five tests containing items of varying precision and difficulty were compared. The five tests were built in the electronic computer and varied in precision and difficulty of items used. The tests were built using Lord's rule in the selection of difficulties so that a normal distribution of scores should be obtained when the distributions of ability were normal. The five precision levels were for r_{bis} equal to .79, .75, .71, .60, and .45. (The .75 precision test was the same as the one constructed above.) For an assumed N of 1000, the .79 and .71 values are one standard error of a r_{bis} above and below .75. The .60 value was selected as it is a value common in the literature; the .45 to show the effect of meeting low precision standards. The .79 precision level is not considered unrealistic if the spread of ability level is great. Precision was hypothesized to be one of the most important parameters in the behavior of the sequential test model.

To examine the hypothesis that the more precise items would produce a better separation of people, the variances of scores for category eight ability level (the middle ability level) individuals were compared for each of the five tests by use of Bartlett's test for homogeneity of variance. This

test was repeated for the combination of categories 14 and 15 (the most extreme categories) for the five test models. It was hypothesized that there would be a difference in the variance of scores, with the more precise items producing the scores with the smaller variances. Since a lower precision of items means that the effective difficulty level regresses toward the mean and, therefore, is closer to the 50 per cent level, the middle difficulty items should increase the precision of scores at the extremes--although not the ability to classify individuals. Thus, the extreme scores would have small variance of ability levels for both precise and less precise items and it was hypothesized that the variances of ability level scores would be most different at the middle score values.

The second hypothesis stated that a test consisting of more precise items would have the ability to discriminate evenly over the entire range of ability rather than making finer discriminations at the middle of the ability range. This hypothesis was tested by examining differences in the means of test scores for each category of ability. A table was made of the means and variances of test scores for each of the fifteen ability levels and for each of the five levels of precision. The discrimination index for adjacent ability levels was computed as suggested by Lord.[4] The higher the

---

[4]Lord, A Theory of Test Scores, op. cit., p. 24.

index the better the discrimination; values may range from zero to infinity. Lord's discrimination index was computed as follows:

$$D' = \frac{M_{s.c_0} - M_{s.c_1}}{\sigma^*}$$

$M_{s.c_0}$ = mean of score values for ability level co

$M_{s.c_1}$ = mean of score values for ability level c_1

σ^* = some appropriate average of the standard deviation of the two score distributions

Lord stated that this discrimination index is completely independent of the distribution of ability in the group tested:[5]

> This is an advantage when a general description of the test is desired without reference to any particular group of examinees; it is a disadvantage if the effective discrimination of the test for a specified group of examinees is desired.

IV. ERRORS IN SEQUENTIAL TEST PARAMETER ESTIMATES

The procedures used to determine the effects of errors in estimating the parameters of precision and difficulty for the sequential test items are related to the nature of the error involved. The difficulty of an item is usually specified in terms of the proportion of the group passing the item. This test model, however, uses difficulty specified in standard scores, so the standard error of a proportion must be translated into standard score terms. The standard

---

[5] Ibid.

error of a proportion ($\sqrt{(PQ)/N}$) is greatest when P = Q = .50. Thus the greatest error in estimating difficulty in terms of proportion passing an item would occur at the 50 per cent level of difficulty. The value of $\sqrt{(PQ)/N}$ is smallest at the extreme values of P or Q. The error in terms of proportion passing an item was thus investigated at .50 and .90. These errors were then translated into standard score units. The values of $\sqrt{(PQ)/N}$ (when N = 1000) were .016 and .010 for .50 and .90, respectively. When the values necessary to encompass two standard errors of the proportion were translated to standard score form the values were quite similar and equal to about $\pm$.10. The error for estimating difficulty was thus assumed to be less than or equal to $\pm$ .10 no matter what the difficulty level of the item.

The error made in precision depends upon the estimate of r_{bis}, which has a sampling error as follows:[6]

$$\sigma_{r_{bis}} = \frac{\sqrt{PQ}/Z - r_{bis}^2}{\sqrt{N}}$$

Terms as defined before.

Thus for r_{bis} equal to .75 (which was the only precision level for which the error was studied), and assuming P = Q = .50, and N = 1000; then $\sigma_{r_{bis}}$ = .02. Since the error in r_{bis} is not likely to be greater than $\pm$ .04, then r_{bis}

---

[6]Quinn McNemar, Psychological Statistics (second edition; New York: John Wiley and Sons, 1955), p. 194.

of .75 is not likely to be outside of the interval of .71 to .79. The σ_d value for .71 is .99 and the σ_d value for .79 is .78. Thus the error in terms of σ_d is not likely to be greater than $\pm$ .10. These estimates were the values used to determine the effect of parameter estimation on output.

The testing of the first hypothesis as to the effect of errors of item difficulty was done with a normal distribution of ability; test items designed for r_{bis} equal to .75; and by the least squares of deviations method described in "Test Model Construction." It was hypothesized that if one were to use at the second stage an item which was .40 more sigma units away from the mean than the items selected as above, then more people should be directed toward mean scores than if the ideal difficulty were used. This would imply fewer people at the extreme values than usual if the rest of the test did not correct this trend. It was hypothesized that the opposite should happen if the item were .40 sigma units toward the mean at the second stage. These changes were tested by use of the chi-square technique. If a difference of .40 did not make any difference it would seem obvious that errors of estimate (about .10) would not make any difference.

The errors of estimate in the fifth stage were determined when the item difficulties were shifted .40 sigma units away from the mean in one problem and .40 sigma units toward the mean on another problem. As the hypotheses on the effects

of error at the second and fifth stages derive from the same rationale,and as the effects of the fifth stage were expected to be in the same direction as the second stage effects only larger, the hypotheses on the second stage errors require only analysis of direction of change (i.e. chi-square) while the hypotheses on the fifth stage errors require more extensive variance analysis.

It was hypothesized that the variance of ability level for the top 84 individuals would be greater for tests with the shifted difficulties than for the test where the items were at the ideal difficulty level. The significance of differences in variances was tested by use of Bartlett's test for homogeneity of variance.

The discrimination of the tests for an ability level was determined by examining, for these same tests, the variance of test scores for the category fifteen ability level individuals. It was hypothesized that the variance of scores for category 15 individuals would be highest when difficulties were closest to the mean value. Variances for the three tests were compared by use of Bartlett's test for homogeneity of variance.

For the test with difficulties at the fifth stage displaced away from the mean by .40 sigma units, it was hypothesized from Lawley's work that the variance of ability level for the 100 middle-scoring individuals would be lower than the variance of these individuals on the other tests. Again

Bartlett's test for homogeneity of variance was used as the test.

Ability level discrimination was similarly determined by examination of the variance of test scores for category eight of ability. It was hypothesized that the original test (with ideal difficulties) would have better discrimination than the modified tests. Again Bartlett's test was used to compare variances.

The third hypothesis--that errors in estimating the precision of the items would be more serious in the initial stages than at later stages--was tested by placing items of $r_{bis} = .71$ (instead of $r_{bis} = .75$) at the second stage. Since subsequent items were designed with the assumption that the second item had $r_{bis} = .75$, the spread of ability should be greater than ideal for discriminating among individuals arriving at subsequent items. These subsequent items are more difficult than ideal and this increased difficulty should thus force the individuals toward the center of the distribution. The greatest increase in variance of test scores should thus be noticed for high and low ability groups; middle ability groups should not change in variance of test scores produced. The variances of scores for extreme and middle ability levels were compared by use of the F ratio. Also, the variances of ability level scores for individuals ranked in top 8.4 per cent of the score distribution were tested by the F ratio.

The fourth hypothesis that errors in estimate of precision should make little difference at the fifth stage was examined by placing items of r_{bis} equal to .71 at the fifth stage. The difficulty of the items remained the same. The effect of this should be that again the item would be more difficult than the Lord formula would suggest as ideal, because difficulty should be regressed toward the mean depending upon the r_{bis} value. The lower the r_{bis} the more the ideal difficulties should be regressed toward the mean. The results should be that more individuals than ideal would take an easier sixth item which, according to Lawley, should increase the precision of high ability scores. It was also hypothesized that this change in fifth item precision would increase the variance of score levels for high ability individuals. These results were hypothesized to be in the same direction as results from changes at the second stage, and the F ratio was likewise used to test these hypotheses.

V. GENERAL COMPARISONS

A general comparison of the relationship between input distribution and output distribution of scores was felt to be of value even though no specific hypotheses were advanced due to the number of variables involved. The difficulty of the items, the precision of the items and the pattern of items taken by individuals of different ability levels all interact to affect the score distribution.

The effect of difficulty of items was noted for the nine tests described in "Effect of Ability Distribution for Additional Sequential Tests." As the difficulties of items in each test do not regress toward the mean at the same rate, no clear conclusion can be made as to the effect of difficulty on output characteristics.

The effect of difficulty can thus be determined only for certain ability levels. (The data for the distributions of only one-half of the scores were presented as the other half was symmetrical.)

In addition to the distribution of scores, the correlation ratios were reported as these give information as to the general relationship between the input distribution of ability and the output distribution of scores. In former unpublished trials of the sequential test the value of the Pearson Product-Moment r was made to closely approach that for eta, by assigning the scores to the 64 different sequences of items from the rank of the mean ability level of the individuals at the score. (Another alternative would have been to assign scores according to rank of the sequence if ideal items had been used in the test model.)

The best general comparison of output to input in regard to precision of item came from the five sequential tests described in "Item Precision and Difficulty for the Sequential Test," where item difficulties and type of distribution remained constant over all five sequential tests. The general

comparisons were made in terms of correlation ratio; the data were reported for one-half of the output distribution of scores for the five tests.

A comparison of output to input in regard to the pattern of items taken by an individual came from using the Lord difficulties which yielded a rectangular output of scores when a rectangular distribution of ability was input. The rectangular distribution was used because this best approximated the "least squares" solution. Two new test models were constructed: each had exactly the same items with same difficulties and same precision ($r_{bis} = .75$); one test had items distributed as in Figure 1, and the other test had items distributed by one item at first stage, two items at the second, three at the third, and continued until it had six-items at stage six. Only the pattern of items taken by the individuals was different in the two tests. Again eta and the distribution of one-half of the output distribution of scores for each of the two test models were reported.

VI. SUMMARY OF PROCEDURES AND HYPOTHESES

One sequential test model was constructed by the "least squares" (of the deviations from the mean ability level) rule for a rectangular distribution of ability over 15 ability categories and r_{bis} equal to .75 for item precision. (Ability level one represented lowest ability level and ability level 15 represented highest ability level.)

The above test was then used with an input of normal and U-shaped distributions of ability. A six-item cumulative test with all items at the 50 per cent level of difficulty and a precision level of the item-total r_{bis} equal to .75 was likewise used with normal and U-shaped distributions of ability. The output distributions for comparable tests were then examined.

The null statistical hypotheses concerning the effect of the normal ability distribution on output of scores stated that the cumulative and sequential test models should have the following: (The alternative hypothesis expected from the rationale is given in parentheses.)

(1) equal means for the comparable normalized scores for category eight individuals (no alternate, hope to accept null);

(2) equal variances for the comparable normalized scores for category eight individuals (hope to accept null; cumulative may be smaller);

(3) equal means for the comparable normalized scores for combined category 14 and 15 individuals (cumulative lower);

(4) equal variances for the comparable normalized scores for combined category 14 and 15 individuals (sequential smaller);

(5) equal means for the ability level scores for the individuals ranked in the top 8.4 per cent of the score distribution (cumulative lower); and

(6) equal variances for the ability level scores for the individuals ranked in the top 8.4 per cent of the score distribution (sequential smaller).

The null statistical hypotheses concerning the effect of the U-shaped ability distribution on output stated that

the cumulative and sequential test models should have the following:

(1) equal means for the comparable normalized scores for category 13 individuals (cumulative lower);

(2) equal variances for the comparable normalized scores for category 13 individuals (sequential smaller);

(3) equal means for the comparable normalized scores for category 15 individuals (cumulative lower);

(4) equal variances for the comparable normalized scores for category 15 individuals (sequential smaller);

(5) equal means for the ability level scores for the individuals ranked in the top 13.5 per cent of the score distribution (cumulative lower); and

(6) equal variances for the ability level scores for the individuals ranked in the top 13.5 per cent of the score distribution (sequential smaller).

In addition to the hypotheses listed above, mean score values for each ability level, and mean ability level for each score value were plotted for both the normal and U-shaped distributions of ability. Additional information as to effect of distribution of input on output is presented as part of the general comparisons.

Three tests were constructed by Lord's rules and each of these was used with normal, rectangular, and U-shaped distributions of ability, although each test was designed to reflect only one of the input distributions. Eta was used to compare the input distribution with output distribution for these nine tests. In addition, the actual output distribution of each of the nine tests was tabled. These tests were built for information, and no hypotheses were made as to results.

To determine the effect of item precision on the output of the sequential test, four test models were constructed with an input of a normal distribution of ability and item precision taking the values of r_{bis} equal to .79, .71, .60, and .45. Item difficulties were those determined by Lord's procedure to be most appropriate for a given precision level when assuming a normal distribution of scores desired. The variances of ability levels for extreme and middle scores, and the variances of scores for extreme and middle ability levels were examined by use of Bartlett's test.

The null statistical hypotheses (and expected alternatives) concerning the effect of item precision and difficulty stated that tests which use a normal distribution of ability for input and a nearly normal output of scores should yield the following: (The alternative hypothesis is given in parentheses:)

(1) equal variances of scores for category eight ability level individuals for all five tests of different precision levels (most precise test smallest);

(2) equal variances of scores for category 14 and 15 ability level individuals for all five tests of different precision levels (most precise test smallest);

(3) equal variances of ability level scores for the individuals ranked in the top 8.4 per cent by each of the five tests of different precision levels (most precise test smallest); and

(4) equal variances of ability level scores for the individuals ranked in the middle 10 per cent by each of the five tests of different precision levels (most precise test smallest).

In addition to these hypotheses, the means and variances of the test scores, and the discrimination indices between each of the adjacent ability levels were computed for each of the five different precision tests.

To determine the effect of errors of using other than the difficulty level computed by "least squares" method for certain items, four sequential tests were constructed. One had the second item shifted away from the sample mean in difficulty; another had the second item toward the mean value. The fifth item encountered by the individual was likewise displaced toward or away from the mean difficulty value in the third and fourth test models, respectively. Again the characteristics of the "error" and "error free" output distributions were examined.

The null statistical hypotheses that were tested concerning the effect of errors in estimating the difficulty of the item at the second stage are as follows: (These hypotheses were used to determine if differences were in direction hypothesized.)

(1) the number of people in each of a set of score categories would be independent of whether distributed by an "error free" difficulty test or one in which difficulties at the second stage were away from the mean (50 per cent) difficulty (more people at middle for "error" test);

(2) the number of people in each of a set of score categories would be independent of whether the people were distributed by an "error free" difficulty test or one in which difficulties at the second stage were toward the mean (50 per cent) level of difficulty (more people at extreme for "error" test).

The null statistical hypotheses that were tested concerning the effect of errors in estimating the difficulty of the item at the fifth stage predicted the following: (These hypotheses were deduced from same rationale as ones above, and data were examined more closely as it was hypothesized that those differences would be in same direction as differences above and of a larger magnitude.)

(1) equal variances for the ability level scores for the individuals ranked in the top 8.4 per cent of the score distribution ("error free" test smallest);

(2) equal variances of test scores for individuals in ability category 15 (test with items near 50 per cent largest);

(3) equal variances of ability level scores for the individuals ranked in the middle 10 per cent of the score distribution (test with items away from mean smallest); and

(4) equal variances of test scores for individuals in ability category 8 ("error free" test smallest).

The effect of error in estimating the precision of items was examined by constructing two additional "least squares" test models. One test had less precise items for the second item encountered; the other had less precise items substituted for the fifth item encountered. Again the distributions of scores for the "error" and "error free" tests were examined.

The null statistical hypotheses concerning the effect of error in estimating the precision of items at the second stage predicted the following:

(1) equal variances of test scores for individuals in ability category 15 ("error free" test smaller);

(2) equal variances of test scores for individuals in ability category 8 ("error free" test smaller); and

(3) equal variances of ability level scores for individuals ranked in top 8.4 per cent of the score distribution ("error free" test smaller).

The null statistical hypotheses concerning the effect of error in estimating the precision of items at the fifth stage predicted the following:

(1) equal variances of test scores for individuals in ability category 15 ("error free" test smaller); and

(2) equal variances of ability level scores for individuals ranked in top 8.4 per cent of the score distribution ("error free" test smaller).

The general comparison examined the effect of difficulty on score output, the effect of precision of items, and the effect of the pattern of items. Difficulty effects were examined for normal, rectangular, and U-shaped inputs on tests with item precision of r_{bis} equal to .75 and item difficulties as listed in Table 20 of the Appendix. (The rule for selection of difficulties of items is that one should use an item not at the difficulty level equal to ability level where split between groups is desired, but difficulty level should be regressed toward the mean value of 50 per cent. The lower the r_{bis} the greater should be the regression.) The distributions and mean ability level scores for each score were tabled.

Distributions and mean scores were also tabled for five tests with different item precision and for two tests with different patterns of items. In addition to these tables eta

between input and output scores was reported for each of these tests.

CHAPTER IV

ANALYSES AND RESULTS

There are six sections to this chapter. Section one gives the results of building the six-item sequential test model. Section two reports results of the input distribution on the score distribution of both the sequential and the six-item cumulative test models. Section three presents the effects of item precision and difficulty on the score distribution of the six-item sequential test model. Section four gives the effects of errors of estimating precision and difficulty parameters on the score distribution of the sequential test. Section five gives some general results of changes in difficulty of items, precision of items, and pattern of items. Section six is a summary of the analyses and results. In all sections results are simply reported; interpretation is reserved for Chapter V.

I. SEQUENTIAL TEST CONSTRUCTION

As stated in Chapter III, the sequential test model was constructed so that the $\Sigma (\Sigma X)^2/N$ was maximized; graphic methods were used to aid in determination of maximum values. (ΣX refers to sum of ability level scores for any one group.

$\Sigma\ (\Sigma X)^2/N$ refers to squaring the sum of scores for the group dividing by the number in the group and then summing over the two or more groups that used the particular item.) The only restriction was that any item difficulty had to be more than .20 standard score units away from other difficulties to be considered different from them, and thus to be used. (The reader will be aided in following the item decisions given below by referring to Figure 3.)

First Item Decision

The values of $\Sigma\ (\Sigma X)^2/N$ for + .01, .00, and -.01 difficulty items were as follows: 109073.85, 179931.86, and 109073.85. The maximum value was thus obtained from a .00 difficulty level item and this item fulfilled the criterion of selection. Thus out of the 1500 people taking the hypothetical test, 750 would pass and 750 would fail this item. The mean ability level of these groups was $\pm$ .73.

Second Item Decision

The second item produced four groups over which $\Sigma\ (\Sigma X)^2/N$ was maximized. The three strategic values for this item were $\pm$.23, $\pm$.24, and $\pm$.25 which had values for $\Sigma\ (\Sigma X)^2/N$ of 113796.15, 113796.21, and 113796.00. (Strategic values were determined by estimating values and plotting these values of $\Sigma\ (\Sigma X)^2/N$ until the maximum value was stradled by three points that could be read from the graph.) The $\pm$.24 items were selected for the second stage. The resulting four

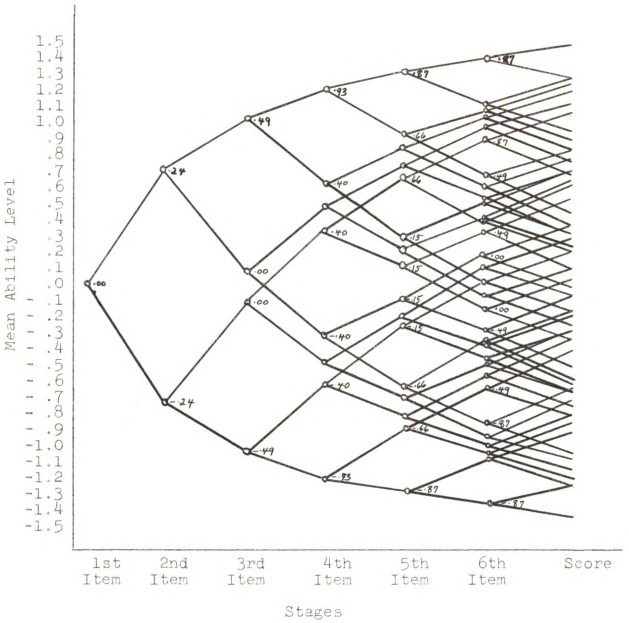


Fig. 3.--Mean Ability Level of Groups Separated Out by Sequential Test and Difficulties of Items Used

groups had mean ability levels of +1.04, +.10, -.10, and -1.04. At this point 504 individuals had passed both the first and second items; 246 had passed the first and failed the second; and like numbers had failed both, and the first only.

Third Item Decision

The third stage items were reduced to three in number as the two middle groups were both given the same difficulty. Both of these middle groups took the same difficulty because each has the sum of $(\Sigma X)^2/N$ of 32878.55 for $\pm$.09 items and 32878.02 for $\pm$.10 items. As $\Sigma (\Sigma X)^2/N$ maximized at less than .10, the ideal difficulty levels would be less than .20 sigma units apart. As this would violate a condition of the test construction, the two middle groups were given the same item which yielded a $\Sigma (\Sigma X)^2/N$ of 32885.60.

The two extremes ability groups produced $\Sigma(\Sigma X)^2/N$ equal to 83416.29, 83417.00, and 83416.84 for .48, .49, and .50 difficulty items, respectively. Thus the three difficulty levels used at the third stage are +.49, .00, and -.49. The mean ability levels of the eight resulting groups were from highest to lowest 1.21, .62, .48, .33, -.33, -.48, -.62, and -1.21.

Fourth Item Decision

At this stage there were eight groups taking four different difficulty items ($\pm$.73 and $\pm$.40) and resulting in

sixteen groups. Those individuals who had passed (or failed) the first three items had $\Sigma (\Sigma X)^2/N$ equal to 62860.59, 62860.69, and 62860.50 for $\pm.72$, $\pm.73$, and $\pm.74$ items respectively. The $\pm.73$ item difficulty was selected. The second group (PPQ or QQP) had maximum values between $\pm.45$ and $\pm.50$ which were more than .20 standard score units away from $\pm.73$. However, the third group (PQP or QPQ) had $\Sigma (\Sigma X)^2/N$ that maximized above .32. The similarity of the groups is shown in that while a .32 maximum is 18242.13, the .41 maximum is 18243.05. Since such values would give items less than .20 standard deviation units away, the second and third groups were each given the same difficulty. The remaining group (QPP or PQQ) maximized between .29 and .35 for $\Sigma (\Sigma X)^2/N$ of 15387.12 and 15386.92, respectively. Since the best difficulty level for the previous two groups would be less than .20 standard deviation units away, all three groups were given an item of the same difficulty level. The strategic values for difficulty of item assigned to the three groups were $\pm.39$, $\pm.40$, and $\pm .41$ which had $\Sigma (\Sigma X)^2/N$ values of 54983.63, 54983.81, and 54983.57. Thus the $\pm.40$ item difficulties were used. Of the eight groups at this stage, one group took +.73, three took +.40, three took -.40 and one took an item of -.73 difficulty level.

Fifth Item Decision

The fifth stage decisions resulted in sixteen groups taking six items of different difficulty thus producing thirty-

two new groups. The groups that took the different difficulty levels were as follows: The PPPP and QQQQ groups took items of +.87 and -.87 difficulty. The PPPQ, PPQP, PQPP, and QPPP groups took an item of +.66 level of difficulty. (The QQQP, QQPQ, etc. opposites of above took an item at -.66.) The PPQQ, PQPQ, and QPPQ groups each took an item of +.15 difficulty level. (Opposite groups took -.15 difficulty item.) In other words for the eight groups above the mean, one group took an item of .87 difficulty, four groups took an item of .66, and three groups took an item of .15 level of difficulty.

The PPPP (and opposite) had $\Sigma (\Sigma X)^2/N$ values of 45791.18, 45791.20, and 45791.18 for .86, .87, and .88 levels of difficulty. The PPPQ group maximized the $\Sigma(\Sigma X)^2/N$ just above the .71 difficulty level, thus the decision had to be made to give this group either the same difficulty item as the PPPP group or the difficulty of the PPQP group. The PPQP group maximized between .67 and .71-- $\Sigma (\Sigma X)^2/N$ values of 13411.14 and 13411.11, respectively. These two groups were thus given the same difficulty level as their curves remained fairly near maximum for the difficulty level common to both. The PQPP group maximized $\Sigma (\Sigma X)^2/N$ between .60 and .65 with values of 10395.92 and 10395.98. the QPPP group maximized at about .60 with $\Sigma (\Sigma X)^2/N$ of 7826.26. Since none of these was .20 standard score units apart in difficulty, the one difficulty value that would maximize $\Sigma(\Sigma X)^2/N$ for all four groups was determined.

The difficulties of .65, .66, and .67 had $\Sigma(\Sigma X)^2/N$ values of the eight groups of 49000.44, 49000.65, and 49000.60. The item of $\pm$.66 difficulty level was thus used for these eight groups.

The PPQQ group maximized $\Sigma(\Sigma X)^2/N$ values between .20 and .28--8208.84 and 8298.88, respectively. This was more than .20 standard deviation units from .66, so this group was not given the item of .66 difficulty level. The PQPQ group maximized $\Sigma(\Sigma X)^2/N$ at .15 with 8096.18. (Difficulty levels .10 and .14 had $\Sigma(\Sigma X)^2/N$ values of 8096.15 and 8096.17, respectively.) The QPPQ group maximized between .00 and .10 difficulty levels. This was not .20 standard score units of difficulty away, so the one difficulty level that would maximize the sum of $(\Sigma X)^2/N$ for these six groups was determined. The strategic difficulty levels of .14, .15, and .16 had $\Sigma(\Sigma X)^2/N$ for six groups of 24092.29, 24092.36, and 24092.30.

Sixth Item Decision

The sixth stage had 32 groups taking items at five different difficulty levels ($\pm$.87, $\pm$.49, and .00). The PPPPP group had maximized $\Sigma(\Sigma X)^2/N$ between .90 and 1.00 difficulty--the respective $\Sigma(\Sigma X)^2/N$ values are 32852.58 and 32852.64. The group PPPPQ had $\Sigma(\Sigma X)^2/N$ values of 13051.00, 13051.01, and 13051.00 at .86, .87, and .88, respectively. Thus it was clear that these two would not use different difficulty of item and neither would any group that maximized

above .75. The other groups which maximized about.75 were as follows: the PPPQP group which for .85, .86, .87, and .88 had $\Sigma (\Sigma X)^2/N$ values of 11334.16, 11334.17, 11334.17, and 11334.16, respectively; the PPQPP group which for .85, .86, .87, and .88 had 8373.86, 8373.86, 8373.86, and 8373.84, respectively; the PQPP group which maximized $\Sigma (\Sigma X)^2/N$ between .80 and .85 with values of 6059.83 and 6059.79, respectively; and the QPPPP group which maximized between .74 and .80 both with $\Sigma (\Sigma X)^2/N$ value of 4227.53.

The $\Sigma (\Sigma X)^2/N$ for the 12 groups using the same difficulty level of item were 75898.65, 75898.66, and 75898.60 for the .86, .87, and .88 level of difficulty, respectively. The decision was thus to use a .87 difficulty item for these groups.

The PPPQQ group (the next highest ability level group) maximized between .55 and .65 with $\Sigma (\Sigma X)^2/N$ values of 6150.61 and 6150.64, respectively. (The approximate value for maximum was determined by plotting of the curve from six points.) Since the group maximized more than .20 standard deviation units away from the .87 groups and also maximized within five points of the next lowest group, the decision was made to use a new difficulty for all remaining groups that maximized above .40. The remaining groups which maximized at difficulty levels greater than .40 (but below.60) were as follows: The PPQPQ group which for difficulty levels of .50, .55, and .65, had $\Sigma (\Sigma X)^2/N$ values of 5135.92, 5136.00, and

5135.88, respectively; the PQPPQ group which for difficulty levels of .43, .48, .49, and .50 had $\Sigma(\Sigma X)^2/N$ values of 4422.34 4422.41, 4422.41, and 4422.40, respectively; the PPQQP group which for difficulty levels of .43, .48, .49, and .50 had $\Sigma(\Sigma X)^2/N$ values of 4680.27, 4680.33, 4680.33, and 4680.32, respectively; the PQPQP group which for difficulty levels of .32, .43, and .48 had $\Sigma(\Sigma X)^2/N$ values of 4198.70, 4198.83, and 4198.77, respectively; and the QPPPQ group which for .32, .43, and .48 had $\Sigma(\Sigma X)^2/N$ values of 3670.06, 3670.19, and 3670.14, respectively.

The QPPQP group maximized $\Sigma(\Sigma X)^2/N$ between .32 and .43. Difficulty levels of .29, .32, and .43 had values of 3628.46, 3628.48, and 3628.35, respectively. A decision thus was whether to include this group with the higher or lower groups. The PQQPP group (next in line) for difficulty levels of .00 and .09 had $\Sigma(\Sigma X)^2/N$ values of 4244.60 and 4243.82 and maximized below .09. For this reason the QPPQP group was included with the higher group instead of .20 units lower in difficulty which would have yielded a lower $\Sigma(\Sigma X)^2/N$ value.

The sum of $\Sigma(\Sigma X)^2/N$ for the 14 groups for difficulty levels of .48, .49, and .50 were 31886.02, 31886.04, and 31885.95. Thus a difficulty of $\pm$.49 was used with each of these groups.

The remaining six groups all maximized between $\pm$.09, thus .00 item was used here. The QPPQQ group for difficulty

levels of .00 and .09 had $\Sigma(\Sigma X)^2/N$ values of 4244.60 and 4243.82, respectively. The PQPQQ group had 3983.96 and 3983.54 for these same values, and the PPQQQ group for difficulty levels of .00 and .09 had $\Sigma(\Sigma X)^2/N$ values of 3615.52 and 3615.32, respectively.

Thus of the 16 groups above the mean, six groups took the .87 difficulty item, seven groups took the .49 difficulty item, and three groups took the .00 difficulty item at the final stage.

The above sequential test was compared with the cumulative test to determine how well the score differentiated individuals of different ability levels and to determine the range of ability levels assigned to any one score.

The above sequential test was also used in the determination of the effects of errors in estimating the parameter values for the items in this test. Parameter values considered were difficulty and precision.

This pattern of items determined above was also used with different difficulties to determine how a test with an arbitrary pattern and easily computed difficulties compared with a test using pattern of items determined above.

II. INPUT DISTRIBUTION EFFECTS

Normal and U-shaped distributions were each used with the cumulative and the "least squares" sequential test. The results from the two distributions are presented separately.

Results from the Normal Distribution

The first null hypothesis was that there should be equal means of the comparable normalized scores for the middle category, category eight, individuals taking the sequential and cumulative test both with a normal distribution of ability input. Results are shown in Table 1. As can be seen from the table, the null hypothesis tested by a "t" test must be accepted. This was expected as both have a symmetrical distribution of scores. This hypothesis was included as a parallel hypothesis to hypothesis one on U-shaped distribution (and as a check on the accuracy of computer computations). In this, and all other hypotheses, the reader should be aware of the fact that the number of individuals is dependent only upon the accuracy of the calculations. Since the figures were carried to between eight and twelve places a larger N could well be assumed. This would make the error terms smaller and differences significant. The theoretical 1000 individuals were used to give the reader a point of reference. If the differences exist in the proper direction, the rationale may be said to be supported.

The second null hypothesis was that there would be equal variances of the comparable normalized scores for middle category (number 8) individuals taking the sequential test and the cumulative test both with a normal distribution of ability input. Results are shown in Table 1. Again

the null hypothesis tested by a F ratio test must be accepted. This was expected from the rationale.

The third null hypothesis was that there should be equal means of the comparable normalized scores for combined category 14 and 15 individuals taking the sequential and cumulative tests both with a normal distribution of ability input. Results are shown in Table 2. The null hypothesis was based upon 1000 individuals and accepted. The scores were in the expected direction with the sequential test assigning the more extreme value; therefore, the rationale tends to be supported.

The fourth null hypothesis was that there should be equal variances of the comparable normalized scores for combined category 14 and 15 individuals taking the sequential and cumulative tests both with a normal distribution of ability input. Results are shown in Table 2. The null hypothesis was rejected at the .01 level of significance. The sequential test had lower variance for high ability individuals as was predicted from the rationale.

The fifth null hypothesis was that there should be equal means of ability level scores for the individuals in the top 8.4 per cent of the score distributions taking the sequential and cumulative tests both with a normal distribution of ability input. Results are shown in Table 3. The null hypothesis was rejected at the .01 level of significance. The sequential test had a higher mean ability level for the

TABLE 1

ANALYSIS OF MEANS AND VARIANCES OF NORMALIZED SCORES FOR CATEGORY 8 INDIVIDUALS WHEN NORMAL DISTRIBUTION OF ABILITY IS INPUT INTO SEQUENTIAL AND CUMULATIVE TEST MODELS

Parameter	Sequential Test	Cumulative Test	Significance Between Tests
Mean	50.00	50.00	n.s.
Variance	16.37	21.22	n.s.

TABLE 2

ANALYSIS OF MEANS AND VARIANCES OF NORMALIZED SCORES FOR CATEGORY 14 AND 15 INDIVIDUALS WHEN NORMAL DISTRIBUTION OF ABILITY IS INPUT INTO SEQUENTIAL AND CUMULATIVE TEST MODELS

Parameter	Sequential Test	Cumulative Test	Significance Between Tests
Mean	63.40	62.96	n.s.
Variance	3.87	6.77	$p < .01$

TABLE 3

ANALYSIS OF MEANS AND VARIANCES OF ABILITY LEVEL SCORES FOR THE TOP 8.4 PER CENT OF THE SCORE DISTRIBUTION WHEN NORMAL DISTRIBUTION OF ABILITY IS INPUT INTO SEQUENTIAL AND CUMULATIVE TESTS

Parameter	Sequential Test	Cumulative Test	Significance Between Tests
Mean	13.66	12.92	$p < .01$
Variance	2.35	3.47	$p < .05$

top 8.4 per cent of the score distribution as had been predicted.

Null hypothesis six was that there should be equal variances of ability level scores for the individuals in the top 8.4 per cent of score distributions taking sequential and cumulative tests, both with a normal distribution of ability input. The results are shown in Table 3. The null hypothesis was rejected at the .05 level of significance. The sequential test had smaller variance of ability level scores for the top 8.4 per cent of the score distribution as had been predicted.

To examine the hypothesis that the six-item cumulative test model would have smaller differences in mean ability levels between the middle and adjacent scores than between the extreme and adjacent scores, the differences in mean ability level for adjacent scores were computed. These differences are reported in Table 5, column 3. As was hypothesized, the smaller differences in ability level were between the middle score 4, and the adjacent score 5. However, it should be noted that the differences between ability level scores for adjacent scores for the sequential test model (shown in Table 6) were not equal interval and there is no pattern to the differences shown, although in both cases the differences were greatest for the extreme scores.

If one wishes to examine the mean ability level and number of individuals at each score, these values are shown

TABLE 4

DIFFERENCES BETWEEN NORMALIZED "T" SCORES FOR ADJACENT TOP ABILITY LEVELS FOR NORMAL AND U-SHAPED INPUT

Between Ability Levels	Ideal Difference	Normal Input		U-Shaped Input	
		Cumulative	Sequential	Cumulative	Sequential
15-14	4.5	2.3	2.5	1.2	2.6
14-13	2.5	1.1	2.1	1.1	1.7
13-12	2.5	1.9	2.1	1.5	1.8
12-11	2.5	2.0	2.3	1.7	1.7
11-10	2.4	2.4	2.3	1.7	1.4
10- 9	2.5	2.3	2.3	1.6	1.4
9- 8	2.5	2.5	2.3	1.6	1.2

TABLE 5

DIFFERENCES BETWEEN ABILITY LEVEL SCORES FOR ADJACENT TOP SCORES FOR CUMULATIVE TEST MODEL FOR NORMAL AND U-SHAPED INPUT

Between Scores*	Ideal Difference	Input	
		Normal	U-Shaped
7-6	2.33	2.1	1.9
6-5	2.33	1.5	2.0
5-4	2.33	1.3	2.0

*Scores range from 1-7.

TABLE 6

DIFFERENCES BETWEEN ABILITY LEVEL SCORES FOR ADJACENT TOP SCORES FOR SEQUENTIAL TEST MODEL FOR NORMAL AND U-SHAPED INPUT

Between Scores*	Input		Between Scores*	Input	
	Normal	U-Shaped		Normal	U-Shaped
64-63	1.4	.8	48-47	.3	.1
63-62	.0	.0	47-46	.3	.3
62-61	.2	.1	46-45	.2	.5
61-60	.1	.2	45-44	.3	.2
60-59	.3	.2	44-43	.2	.0
59-58	.1	.2	43-42	.0	.4
58-57	.6	.6	42-41	.0	.1
57-56	.3	.1	41-40	.0	-.1
56-55	.0	.2	40-39	.5	.5
55-54	.0	.1	39-38	-.4	-.1
54-53	.5	.2	38-37	.4	.1
53-52	.0	.0	37-36	.0	.0
52-51	.0	.3	36-35	.0	.2
51.50	.0	.1	35-34	.0	.2
50-49	.2	-.1	34-33	.3	.3
49-48	.0	.4			

*Ideal difference if all had been equal intervals would be .22.

in Tables 25 and 26 of the Appendix. The mean normalized "T" score for each ability level is reported in Table 24 of the Appendix.

Results from the U-Shaped Distribution

The first null hypothesis was that there should be equal means of the comparable normalized scores for category 13 individuals taking the sequential and cumulative tests both with an input of a U-shaped distribution of ability. Results are shown in Table 7. As can be seen, the null hypothesis must be accepted. The sequential test did have the higher mean value as expected, but not significantly so if 1000 individuals are assumed to have taken the test. Rationale would tend to be supported though the effect is small. (See comments on size of N under "Results from Normal Distribution.")

The second null hypothesis was that there should be equal variances of the comparable normalized scores for category 13 individuals taking the sequential and cumulative tests each with an input of a U-shaped distribution of ability. From Table 7 one can determine that the null hypothesis must be accepted if only 1000 individuals are assumed to have taken the test. The variance of the sequential test was less, however, than the cumulative test as anticipated though the effect was small.

The third null hypothesis was that there should be equal means of the comparable normalized scores for category

TABLE 7

ANALYSIS OF MEANS AND VARIANCES OF NORMALIZED SCORES FOR CATEGORY 13 INDIVIDUALS WHEN A U-SHAPED DISTRIBUTION OF ABILITY IS INPUT INTO SEQUENTIAL AND CUMULATIVE TEST MODELS

Parameter	Sequential Test	Cumulative Test	Significance Between Tests
Mean	58.44	58.03	n.s.
Variance	13.99	14.00	n.s.

TABLE 8

ANALYSIS OF MEANS AND VARIANCES OF NORMALIZED SCORES FOR CATEGORY 15 INDIVIDUALS WHEN A U-SHAPED DISTRIBUTION OF ABILITY IS INPUT INTO SEQUENTIAL AND CUMULATIVE TEST MODELS

Parameter	Sequential Test	Cumulative Test	Significance Between Tests
Mean	60.73	60.44	n.s.
Variance	1.96	3.62	$p < .01$

TABLE 9

ANALYSIS OF MEANS AND VARIANCES OF ABILITY LEVEL SCORES FOR THE TOP 13.5 PER CENT OF THE SCORE DISTRIBUTION WHEN A U-SHAPED DISTRIBUTION OF ABILITY IS INPUT INTO SEQUENTIAL AND CUMULATIVE TESTS

Parameter	Sequential Test	Cumulative Test	Significance Between Tests
Mean	14.43	13.87	$p < .01$
Variance	.77	1.86	$p < .01$

15 individuals taking the sequential and cumulative tests both with an input of a U-shaped distribution of ability. The results are shown in Table 8. The null hypothesis must be accepted, although the results were in the direction indicated by the research hypothesis. The cumulative had a lower value for the mean. Again significance depends upon number of individuals assumed to have taken the test.

The fourth null hypothesis was that there should be equal variances of the comparable normalized scores for category 15 individuals taking the sequential and cumulative tests both with an input of a U-shaped distribution of ability. As shown in Table 8, the null hypothesis was rejected at the .01 level of significance. The sequential test had less variance of scores for the highest ability level individuals than did the cumulative test.

The fifth null hypothesis was that there should be equal means of ability level scores for the individuals in the top 13.5 per cent of the score distribution taking the sequential and cumulative tests both with an input of a U-shaped distribution of ability. The results are shown in Table 9. The sequential test had a significantly higher mean ability level for the top 13.5 per cent of the score distribution than did the cumulative. This was in the direction hypothesized.

The sixth hypothesis was that there should be equal variances of ability level scores for the individuals in the top 13.5 per cent of the score distribution taking the sequential and cumulative tests both with an input of a U-shaped

distribution of ability. The results in Table 9 indicate that the sequential test had at the .01 level of significance, a smaller variance of ability level scores for the top 13.5 per cent of the score distribution than did the cumulative test. This was in the direction hypothesized.

The difference in mean ability level between adjacent top scores for the cumulative and sequential test models are shown in Tables 5 and 6, respectively. The scores on the sequential test did not yield equal intervals on the ability level scale as had been hypothesized. The cumulative scores are a good approximation of equal intervals on the ability level scale.

To examine the hypothesis that the sequential test model should have approximately equal distance between test score means for each of the ability categories, while the six-item cumulative would have larger differences in mean test scores for the middle ability levels than for extreme ability levels, the differences between adjacent scores were computed. These differences are reported in Table 4. The cumulative test did have smaller score differences between the extreme ability levels than any other point in ability distribution. However, the sequential test did not have an equal interval scale, but in general decreased in size of difference between mean scores of adjacent ability levels from extreme ability category to middle ability category. It should be noted that neither test represented the ability

levels with any real accuracy. The top ability level shown had an ideal "T" score of 69 instead of the 61.8 assigned by the sequential or the 60.4 assigned by the cumulative test. (See Table 24.)

III. ITEM PRECISION AND DIFFICULTY FOR THE SEQUENTIAL TEST

Five levels of precision and the appropriate levels of difficulty for each were used in the construction of five sequential test models. For these tests the variances of scores for the extreme and middle ability levels and the variances of ability level for extreme and middle scores were examined.

Variance of Scores

The first null hypothesis was that there would be equal variances of scores for category 8 ability level individuals for all five tests of different precision level. Data and results are shown in Table 10. The null hypothesis was rejected at the .001 level of significance. As was hypothesized, the more precise tests had smaller variances.

The second null hypothesis was that there would be equal variances of scores for a combination of ability level categories 14 and 15 for all five tests of different precision level. From data in Table 10 it can be seen that the null hypothesis was rejected at the .001 level of significance; the more precise tests had smaller variances of scores.

TABLE 10

ANALYSIS OF THE VARIANCE OF SCORES FOR INDIVIDUALS AT SPECIFIED ABILITY LEVELS FOR FIVE TESTS OF DIFFERENT PRECISION

Ability Category	Precision of Test					Significance of Difference
	.45	.60	.71	.75	.79	
8	260.03	198.70	147.52	127.33	111.65	$p < .001$
14 and 15	94.74	40.90	19.69	15.50	11.86	$p < .001$

TABLE 11

ANALYSIS OF THE VARIANCE OF ABILITY LEVEL SCORES FOR INDIVIDUALS AT SPECIFIED SCORE LEVELS FOR FIVE TESTS OF DIFFERENT PRECISION

Score Level (Per Cent)	Precision of Test					Significance of Difference
	.45	.60	.71	.75	.79	
Top 8.4	5.88	3.71	2.45	2.10	1.77	$p < .001$
Middle 10	8.22	5.37	3.62	3.06	2.56	$p < .001$

As was hypothesized, the precision of the item was an important variable in precision of scores.

Variance of Ability Levels

The third null hypothesis was that there would be equal variances of ability level scores for the individuals ranked in the top 8.4 per cent of the score distribution by each of the five tests of different precision level. Data and results are shown in Table 11. The null hypothesis was rejected at the .001 level of significance. As can be seen, the precision of item was important in determining the precision of the scores as hypothesized. The individuals assigned to the top 8.4 per cent of the score distribution were not as variable in ability level when assigned by a test with items having an r_{bis} of .79 as when assigned by a test with items having an r_{bis} of .45.

The fourth null hypothesis was that there would be equal variances of ability level scores for the individuals ranked in the middle 10 per cent of the score distribution by each of the five tests of different precision level. As can be seen in Table 11, the null hypothesis was rejected at the .001 level of significance. The results were in the direction hypothesized--the more precise tests had smaller variance of ability levels. However, it should be noted that for the middle 10 per cent of the score distribution, the variances were 8.22 and 2.56 for the .45 and .79 tests, respectively. The one variance is 3.21 times larger than the

other. For the top 8.4 per cent of the score distribution the variances were 5.88 and 1.77 for the .45 and .79 tests, respectively. The larger variance is 3.32 times the other. Greater differences at the top than at the middle of the score distribution were contrary to what had been expected.

Table 12 gives the means and variances of rank scores assigned to each ability level by the five tests of different precision. The means for category 8 individuals were always the same. However, the mean rank scores assigned to category 1 individuals were lower as the precision of the item increased. This was especially noticeable at the lower precision levels. The variances of the test scores for each ability level decreased with the precision of item as was hypothesized. Also, it should be noted that the variances of extreme scores were much lower than the variances of the middle value scores.

The discrimination indices are reported in Table 13. (Only one-half of the score distribution is tabled because the two halves are symmetrical about the mean.) The higher the value, the better the discrimination. The test consisting of the most precise items had the highest discrimination index. The test was more discriminating for the extremes in ability than it was for the other ability values. However, the other values of the discrimination index were remarkably close to each other for all ability levels other than the extremes. This was what had been hoped for with the sequential test.

TABLE 12

THE MEANS AND VARIANCES OF RANK SCORES ASSIGNED TO EACH ABILITY LEVEL BY FIVE TESTS OF DIFFERENT PRECISION

Ability Level	Mean of Rank Scores for Different Precision Levels					Variances of Rank Scores for Different Precision Levels				
	.45	.60	.71	.75	.79	.45	.60	.71	.75	.79
1	8.4	5.1	3.6	3.2	2.8	72.2	25.8	11.0	7.8	5.9
2	12.4	8.6	6.5	6.0	5.4	123.8	58.9	29.0	23.5	17.6
3	15.1	11.2	9.2	8.4	8.0	159.6	83.3	49.1	38.6	31.6
4	18.0	14.5	12.5	11.8	11.3	188.8	116.0	72.2	58.8	49.5
5	21.4	18.4	16.5	16.0	15.5	218.3	146.3	97.4	82.9	68.7
6	25.0	22.7	21.4	21.1	20.7	241.9	171.8	124.3	107.0	89.6
7	28.7	27.6	26.8	26.4	26.6	255.1	192.5	141.9	123.4	104.3
8	32.5	32.5	32.5	32.5	32.5	260.0	198.7	147.5	123.7	111.7
9	36.3	37.4	38.2	38.4	38.6	255.1	192.5	141.9	123.4	104.3
10	40.0	42.3	43.6	43.9	44.3	241.9	171.8	124.3	107.0	89.6
11	43.6	46.6	48.5	49.0	49.5	218.3	146.3	97.4	82.9	68.7
12	47.0	50.5	52.5	53.2	53.7	188.8	116.0	72.2	58.8	49.5
13	49.9	53.8	55.8	56.6	57.0	159.6	83.3	49.1	38.6	31.6
14	52.6	56.4	58.5	59.0	59.6	123.8	58.9	29.0	23.5	17.6
15	56.6	59.9	61.4	61.8	62.2	72.2	25.8	11.0	7.8	5.9

TABLE 13

THE DISCRIMINATION INDICES BETWEEN ADJACENT ABILITY LEVELS FOR THE INPUT OF A NORMAL DISTRIBUTION OF ABILITY INTO TESTS OF DIFFERENT PRECISION

Between Ability Levels	Precision Level of Test				
	.45	.60	.71	.75	.79
1 and 2	.42	.56	.69	.76	.81
2 and 3	.22	.30	.42	.43	.50
3 and 4	.22	.33	.43	.47	.51
4 and 5	.24	.34	.43	.50	.54
5 and 6	.23	.34	.46	.52	.57
6 and 7	.24	.36	.46	.52	.59
7 and 8	.24	.35	.48	.52	.58

IV. ERRORS IN THE SEQUENTIAL TEST PARAMETER ESTIMATES

Errors in estimating the difficulty level of items and errors in estimating the precision of items were investigated. Four different tests with errors in estimates of difficulty were constructed, and two tests with errors in precision were built. All tests used the "least squares" difficulties as the base for comparison. The results of investigating these two types of errors will be discussed separately.

Errors in Estimating Difficulty

Of the four tests with errors in estimates of item difficulty, two had the error at the second item encountered and two had the error at the fifth item encountered.

Second item error.--The first null hypothesis was that the number of people in each set of score categories would be independent of whether the people were classified by an "error free" test or one which had items too far from the mean at the second stage. The distributions are reported in Table 27 of the Appendix. The number of individuals at 12 selected categories, the expected values from an independence assumption, and the chi-square value are reported in Table 14. The null hypothesis had to be accepted. There were more people at the middle values as hypothesized, but the differences were not significant if 1000 people were

assumed to have taken the test. It can be concluded that the effects of second-item errors are small.

The second null hypothesis was that the number of people in each set of score categories would be independent of whether the people were classified by an "error free" test, or by a test which had the second item encountered too near the mean value. The distribution is reported in Table 27 in the Appendix. The number of individuals at 12 selected categories, the expected values from an independence assumption, and the chi-square value are reported in Table 15. The null hypothesis had to be accepted However, there were more people at the extreme categories in the modified test than in the "error free" test, as was hypothesized. The differences were not significant due to the assumption of 1000 individuals.

Fifth item error.--The first null hypothesis was that of equal variances of the ability level scores for the individuals ranked in the top 8.4 per cent of the score distribution by the "error free" difficulty test and the tests which had the fifth item too far and too near the mean value. The variances of the ability level scores for the top 8.4 per cent in each of the tests are reported in Table 16. The differences in variances were not significantly different from each other. However, the "error free" test did not have the smallest variance as was hypothesized. The test with

TABLE 14

DISTRIBUTION OF INDIVIDUALS BY TWO TESTS--ONE TEST WITH SECOND ITEM DIFFICULTIES FARTHER FROM 50 PER CENT LEVEL THAN THE "ERROR FREE" TEST*

Test	Rank Scores					
	64	58-63	54-57	46-53	40-45	33-39
2nd Item extreme	(62.44) 59	(100.20) 97	(57.91) 50	(96.68) 106	(86.61) 86	(94.16) 100
"Error Free"	(61.56) 65	(98.80) 102	(57.09) 65	(95.32) 86	(85.39) 86	(92.84) 87

X^2 = 10.624 d.f. = 11

TABLE 15

DISTRIBUTION OF INDIVIDUALS BY TWO TESTS--ONE TEST WITH SECOND ITEM DIFFICULTIES NEARER TO 50 PER CENT LEVEL THAN THE "ERROR FREE" TEST*

Test	Rank Scores					
	64	58-63	54-57	46-53	40-45	33-39
2nd Item near 50	(67.24) 68	(104.14) 104	(73.81) 81	(82.40) 77	(88.47) 89	(85.94) 83
"Error Free"	(65.76) 65	(101.86) 102	(72.19) 65	(80.60) 86	(68.53) 86	(84.06) 87

X^2 = 10.624 d.f. = 11

*NOTE: The rank scores are broken to make approximately equal intervals on the ability scale. The scores 1-32 are not reported in the table but are symmetrical about 32.5. All values were used in the calculations of chi-square. Expected cell frequencies are given in parentheses.

items nearer to 50 per cent level of difficulty had least variance of ability represented in the top 8.4 per cent of the score distribution. Rationale was not supported here.

The second null hypothesis was that there would be equal variances of test scores for individuals in ability category 15 on the "error free" test and the tests which had the fifth item too far and too near the mean value. The results in Table 17 show that the null hypothesis must be accepted. The largest variance was for the test with items nearer the 50 per cent level of difficulty as was hypothesized even though the results were not significant due to the assumptions of only 1000 individuals.

The third null hypothesis was that there would be equal variances of ability level scores for the individuals ranked in the middle 10 per cent of the score distribution by the three tests. The results of these tests are shown in Table 16. The test with the items at the fifth stage near the 50 per cent level of difficulty had lower variance than other tests, but not significantly so. It was hypothesized from Lawley's work on the cumulative that the test with the difficulties away from the mean would have had the smallest variance. Rationale was not supported.

The fourth null hypothesis was that there would be equal variances of test scores for individuals in ability category 8 on all three tests. Again the null hypothesis had to be accepted. The lowest variance was for the test with

TABLE 16

ANALYSIS OF THE VARIANCE OF ABILITY LEVEL SCORES FOR INDIVIDUALS AT SPECIFIED SCORE LEVELS FOR ONE "ERROR FREE" TEST AND TWO "ERROR IN DIFFICULTIES OF FIFTH ITEMS" TESTS

Score Level	"Error Free" Test	5th Items Nearer 50%	5th Items Away from 50%	Significance of Differences
Top 8.4 %	2.30	2.25	2.33	n.s.
Middle 10%	3.08	3.06	3.30	n.s.

TABLE 17

ANALYSIS OF THE VARIANCE OF RANK SCORES FOR INDIVIDUALS AT SPECIFIED ABILITY LEVELS FOR ONE "ERROR FREE" TEST AND TWO "ERROR IN DIFFICULTIES OF FIFTH ITEMS" TESTS

Score Level	"Error Free" Test	5th Items Nearer 50%	5th Items Away from 50%	Significance of Differences
15	148.08	173.34	129.01	n.s.
8	5.57	4.75	6.67	n.s.

the fifth item nearer the 50 per cent level of difficulty. It had been hypothesized that the "least squares" would have the smallest variance.

Errors in Estimating Precision

Two tests were built to examine the error of estimating precision: one with r_{bis} equal to .71 items at the second stage of the "least squares" (r_{bis} = .75) test, and the other with r_{bis} equal to .71 items at the fifth stage. These are discussed separately.

Errors at the second stage.--The first null hypothesis was that there would be equal variances of test scores for individuals in ability category 15 for the "error free" test and the test where the precision was lowered at the second stage. The variances of the "error free" and "error" tests were 5.57 and 5.94, respectively, for ability category 15. The F ratio was 1.06 and thus the null hypothesis had to be accepted. The variance increased with error as was expected, but not to a significant degree if only 1000 individuals were assumed to have taken the test.

The second null hypothesis was that there would be equal variances of test scores for individuals in ability category 8 for the "error free" test and the test where precision was lowered at second stage test. The variance of the "error free" test was 148.08 and for the "error" test was 149.09. The F ratio was 1.01 and again the variance increased as was

hypothesized, but not significantly so if an N of 1000 was assumed. It should be noted that the F ratio for the variances at ability category 15 was greater than the F ratio for variances at level 8--i.e., errors at the second stage seemed to have a greater effect on extreme scores as was anticipated.

The third null hypothesis was that there would be equal variances of ability level scores for individuals ranked in the top 8.4 per cent of the score distribution of each of these two tests. The variance of ability level scores for top 8.4 per cent on the "error free" test was 2.30 and the variances of ability level scores from the "error" test was 2.37. The null hypothesis had to be accepted, but the variance did increase with error in item precision. Again significance depended upon the value assumed for N.

Errors at the fifth stage.--The first null hypothesis was that there would be equal variances of test scores for individuals in ability category 15 for the "error free precision" test and the test with "error" in precision at the fifth stage. The "error free" test had a variance of test scores of 5.57 and the "error" test had a variance of 5.71 for ability category 15. The null hypothesis had to be accepted, but the variance was larger for the test with errors as had been hypothesized. (Changes in assumption of N would change significance test.)

The second null hypothesis was that there would be equal variances of ability level scores for individuals ranked in the top 8.4 per cent of the score distribution by the two tests. The "error free" test had a variance of 2.30 and the "error" test had a variance of 2.40. Again the null hypothesis was accepted; but the variance of the test with the error was larger as hypothesized.

It had been assumed that at the middle ability level the effects of errors in precision would be slight The variance for the "error free" test was 148.08 and for the "error" test was 150.69. The difference between the variances of the two tests is slight and the F ratio for the middle ability variances is the same as the F ratio of variances for category 15 individuals. This was as expected.

It had also been assumed that errors in precision at the second stage would be more serious than those at the fifth stage. In the variances of scores for high ability individuals, the error in precision at the second stage increased variance more than error in precision at fifth stage. (The variances were 5.57, 5.94, and 5.71 for "error free," error at second, and error at fifth stage precision tests, respectively.) However, the variance of scores for middle ability level individuals was higher for the test with error in the fifth stage than the test with error in the second stage. (The variances were 148.08, 149.09, and 150.69 for "error free," error at second, and error at fifth

stage precision tests, respectively.) The error at the fifth stage test also had the highest variance of ability level scores for individuals in top 8.4 per cent of the score distribution. (Variances were 2.30, 2.37, and 2.40 for "error free," error at second, and error at fifth stage precision tests, respectively.) The hypotheses that errors in second stage would be more serious than errors in fifth stage was not confirmed.

General Comparisons

The three areas of general comparisons were effects of difficulty, effects of precision, and effect of pattern of items. There were no hypotheses made about these general comparisons. The information is presented to suggest new hypotheses and to aid in forming tentative conclusions.

Effects of difficulty.--In addition to the hypothesis testing material already reported, examination of Tables 18, 19, and 20 yields information on difficulty. Only the difficulty of the items has changed from column to column within any one of the three tables. It should be noted that the distribution of difficulties to form a normal output of scores yielded the highest mean ability level for the top score, no matter what type of distribution was input. Also, the distribution of difficulties to produce a U-shaped output of scores yielded the greatest number of individuals in the extreme score irrespective of the type of distribution that

TABLE 18

DISTRIBUTION AND MEAN ABILITY LEVEL SCORES FOR TOP SCORE VALUES FOR THREE TESTS WITH DIFFERENT DIFFICULTIES AND NORMAL DISTRIBUTION OF ABILITY INPUT

Rank*	Expected Pattern of Output						Rank*	Expected Pattern of Output					
	Normal		Rectangular		U-Shaped			Normal		Rectangular		U-Shaped	
Score	N	Mean	N	Mean	N	Mean	Score	N	Mean	N	Mean	N	Mean
64	29	14.5	71	13.9	101	13.6	48	21	10.1	11	9.6	10	9.5
63	16	13.7	24	12.6	26	11.9	47	6	10.0	12	9.6	11	9.4
62	20	13.4	25	12.3	27	11.8	46	13	9.8	13	9.5	11	9.4
61	23	13.1	27	12.1	27	11.6	45	16	9.7	12	9.4	11	9.3
60	25	12.8	26	11.9	26	11.4	44	10	9.5	11	9.2	10	9.2
59	21	12.3	24	11.6	24	11.3	43	14	9.4	7	9.1	8	9.0
58	9	12.1	16	11.2	19	11.0	42	12	9.2	8	8.9	7	8.6
57	13	11.7	10	10.8	10	10.3	41	16	9.0	9	8.8	8	8.5
56	9	11.6	12	10.6	11	10.1	40	12	9.0	10	8.6	8	8.3
55	13	11.4	12	10.4	11	10.0	39	10	8.9	11	8.5	9	8.3
54	17	11.2	13	10.3	12	10.0	38	18	8.7	9	8.5	8	8.3
53	18	10.9	14	10.2	12	9.8	37	6	8.5	11	8.4	9	8.2
52	18	10.8	14	10.1	12	9.7	36	13	8.4	11	8.3	9	8.1
51	17	10.6	13	10.0	12	9.7	35	17	8.4	12	8.2	8	8.1
50	21	10.4	15	9.9	12	9.6	34	17	8.3	11	8.1	9	8.1
49	11	10.1	14	9.7	12	9.6	33	17	8.0	11	8.0	9	8.0

*Rank of the mean ability level of individuals at a score.

TABLE 19

DISTRIBUTION AND MEAN ABILITY LEVEL SCORES FOR TOP SCORE VALUES FOR THREE TESTS WITH DIFFERENT DIFFICULTIES AND RECTANGULAR DISTRIBUTION OF ABILITY INPUT

Rank* Score	Expected Pattern of Output: Normal N	Normal Mean	Rectangular N	Rectangular Mean	U-Shaped N	U-Shaped Mean	Rank* Score	Normal N	Normal Mean	Rectangular N	Rectangular Mean	U-Shaped N	U-Shaped Mean
64	40	14.5	100	14.0	139	13.8	48	18	10.5	9	10.1	8	9.8
63	23	13.8	31	13.0	30	12.4	47	5	10.4	10	9.9	8	9.8
62	27	13.6	31	12.8	31	12.3	46	11	10.2	10	9.8	9	9.7
61	31	13.4	32	12.6	29	12.2	45	13	10.1	10	9.7	8	9.6
60	32	13.1	30	12.4	29	11.9	44	9	9.9	8	9.4	8	9.5
59	26	12.7	26	12.1	25	11.8	43	12	9.8	5	9.3	7	9.3
58	11	12.5	16	11.7	19	11.6	42	9	9.5	6	9.2	5	8.8
57	16	12.4	10	11.3	9	10.7	41	12	9.2	7	9.0	6	8.5
56	10	12.1	11	11.0	10	10.5	40	9	9.2	7	8.7	6	8.4
55	14	11.9	11	10.9	10	10.4	39	8	9.1	8	8.6	8	8.4
54	18	11.8	12	10.8	10	10.4	38	14	8.8	7	8.6	6	8.4
53	18	11.5	12	10.6	10	10.2	37	4	8.7	8	8.5	7	8.3
52	18	11.3	12	10.5	10	10.2	36	10	8.5	8	8.3	6	8.2
51	16	11.1	12	10.4	10	10.1	35	13	8.4	8	8.3	6	8.2
50	19	11.0	13	10.3	10	10.0	34	13	8.3	8	8.2	7	8.1
49	10	10.6	12	10.1	10	9.9	33	12	8.0	8	8.0	6	8.0

*Rank of the mean ability level of individuals at a score.

TABLE 20

DISTRIBUTION AND MEAN ABILITY LEVEL SCORES FOR TOP SCORE VALUES FOR THREE TESTS WITH DIFFERENT DIFFICULTIES AND U-SHAPED DISTRIBUTION OF ABILITY INPUT

Rank* Score	Expected Pattern of Output: Normal N	Normal Mean	Rectangular N	Rectangular Mean	U-Shaped N	U-Shaped Mean	Rank* Score	Normal N	Normal Mean	Rectangular N	Rectangular Mean	U-Shaped N	U-Shaped Mean
64	71	14.7	158	14.4	205	14.2	48	14	11.0	6	10.6	6	10.4
63	35	14.2	37	13.5	32	13.0	47	4	10.8	7	10.3	6	10.0
62	38	14.0	36	13.3	32	12.9	46	8	10.7	7	10.2	6	10.0
61	42	13.9	35	13.2	30	12.8	45	10	10.5	6	10.0	6	9.9
60	40	13.6	32	13.0	27	12.6	44	6	10.3	6	9.8	5	9.8
59	30	13.3	26	12.7	23	12.5	43	8	10.2	4	9.6	4	9.4
58	11	13.0	15	12.4	17	12.2	42	6	9.8	4	9.4	4	9.0
57	14	12.7	8	11.8	7	11.2	41	8	9.6	4	9.4	4	8.7
56	10	12.7	9	11.6	7	10.9	40	6	9.6	5	9.0	4	8.6
55	13	12.5	9	11.4	7	10.9	39	5	9.3	4	8.7	4	8.5
54	16	12.4	9	11.3	7	10.8	38	8	9.0	4	8.7	4	8.4
53	16	12.0	9	11.2	7	10.7	37	3	8.9	5	8.5	4	8.4
52	15	11.8	9	11.0	7	10.6	36	8	8.6	5	8.3	4	8.3
51	13	11.6	8	11.0	7	10.6	35	6	8.5	5	8.3	4	8.2
50	15	11.5	9	10.8	7	10.6	34	8	8.4	5	8.3	4	8.0
49	7	11.1	8	10.6	7	10.4	33	8	8.1	5	8.0	4	8.0

*Rank of the mean ability level of individuals at a score.

was input. In general, it can be seen that the number of people at each score is controlled by the distribution of difficulties. A U-shaped input of ability with difficulties calculated to yield a normal distribution of scores did yield scores with more people at the extremes. However, not as many were assigned the extreme score values by the test designed for normal output of scores as were assigned by tests designed for rectangular or U-shaped distribution of scores.

Changes in difficulty made no difference on the value for eta (ability predicted from scores). For a normal distribution of ability into any set of difficulties, the value for eta was .89; for a rectangular distribution of ability into any set of difficulties, the value for eta was .92; and for a U-shaped distribution of ability into any of the three sets of difficulties, the value for eta was .95.

Effects of precision.--In addition to the hypotheses about precision, examination of Table 21 yields information about the numbers and types of individuals at each score value when the precision (and difficulty) of a test is changed. As may be noted in Table 23 of the Appendix, the difficulties of items for each test were quite different. However, the number of people at each score level remained relatively constant. It is thus obvious that the test with more precise items may effectively use a greater range of difficulties than the less precise item tests, and still produce similar distribution

TABLE 21

DISTRIBUTION AND MEAN ABILITY LEVEL SCORES FOR TOP SCORES OF TESTS WITH DIFFERENT LEVELS OF PRECISION AND WITH AN INPUT OF NORMAL DISTRIBUTION OF ABILITY

Rank* Score	Level of Precision									
	.45		.60		.71		.75		.79	
	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
64	23	13.3	27	14.0	28	14.4	29	14.5	29	14.6
63	16	12.2	16	13.0	17	13.5	16	13.7	16	13.9
62	17	12.1	18	12.9	19	13.3	20	13.4	20	13.6
61	18	12.0	20	12.7	22	13.0	23	13.1	24	13.2
60	20	11.9	22	12.4	24	12.7	25	12.8	26	12.8
59	20	11.7	21	12.2	22	12.3	21	12.3	21	12.2
48	18	11.5	15	11.8	10	11.8	9	12.1	8	12.2
57	13	10.5	11	11.4	11	11.7	13	11.7	12	11.8
56	14	10.4	13	11.1	13	11.6	9	11.6	7	11.5
55	14	10.4	14	10.9	13	11.3	13	11.4	13	11.4
54	15	10.3	15	10.8	16	11.2	17	11.2	18	11.3
53	15	10.3	16	10.7	17	10.9	18	10.9	19	11.0
52	16	10.2	17	10.5	18	10.7	18	10.8	19	10.8
51	15	10.2	16	10.4	17	10.6	17	10.6	17	10.6
50	16	10.1	18	10.3	20	10.4	21	10.4	22	10.5
49	15	10.0	14	10.1	20	10.1	11	10.1	5	10.2
48	16	9.9	18	10.1	12	10.1	21	10.1	22	10.1
47	15	9.8	15	9.9	14	9.8	6	10.0	10	10.1
46	16	9.7	16	9.9	16	9.8	13	9.8	12	9.8
45	16	9.7	15	9.7	7	9.7	16	9.7	10	9.7
44	15	9.5	14	9.3	15	9.5	10	9.5	16	9.6
43	13	9.4	9	9.3	11	9.4	14	9.4	14	9.4
42	12	8.7	12	9.0	13	9.1	12	9.2	11	9.3
41	13	8.6	10	9.0	12	9.1	16	9.0	16	9.2
40	14	8.4	13	8.8	15	8.9	12	9.0	10	9.0
39	14	8.4	14	8.7	11	8.8	10	8.9	11	8.9
38	14	8.3	12	8.6	7	8.6	18	8.7	20	8.7
37	15	8.2	16	8.5	17	8.6	6	8.5	13	8.5
36	15	8.2	14	8.2	17	8.3	13	8.4	18	8.4
35	14	8.2	15	8.2	14	8.3	17	8.4	18	8.3
34	15	8.1	15	8.2	17	8.2	17	8.3	4	8.3
33	15	8.0	15	8.0	16	8.0	17	8.0	18	8.0

*Rank of the mean ability level of individuals at a score.

if only the number of people at each score level is considered. However, the mean of ability levels for each score indicates that as precision increases the extreme scores do represent individuals who are more extreme in ability.

The precision of the item also effects the value obtained for eta (ability levels predicted from scores). Eta has the values of .68, .81, .87, .89, and .91 for biserials of .45, .60, .71, .75, and .79, respectively. This was as expected.

Effect of pattern.--Table 22 reports the distribution and mean ability level scores when a normal distribution of ability is input into two tests that have exactly the same difficulty of items, but have these items taken in a different pattern. The difficulties are those reported in Table 23 of the Appendix, for the Lord rectangular. The two patterns are those for the "least squares" and the Lord rectangular. The differences in the two patterns were not great and neither was the difference in the score distributions. The eta (ability level predicted from scores) for both tests was .89.

TABLE 22

DISTRIBUTION AND MEAN ABILITY LEVEL SCORES FOR TOP SCORES OF TESTS WITH DIFFERENT PATTERNS OF ITEMS ENCOUNTERED AND WITH AN INPUT OF A NORMAL DISTRIBUTION OF ABILITY

Rank Score	Pattern of Items (1)* N	Mean	(2)** N	Mean	Rank Score	Pattern of Items (1)* N	Mean	(2)** N	Mean
64	72	13.9	71	13.9	48	12	9.9	11	9.6
63	20	12.5	24	12.6	47	13	10.0	12	9.6
62	20	12.5	25	12.3	46	10	9.6	13	9.5
61	22	12.5	27	12.1	45	11	9.4	12	9.4
60	17	12.2	26	11.9	44	12	9.2	11	9.2
59	15	11.9	24	11.6	43	17	9.2	7	9.1
58	10	11.6	16	11.2	42	14	8.9	8	8.9
57	13	11.0	10	10.8	41	11	8.9	9	8.8
56	14	11.0	12	10.6	40	15	8.9	10	8.6
55	13	10.8	12	10.4	39	12	8.5	11	8.5
54	17	10.8	13	10.3	38	12	8.6	9	8.5
53	15	10.6	14	10.2	37	11	8.7	11	8.4
52	13	10.5	14	10.1	36	11	8.6	11	8.3
51	16	10.2	13	10.0	35	12	8.4	12	8.2
50	12	10.3	15	9.9	34	13	8.5	11	8.1
49	10	10.2	14	9.7	33	13	8.1	11	8.0

*(1) Pattern used in "least squares" solution.

**(2) Pattern used in all other sequential tests.

CHAPTER V

CONCLUSIONS

The conclusions are separated into two sections. The conclusions related to testing problems as discussed in "Need for Test Improvements" in Chapter I are stated in the first section. The conclusions as to the three major hypotheses are presented in the second section.

I. SEQUENTIAL TESTING AND TESTING PROBLEMS

Efficiency of Items

The lack of a monotonic relationship between the reliability of items and the validity of scores produced has been remarked upon by many authors as has the problem of the definition of validity. If one uses eta for the prediction of ability levels from the scores, one finds the value of eta steadily increasing from .68 to .91 as r_{bis} increased from .75 to .79. The more precise the items the higher the correlation between the ability level input and the resulting scores. If one uses the variance of scores assigned to a given ability level as measure of validity (see Table 12), the variance decreased for every ability level as the precision of the item increased. If one uses

the discrimination index as the measure of efficienty (see Table 13), the index values increased with increases in precision.

Rather than using the cumulative test as a comparison with the sequential to determine the efficiency of use of items, it would probably be better to use Lord's work with the discrimination index as a guide to the efficiency of use of items. There has been no attempt here to find and compare the sequential test with the best possible cumulative test. The expected value for the discrimination index was determined (using an assumed r = .91, because this was highest value for r used in the study) by calculating the predicted score value for ability levels 8 and 9. These values were subtracted from each other and the difference divided by an estimate of the standard error of scores for a given ability level (estimated from the correlation). The predicted score values for ability level 8 and 9 were 32.5 and 37.1, respectively. The estimate of the standard error of scores for a given ability level (with r = .91) was 7.95. This yielded a discrimination index of .59. However, if the calculations had been using an assumed r = .81, the discrimination index would have been only .25.

Using .59 as the expected value of the discrimination index for r_{bis} = .79, and .25 as the expected index when r_{bis} = .45, one can compare these values with those actually obtained for different ability levels in Table 13. One notes

that the sequential test discriminated better than expected at the extreme values, and never dropped much below the expected value. Lord has shown that the usual cumulative test (with items all at the 50 per cent level of difficulty) has the highest discrimination at middle score values and is much more discriminating than tests with a spread of item difficulties.[1] He has also shown that even cumulative tests with a spread of item difficulties discriminated better in the middle ability areas than at the extremes of ability level. No conclusions can be drawn as to the efficiency of the sequential compared with the cumulative test for there has been no comparable study of the cumulative test model. However, the fact that the sequential has here been shown the better discriminator for extreme ability levels may be of potential value for test constructors.

It is obvious that efficiency of items must be determined by the use of the test. The sequential tests of different precision levels had difficulties selected by Lord's formula which assumed an output of a normal score distribution at each stage. A sequential test could be built similar to the one built by the "least squares" method except that the difficulty would be that difficulty which would maximize the discrimination index for the divisions that were most

---

[1]Frederic M. Lord, A Theory of Test Scores. Psychometric Monograph No. 7 (Chicago: University of Chicago Press, 1952).

desired. There are many more possibilities, but these are not investigated in this paper.

Control of the Score Distribution

The sequential test can be constructed to yield any type of score distribution. However, the one test built with "least squares" for a rectangular input of ability did not seem to control the score distribution as well as might be expected. (See Table 26.) A normal input of ability into the "least squares" model did not have an automatic normal output of scores, and a U-shaped input into this model did not yield a similar U-shaped output. However, since there are 64 scores for the sequential test, any number of these could be combined to yield the type of score distribution that one wished. One logical combination of scores might be to combine score categories until each separate combination represented equal intervals on the ability level scale. However, the stability of these combinations would need to be investigated before any conclusions could be reached.

The "least squares" sequential did yield a distribution that was somewhat like the input distribution of ability as can be seen in Table 26. The cumulative also yielded a distribution that was somewhat like the input distribution as can be seen in Table 25. However, the important fact is not the number of people at each score, but the type of people at each score.

As can be seen from Table 8, when the U-shaped distribution was input into the sequential the top ability level individuals were assigned to the higher "T" score values than when the U-shaped distribution was input into the cumulative test. The same was true for a normal distribution as is shown in Table 2. Thus, this sequential did control the top ability level better than did the cumulative. Tables 1 and 7 indicate that this conclusion may be valid for average ability level individuals also. However, it should be remembered that this is a single case of the sequential and cumulative test. It can only be concluded that a single sequential test can control several ability distributions--that is not to say that it is better than the cumulative at controlling the ability level distribution.

In addition to the comparison of the "least squares" sequential with the cumulative, it is interesting to compare the "least squares" test with the "Lord rectangular" test--its closest counterpart. These two tests yielded very similar results as can be seen in Table 22. If the "Lord rectangular" test should continue to perform as well as the "least squares" test, then it would be concluded that the "Lord rectangular" procedure would be preferable because of the comparative convenience of calculating the difficulties and pattern of items taken. Further investigation would be needed to determine if this were a valid conclusion.

Meaning of a Score

By comparing Tables 5 and 6 it can be seen that what the score represents is not constant but depends upon the distribution of ability of those taking the test. With a U-shaped distribution of ability input, the top score represented a higher ability level than it did when a normal distribution of ability was input into the test model. This was true for both the sequential and the cumulative tests.

If for a normal distribution of ability one were to combine the sequential scores into score categories such that each category represented a unit of ability, it should be noted that the scores which would be combined to make up a category equivalent to an ability unit would be different than those that would be similarly combined for a U-shaped distribution of ability. This indicates that scores have to be interpreted in terms of the distribution of ability that is likely to be encountered. This was true for both the sequential and cumulative tests. It had been hoped that the "least squares" sequential would have been more stable, thus an equal interval on ability scale could have been set up and used over very different inputs.

It can be concluded that if a normal distribution of ability is input into sequential and cumulative tests comparable to those in this study, that the normalized scores assigned by the sequential to the highest ability level will

be closer to criterion value than will be the normalized scores assigned by the cumulative test. (See Table 24.)

Also, it can be concluded that the "least squares" sequential and the cumulative (but especially the sequential) better control the output of scores for a normal distribution of ability than they control scores for a U-shaped distribution of ability. Table 25 shows that the top ability level individuals came closer to being assigned the criterion score of 69 in the case of the normal distribution than in the case of the U-shaped distribution.

It should be remembered that these conclusions are based upon the assumption that one wishes to reflect ability level rather than rank individuals. Judging from the values for eta (ability levels predicted from scores) the U-shaped distribution was the best ranked distribution by both sequential and cumulative tests. The value of eta reached .95 when a U-shaped distribution of ability was used with all of the $r_{bis} = .75$ sequential and cumulative models studied here. The input of a normal ability distribution ($r_{bis} = .75$) yielded an eta of .89 for all sequential models, and an eta of .88 for the cumulative model. This cannot be attributed to the large number at one value, because the rectangular input likewise had a lower value for eta on all tests than did the U-shaped distribution of ability. The actual ranking of individuals was not investigated in this paper.

II. SEQUENTIAL TESTING HYPOTHESES

The three major hypotheses were (1) that different ability distributions would be translated into score distributions that were not too different from the ability distribution; (2) that the more precise tests (with a spread of item difficulties) would produce the more accurate scores; and (3) that errors in estimating the difficulty of an item would have a greater effect at the fifth stage than at the second stage, while errors in estimating the precision of an item would have a greater effect at the second stage than at the fifth stage. The conclusions for each hypothesis are considered separately.

Effect of Ability Distribution

It can be concluded that the "least squares" sequential test did perform better than the cumulative in respect to the variance of scores for top ability level individuals. For both the normal and U-shaped distributions the sequential test had significantly lower variance of scores for these individuals. The "least squares" test did have a lower variance of scores for middle ability level individuals, but not significantly lower when a N of 1000 was assumed.

If one determines the variance of the ability level for individuals assigned the top 8.4 per cent of the scores by the sequential and cumulative tests, the sequential again had significantly less ability level variance.

If one asks the question about the mean normalized "T" scores assigned to each ability level as compared with the "T" score value for that ability level (see Table 24), it can be seen that for a normal input the sequential assigned for the top and bottom three ability levels "T" scores that were close to the "T" score value for the ability level. The cumulative assigned "T" scores for ability levels five through seven and nine through eleven that were more nearly like the ability level "T" score values than did the sequential. The two tests shared honors. For the U-shaped distribution of ability, the cumulative did as well or better than the sequential except at the most extreme scores.

If test efficiency depends upon the value of eta (ability level predicted from scores), the two tests were identical in their ability to perform with a U-shaped distribution (eta = .95), and the sequential was slightly better than the cumulative when a normal distribution of ability was used (eta = .89 and .88 for the sequential and cumulative, respectively).

Thus, the effectiveness of the test also depends upon the criterion used. However, the sequential seemed to compare favorably with the cumulative. Considering the amount of work needed to prepare the sequential and considering the fact that this cumulative may not be the best possible cumulative, one must determine for himself if the sequential is worthwhile.

Effect of Precision and Difficulty

Conclusions must be made about the joint effect of precision and difficulty as there is no reasonable way to separate these in a sequential test. It must be concluded that precision (used with appropriate difficulties) was an important variable as it resulted in lowering the variance of scores for ability levels (see Tables 10 and 12), and lowering the variance of ability levels for scores (see Table 11).

The contribution of difficulty as a separate factor is not clear. If one uses eta as the criterion, the difficulty of items seemed to make no difference. The three tests in Table 18 differed as to difficulty of items, but did not differ as to the value for eta. The same is true for Tables 19 and 20. However, the number of people at each score and the mean ability level of individuals at a score did change, so one cannot conclude that difficulty does not have an effect.

Another clue as to the contribution of difficulty can be obtained from Tables 16 and 17. In Table 16 the top 8.4 per cent of the score distribution and middle 10 per cent of the score distribution had lower variances of ability level scores when the fifth items were nearer the 50 per cent level of difficulty. This might lead one to believe that Lawley's suggestion that for a cumulative test if the items were near the 50 per cent level of difficulty then the top scores would be more precise was correct. This is not the case, however,

for a sequential test, because if one compares the top score (on the normal distribution of ability) for tests with difficulties calculated so as to yield normal and U-shaped distributions of scores, the variances of ability levels are .88 and 2.40 for the normal and U-shaped difficulty distributions, respectively. The larger variance for the top score came from the test with the item difficulties toward the 50 per cent level.

The above results can probably be explained from Bayes' Theorem. Lawley's work was with a cumulative test where everyone took every item. In the sequential test only those considered to be of high ability took the difficult items. This thus made Lawley's work not directly applicable to the sequential test.

However, it is noted that using items which are nearer the 50 per cent level of difficulty did produce more precise scores. The same change also produced lower variances of scores for middle ability individuals, but the variance of scores for extreme ability people was increased (see Table 17). Thus the more items one has appropriate for the ability level of the individual being classified, the better one can discriminate between these ability levels.

The above conclusion was supported by the test with the fifth item away from the 50 per cent level of difficulty. The more difficult items increased the precision of scores for high ability individuals and decreased the precision for the middle ability individuals (see Table 17).

The general conclusion thus seems to be that one should use more difficult items to distinguish among the more able students.

Effect of Error in Parameters

The data lead to the conclusion that errors in parameter estimates for the "least squares" test did not seem to be too serious. In actual practice one error may tend to cancel another. (In this study, the errors were not made so as to cancel one another.) If one examines the score distribution (see Tables 27 and 28) it can be seen that there was very little shift in the number of people at each score, and little change in the mean of ability level of individuals at the score.

If eta is used as the criterion for the effect of error, this same conclusion is reached. The value for eta was .89 no matter whether the error introduced into the "least squares" sequential test was an error in estimating the difficulty of the item or an error in estimating the precision of the item.

CHAPTER VI

SUMMARY AND RECOMMENDATIONS

The first section of this chapter reviews the different test models constructed and the reasons for their construction, and then reports the conclusions reached from the data analyses. The second section lists the limitations of the study and recommendations for the future study of this and other related problems.

I. SUMMARY

To evaluate the sequential testing procedure, the contribution of the sequential procedure to the solution of test construction problems was examined. An attempt was made to determine how well the sequential test could classify individuals of different ability levels and which parameters seemed to be related to this ability. The effect of error in estimating these parameters was also considered. Adequate control of all variables for this evaluation was obtained by the construction of test models which used hypothetical populations.

The probability of passing a given item in each test was calculated from the ability level of the individual, the difficulty of the item, and the precision of the item.

The probability of passing a sequence of items was determined for each ability level by multiplying together the probabilities of passing or failing a sequence of six-items. Sixty-four differenct sequences were calculated for each of 15 ability levels with 100 hypothetical people at each of the ability levels.

Using the above procedure one cumulative test was constructed with all items at the 50 per cent level of difficulty and with the precision of $r_{bis} = .75$. This was the only cumulative test model used.

Using this same procedure one sequential test was constructed with a precision of $r_{bis} = .75$ and difficulties such that the sum of the squared deviations of the individual's ability level from the mean ability level of the group (i.e., of those who had passed or who had failed the item) would be a minimum. This test was the only sequential test constructed with difficulties computed in this manner.

With the above cumulative and sequential tests, normal and U-shaped distributions of ability were input. The score distribution from the cumulative and sequential tests were then examined. The variance of ability levels of the individuals at a given score value was obtained as one measure of efficiency of the test. The variance of the scores assigned to the individuals of a given ability category was the other measure of efficiency. It was hypothesized that the sequential test would have a lower variance of the scores assigned to

individuals at the highest category of ability than the cumulative test; however, it was hypothesized that the cumulative test would have a lower variance of ability levels of the individuals at the highest score value than would the sequential test.

The conclusion reached was that regardless of the distribution of ability input the sequential test more accurately classified the individuals at extreme ability levels than did the cumulative test. These extreme ability individuals had a variance of 3.87 and 1.96 on the sequential test scores, and a variance of 6.77 and 3.62 on the cumulative scores for a normal and U-shaped distribution of input, respectively. Also, for the sequential test, the normalized "T" scores for extreme ability individuals were more nearly like the criterion "T" score for the extreme ability level than was the case with the cumulative test. However, the "T" scores assigned by the cumulative test for middle ability levels more nearly approximated the criterion middle ability "T" scores.

In addition to the tests constructed for the comparison of the cumulative and the sequential test, six sequential tests were constructed to examine the sequential test output in relation to the parameter values. These tests varied from the above sequential test in that two of them had the difficulty of the second item changed, two had the difficulty of the fifth item changed, one had the precision of the second

item changed and the last had the precision of the fifth item changed. All tests had a normal input of ability. The score distributions resulting from these tests were examined to determine the effect of errors of difficulty (tests 1-4) and errors of precision (tests 5 and 6). The changes in the number of people at each score, the mean ability level of individuals at each score, the variance of scores for top and middle ability level individuals, and the variances of ability level scores for the top and middle scoring individuals were all insignificantly changed.

The above seven sequential tests were constructed with the "least squares" difficulties. Five additional sequential tests were constructed with the level of precision (r_{bis}) taking values of .45, .70, .71, .75, and .79. In contrast to the "least squares" difficulties, the difficulties of items used here were such that they would provide one additional difficulty level at each subsequent stage. Theoretically the items at each stage would separate individuals normally distributed in ability into a normal output distribution. These values were determined from the use of Lord's work. Each of these tests ofdifferent levels of precision was used only with a normal distribution of ability.

Results from these test models indicated that precision of item was an important factor. When used with appropriate difficulties, the high precision tests had significantly lower variances of scores for top and middle ability level individuals and had significantly lower variance of ability

levels for top and middle score level individuals. The values for the discrimination indices and for eta increased with corresponding increases in precision.

The general conclusion as to selection of item difficulty was that the items should be more difficult if used to distinguish among the more able students, and vice versa. When fifth stage items in the sequential test were made closer to 50 per cent level of difficulty, the variances of ability level scores for both top and middle scoring individuals decreased; however, at the same time the score variance for the middle ability level individuals decreased and the score variance for the high ability level individuals increased. The variance of scores for high and middle ability individuals changed in the opposite direction (high ability decreased, middle ability increased) when the difficulties of the fifth items were moved away from the 50 per cent level. Caution is necessary in making conclusions about the variance of ability level for given scores, because when the top scores were considered for two tests which had the same precision but different difficulties, the larger ability level variance came from the test with the items toward the 50 per cent level of difficulty. It was concluded that this was due to the fact that a smaller ability range of individuals took the item in the sequential. (Range of ability will vary with the precision and difficulty of the preceding item.) From Bayes' Theorem one notes that the probability of high

ability people passing the item must be much higher than the probability of low ability people passing the item if 90 per cent of those taking the item have low ability. In the sequential test the base rate (per cent of people who pass the item) is about 50 per cent, even in the case of the "difficult" item considered above instead of a base rate of 10 per cent as in the case for the cumulative test which uses the above item. (More people must pass an item because of their ability than pass the item because of chance. This is not always true in a cumulative test when 90 per cent of the individuals may guess at the answer.)

For descriptive purposes only, three additional sequential tests were constructed. All had $r_{bis} = .75$ for level of item precision; however, the difficulties varied. Each had item difficulties calculated to use one additional difficulty level at each succeeding stage. (The nth stage had N different difficulties for items at that stage.) One test had difficulties calculated to yield a normal distribution of scores at each stage with the input of a normal distribution of ability at each stage. (This test is the same as the .75 test described above and calculated to test the hypotheses about effect of precision.) The second test had difficulties calculated to yield a rectangular distribution of scores when a rectangular distribution of ability was input. The third test had difficulties calculated to yield a U-shaped distribution of scores when a U-shaped distribution of ability was

input. (These difficulties were quite different as can be seen in Table 23.) Each test was used with a normal, rectangular and U-shaped distribution of ability--i.e., with two distributions of ability other than the one used in the original calculation of item difficulties. The difficulties for a U-shaped score distribution yielded the greatest number of individuals at the extreme score irrespective of the type of ability distribution that was input; and the item difficulties for a normal score distribution yielded the fewest individuals at the extreme score each time.

For a test with item precision of $r_{bis} = .75$ for all items, changes in item difficulty made no difference on the value for eta (ability predicted from score). The normal distribution of ability used with any set of difficulties yielded an eta of .89. The rectangular distribution of ability yielded an eta of .92 for all item difficulties, and likewise the U-shaped yielded an eta of .95. Values thus were dependent upon the type of distribution of ability used with the tests and not upon the type of ability distribution that the test had been designed to reflect.

The last sequential test constructed was one which had the difficulties determined for a rectangular score output and rectangular distribution of ability, but with the pattern of items the same as those for "least squares" sequential test. Item precision remained at $r_{bis} = .75$. The results from this test (used with a normal input of ability) were compared with

a test which had the same set of difficulties and precision level, but which had the arbitrary pattern of one more difficulty level used at each stage. These two tests produced approximately the same number at each score value and had approximately the same mean ability levels for individuals at each score value (see Table 22). Results of the comparison of the "least squares" and the "arbitrary pattern" tests indicate that the latter more rapidly calculated test may yield as useful results as the more exact and time-consuming "least squares" method. While the comparative utility cannot be determined except relative to a specific decision to be made, it would seem advisable to explore further the more easily constructed "arbitrary pattern" test.

II. RECOMMENDATIONS

Results of this study indicate that finding applications of the sequential method is not of first importance. The sequential method for the item precision levels studied in this dissertation does allow better classification of individuals in the extreme ability levels than of those in the middle ability levels. Whereas others have used the sequential method to quickly classify the individuals as not being of middle ability, they have not attempted to better classify individuals who are in the extremes. (It may be that such users of the sequential test are primarily concerned with discriminations at the middle ability level where the majority

of the people are.) In any case, the sequential test has not been used in a manner that takes advantage of its ability to quickly pick out and discriminate among the extreme ability individuals.

If one attempts to build a theory of the sequential test that may be used at some later date, then the design needs to be somewhat different from this paper. The building of sequential tests with different decision rules would seem to be more appropriate than testing the behavior of a single test under varying conditions. This would mean that one would not spend time comparing the sequential and cumulative models until much more was understood about the building of the sequential test.

The construction of an electronic computer program which would select item difficulty and precision should be done in line with specific decision rules such as maximizing the $\Sigma (\Sigma X)^2/N$. One possible scheme might be to assign each individual a score equal to the mean ability level of the group in which he is included; then for each ability level calculate the mean and variance of the score values. From these values the discrimination index between each adjacent ability level could be determined. The difficulty of the item which yielded (1) the highest sum of discrimination indices or (2) the highest discrimination index at a particular point (such as between the middle and adjacent ability levels) would then be selected. Still another scheme would

be to calculate the variances of ability levels for each group and maximize the sum of $(\Sigma X)^2/N$ for each group as was done in this paper, but this time making no restrictions on these difficulty levels of items. It might then be decided that when the variance is below a certain specified level, then the group takes no more items. This would allow the length of test to change to suit the ability level of the individuals about whom the decision must be made.

No recommendation is made to study the psychological effects of sequential testing, the type of distributions of ability one is likely to encounter, or precision of items likely to be found in practice. These are practical questions which need to be answered only after the sequential procedure is more fully understood.

SELECTED BIBLIOGRAPHY

SELECTED BIBLIOGRAPHY

American Educational Research Association, Committee on Test Standards, and National Council on Measurements Used in Education, Committee on Test Standards. Technical Recommendations for Achievement Tests. Washington, D. C.: National Education Association, 1955. 36 pp.

Brogden, Hubert E. "Variation in Test Validity with Variation in the Distribution of Item Difficulties, Number of Items, and Degree of Their Intercorrelation," Psychometrika, 11:197-214, No. 4, December, 1946.

Campbell, Donald T. and Fiske, Donald W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," Psychological Bulletin, 56:81-105, No. 2, March, 1959.

Cleeton, Glen U. "Optimum Difficulty of Group Test Items," Journal of Applied Psychology, 10:327-340, No. 3, September, 1926.

Cook, Desmond L. "A Replication of Lord's Study on Skewness and Kurtosis of Observed Test-Score Distributions," Educational and Psychological Measurement, 19:81-87, No. 1, Spring, 1959.

Cowden, Dudley J. "An Application of Sequential Sampling to Testing Students," Journal of the American Statistical Association, 41:547-556, No. 236, December, 1946.

Cronbach, Lee J. "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, 16:297-334, No. 3, September, 1951.

Cronbach, Lee J. and Warrington, Willard G. "Efficiency of Multiple-Choice Tests as a Function of Spread of Item Difficulties," Psychometrika, 17:127-147, No. 2, June, 1952.

Davis, Frederick B. "The Selection of Test Items According to Difficulty Level," American Psychologist, 4:243, No. 7, July, 1949.

Davis, Frederick B. "Item Analysis in Relation to Educational and Psychological Testing," Psychological Bulletin, 49: 97-121, No. 2, March, 1952.

Ferguson, George A. "On the Theory of Test Discrimination," Psychometrika, 14:61-68, No. 1, March, 1949.

Fiske, Donald W. and Jones, Lyle V. "Sequential Analysis in Psychological Research," Psychological Bulletin, 51: 264-275, No. 3, May, 1954.

Flanagan, John C. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from Data at the Tails of the Distribution," Journal of Educational Psychology, 30:674-680, No. 9, December, 1939.

Glaser, Robert, Damrin, Dora E., and Gardner, Floyd M. "The Tab Item: A Technique for the Measurement of Proficiency in Diagnostic Problem Solving Tasks," Educational and Psychological Measurement, 14: 283-293, No. 2, Summer, 1954.

Glaser, Robert and Schwarz, P. A. "Scoring Problem--Solving Test Items by Measuring Information," Educational and Psychological Measurement, 14:665-670, No. 4, Winter, 1954.

Gulliksen, Harold. "The Relation of Item Difficulty and Inter-item Correlation to Test Variance and Reliability," Psychometrika, 10:79-91, No. 2, June, 1945.

Harris, Frank J., Howell, Margaret A., and Newman, Sidney H. "Forced Choice Tetrads--Effect of Scoring Procedure and Key Length on Validity and Reliability," Educational and Psychological Measurement, 16:454-464, No. 4, Winter, 1956.

Humphreys, Lloyd G. "The Normal Curve and the Attenuation Paradox in Test Theory," Psychological Bulletin, 53: 472-476, No. 6, November, 1956.

Jackson, Robert W. B. and Ferguson, George A. "A Plea for a Functional Approach to Test Construction," Educational and Psychological Measurement, 3:23-28, No. 1, Spring, 1943.

Johnson, M. Clemens and Lord, Frederic M. "An Empirical Study of the Stability of a Group Mean in Relation to the Distribution of Test Items Among Students," Educational and Psychological Measurement, 18:325-329, No. 2, Summer, 1958.

Lawley, D. N. "On Problems Connected with Item Selection and Test Construction," Proceedings of the Royal Society of Edinburgh, 61 (Section A, Part III): 273-287, 1942-1943.

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

Loevinger, Jane. "The Attenuation Paradox in Test Theory," Psychological Bulletin, 51:493-504, No. 5, September, 1954.

Lord, Frederic M. A Theory of Test Scores. Psychometric Monograph No. 7. Chicago: University of Chicago Press, 1952. 84 pp.

_____. "The Relation of Test Score to the Trait Underlying the Test," Educational and Psychological Measurement, 13:517-549, No. 4, Winter, 1953.

_____. "Some Perspectives on 'The Attenuation Paradox in Test Theory'," Psychological Bulletin, 52:505-510, No 6, November, 1955.

_____. "A Survey of Observed Test-Score Distributions with Respect to Skewness and Kurtosis," Educational and Psychological Measurement, 15:383-389, No. 4, Winter, 1955.

_____. "An Empirical Study of the Normality and Independence of Errors of Measurement in Test Scores," Psychometrika, 25:91-104, No. 1, March, 1960.

McNemar, Quinn. Psychological Statistics. Second edition. New York: John Wiley and Sons, 1955. 408 pp.

Maxwell, A. E. "Maximum Likelihood Estimates of Item Parameters Using the Logistic Function," Psychometrika, 24:221-227, No. 3, September, 1959.

Meehl, Paul E. and Rosen, Albert. "Antecedent Probability and the Efficiency of Psychometric Signs Patterns or Cutting Scores," Psychological Bulletin, 52:194-216, No. 3, May, 1955.

Merwin, Jack C. "Rational and Mathematical Relationships of Six Scoring Procedures Applicable to Three-Choice Items," Journal of Educational Psychology, 50:153-161, No. 4, August, 1959.

Michael, William B. "Development of Statistical Methods Especially Useful in Test Construction and Evaluation," Review of Educational Research, 29:106-129, No. 7, February, 1959.

Michael, William B. and Perry, Norman C. "A Theory of Item-Analysis Based on the Scoring of Items at Three Levels of Appropriateness of Response," Educational and Psychological Measurement, 15:404-415, No. 4, Winter, 1955.

Milholland, John E. "The Reliability of Test Discriminations," Educational and Psychological Measurement, 15:362-370, No. 4, Winter, 1955.

Mollenkopf, William G. "Variation of the Standard Error of Measurement," Psychometrika, 14:189-229, No. 3, September, 1949.

Moonan, William J. "Some Empirical Aspects of the Sequential Analysis Technique as Applied to an Achievement Examination," Journal of Experimental Education, 18:195-207, No. 3, March, 1950.

Mosier, Charles I. "Psychophysics and Mental Trait Theory: Fundamental Postulates and Elementary Theorems," Psychological Review, 47:355-366, No. 4, July, 1940.

Richardson, M. W. "The Relation Between the Difficulty and the Differential Validity of a Test," Psychometrika, 1:33-49, No. 2, June, 1936.

Ryans, David G. "An Analysis and Comparison of Criterion Techniques for Weighting Criteria Data," Educational and Psychological Measurement, 14:449-458, No. 3, Autumn, 1954.

Swineford, Frances. "Some Relations Between Test Scores and Item Statistics," Journal of Educational Psychology, 50:26-30, No. 1, February, 1959.

Symonds, Percival M. "Factors Influencing Test Reliability," Journal of Educational Psychology, 19:73-87, No. 2, February, 1928.

"Symposium: Standard Scores for Aptitude and Achievement Tests," Educational and Psychological Measurement, 22: 5-39, No. 1, Spring, 1962.

Thurstone, L. L. "The Scoring of Individual Performance," Journal of Educational Psychology, 17:446-457, No. 7, October, 1926.

Thurstone, Thelma Gwinn. "The Difficulty of a Test and Its Diagnostic Value," Journal of Educational Psychology, 23:335-343, No. 5, May, 1932.

Tryon, Robert C. "Reliability and Behavior Domain Validity: Reformulation and Historical Critique," Psychological Bulletin, 54:229-249, No. 3, May, 1957.

Tucker, Ledyard R. "Maximum Validity of a Test with Equivalent Items," Psychometrika, 11:1-14, No. 1, March 1946.

Tyler, Ralph W. Constructing Achievement Tests. Columbus, Ohio: Bureau of Educational Research, Ohio State University, 1934. 102 pp.

Wald, Abraham. Sequential Analysis. New York: John Wiley and Sons, 1947. 212 pp.

Walker, David A. "Answer-Pattern and Score-Scatter in Tests and Examinations," British Journal of Psychology, 22: 73-86 (July, 1931); 26:301-308 (January, 1936); 30:248-60 (January, 1940).

Wherry, Robert J. and Gaylord, Richard H. "The Concept of Test and Item Reliability in Relation to Factor Pattern," Psychometrika, 8:247-264, No. 8, December, 1943.

White, Benjamin W. and Saltz, Eli. "Measurement of Reproducibility," Psychological Bulletin, 54:81-99, No. 2, March, 1957.

APPENDIX A

TABLE 23

PER CENT PASSING ITEMS OF THE DIFFERENT SEQUENTIAL TESTS CONSTRUCTED

		Least Squares Constructed					Lord Constructed						
			2nd Away	2nd Toward	5th Away	5th Toward	Normal					Rectangular	U-Shaped
Stage	Item	.75	.75	.75	.75	.75	.45	.60	.71	.75	.79	.75	.75
1	1	50	50	50	50	50	50	50	50	50	50	50	50
2	2	41	26	56	41	41	36	31	28	27	25	37	42
	3	59	74	44	59	59	64	69	72	73	75	63	58
3	4	31	31	31	31	31	29	23	19	17	16	31	38
	5	50	50	50	50	50	50	50	50	50	50	50	50
	6	69	69	69	69	69	71	77	81	83	84	69	62
4	7	23	23	23	23	23	25	18	14	13	12	26	35
	8	34	34	34	34	34	41	38	36	35	34	42	46
	9	66	66	66	66	66	59	62	64	65	66	58	54
	10	77	77	77	77	77	75	82	86	87	88	74	65
5	11	19	19	19	10	32	23	16	12	11	10	24	32
	12	25	25	25	14	40	36	31	28	27	25	37	42
	13	44	44	44	29	60	50	50	50	50	50	50	50
	14	56	56	56	71	40	64	69	72	73	75	63	58
	15	75	75	75	86	60	77	84	88	89	90	76	68
	16	81	81	81	90	68							
6	16						21	14	10	09	08	21	30
	17	19	19	19	19	19	32	26	22	21	20	34	40
	18	31	31	31	31	31	44	41	40	39	39	44	47
	19	50	50	50	50	50	56	59	60	61	61	56	53
	20	69	69	69	69	69	68	74	78	79	80	66	60
	21	81	81	81	81	81	79	86	90	91	92	79	70

TABLE 24

MEAN NORMALIZED "T" SCORES FOR EACH ABILITY LEVEL FOR CUMULATIVE AND "LEAST SQUARES" SEQUENTIAL TESTS

Ability Level	Ideal "T" Score	Normal Input		U-Shaped Input	
		Cumulative	Sequential	Cumulative	Sequential
1	30.6	35.5	34.1	39.6	38.2
2	35.1	37.8	36.6	40.8	40.8
3	37.6	38.9	38.7	41.9	42.5
4	40.1	40.8	40.8	43.4	44.3
5	42.6	42.8	43.1	45.1	46.0
6	45.0	45.2	45.4	46.8	47.4
7	47.5	47.5	47.7	48.4	48.8
8	50.0	50.0	50.0	50.0	50.0
9	52.5	52.5	52.3	51.6	51.2
10	55.0	54.8	54.6	53.2	52.6
11	57.4	57.2	56.9	54.9	54.0
12	59.9	59.2	59.2	56.6	55.7
13	62.4	61.1	61.3	58.1	57.5
14	64.9	62.2	63.4	59.2	59.2
15	69.4	64.5	65.9	60.4	61.8

TABLE 25

DISTRIBUTION AND MEAN ABILITY LEVEL SCORES FOR CUMULATIVE TEST WITH THE INPUT OF DIFFERENT DISTRIBUTIONS OF ABILITY

Score	Normal		U-Shaped	
	N	Mean	N	Mean
7	166	12.9	284	13.9
6	139	10.8	119	12.0
5	130	9.3	69	10.0
4	129	8.0	58	8.0
3	130	6.7	69	6.0
2	139	5.2	119	4.0
1	166	3.1	284	2.1

TABLE 26

DISTRIBUTION AND MEAN ABILITY LEVELS FOR TOP SCORES ON "LEAST SQUARES" SEQUENTIAL TEST WITH THE INPUT OF DIFFERENT DISTRIBUTIONS OF ABILITY

	Normal		U-Shaped			Normal		U-Shaped	
Score	N	Mean	N	Mean	Score	N	Mean	N	Mean
64	65	14.0	145	14.4	48	9	10.3	6	11.0
63	24	12.6	37	13.6	47	11	10.0	7	10.9
62	24	12.6	37	13.6	46	9	9.7	6	10.6
61	20	12.4	31	13.5	45	12	9.5	8	10.1
60	15	12.3	21	13.3	44	13	9.2	7	9.9
59	10	12.0	14	13.1	43	16	9.0	8	9.9
58	9	11.9	9	12.9	42	15	9.0	6	9.5
57	18	11.3	17	12.3	41	15	9.0	7	9.4
56	16	11.0	15	12.2	40	15	9.0	7	9.5
55	15	11.0	12	12.0	39	12	8.5	6	9.0
54	16	11.0	14	11.9	38	14	8.9	7	9.1
53	12	10.5	11	11.7	37	12	8.5	6	9.0
52	12	10.5	9	11.7	36	12	8.5	6	9.0
51	12	10.5	8	11.4	35	12	8.5	6	8.8
50	12	10.5	9	11.3	34	12	8.5	6	8.6
49	9	10.3	7	11.4	33	13	8.2	5	8.3

TABLE 27

DISTRIBUTION AND MEAN ABILITY LEVELS FOR TOP SCORES WITH DIFFICULTIES OF CERTAIN ITEMS CHANGED IN A SEQUENTIAL TEST WITH AN INPUT OF NORMAL DISTRIBUTION OF ABILITY

	Stage Changed									
	2nd Away		2nd Toward		5th Away		5th Toward		None	
Score	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean
64	59	14.1	68	14.0	52	14.2	74	13.9	65	14.0
63	20	12.8	26	12.6	17	12.9	29	12.5	24	12.6
62	20	12.8	26	12.6	36	13.0	13	12.4	24	12.6
61	17	12.7	23	12.4	15	12.8	25	12.3	20	12.4
60	13	12.5	17	12.2	11	12.6	19	12.1	15	12.3
59	18	12.1	5	11.7	8	12.3	14	11.8	10	12.0
58	9	11.7	7	11.9	6	12.0	10	11.6	9	11.9
57	14	11.4	22	11.0	25	11.4	12	10.9	18	11.3
56	13	11.3	21	10.9	11	11.3	23	10.8	16	11.0
55	11	11.1	18	10.7	9	11.1	20	10.6	15	11.0
54	12	11.1	20	10.7	23	11.0	10	10.5	16	11.0
53	10	10.9	17	10.5	19	10.8	9	10.4	12	10.5
52	19	10.8	7	10.3	8	10.8	17	10.3	12	10.5
51	18	10.6	7	10.1	16	10.6	8	10.1	12	10.5
50	9	10.6	16	10.2	9	10.6	15	10.2	12	10.5
49	11	10.2	8	10.4	6	10.6	14	10.1	9	10.3
48	11	10.0	8	10.2	13	10.3	7	9.9	9	10.3
47	17	10.3	7	9.9	8	10.2	14	9.9	11	10.0
46	11	9.6	7	10.0	7	10.0	12	9.6	9	9.7
45	10	9.7	20	9.2	19	9.6	11	9.3	12	9.5
44	9	9.5	19	9.0	18	9.4	10	9.0	13	9.2
43	22	9.4	10	9.0	19	9.0	12	9.4	16	9.0
42	19	9.2	9	8.8	17	9.1	10	8.8	15	9.0
41	9	9.2	20	8.8	10	9.2	20	8.8	15	9.0
40	17	8.8	11	9.2	18	8.8	10	9.2	15	9.0
39	15	8.5	9	8.9	15	8.9	10	8.5	12	8.5
38	20	8.9	10	8.5	10	8.9	20	8.5	14	8.9
37	8	9.0	17	8.5	17	8.9	8	8.4	12	8.5
36	7	8.4	19	8.9	15	8.5	9	8.8	12	8.5
35	17	8.2	10	8.6	9	8.7	19	8.2	12	8.5
34	17	8.6	9	8.2	17	8.7	8	8.2	12	8.5
33	16	8.0	9	8.4	16	8.4	8	8.0	13	8.2

TABLE 28

DISTRIBUTION AND MEAN ABILITY LEVELS FOR TOP SCORES WITH PRECISION OF ITEMS CHANGED IN A SEQUENTIAL TEST WITH AN INPUT OF NORMAL DISTRIBUTION OF ABILITY

	Stage Changed					Stage Changed			
	2nd		5th			2nd		5th	
Score	N	Mean	N	Mean	Score	N	Mean	N	Mean
64	63	14.0	63	14.0	48	9	10.1	10	10.2
63	23	12.7	23	12.6	47	12	10.2	11	10.0
62	23	12.7	25	12.8	46	9	9.8	9	9.8
61	20	12.6	20	12.5	45	15	9.4	15	9.5
60	15	12.3	15	12.3	44	14	9.2	14	9.3
59	13	12.2	11	12.0	43	16	9.3	16	9.2
58	8	11.8	8	11.7	42	14	9.1	13	9.0
57	18	11.1	18	11.3	41	15	8.8	15	8.9
56	17	11.0	17	11.0	40	14	8.9	14	8.9
55	14	10.8	14	10.8	39	12	8.7	12	8.7
54	16	10.8	17	10.9	38	15	8.8	15	8.6
53	14	10.6	14	10.7	37	12	8.6	13	8.8
52	13	10.7	12	10.5	36	13	8.8	12	8.6
51	13	10.4	12	10.4	35	14	8.4	14	8.4
50	12	10.3	12	10.3	34	13	8.5	13	8.5
49	10	10.3	10	10.2	33	12	8.1	12	8.3

APPENDIX B

GLOSSARY

Ability--the criterion measure. The term ability is used even though the test could be used to measure attitude or interest. Ability is used to refer to the input variable.

Ability level--a point on the criterion. The term "ability category" is used to designate a range of ability levels.

Arbitrary pattern--pattern chosen by logic of the situation alone not from an underlying theory or on empirical grounds. In the use of Lord's work one must decide where to divide a group and if every possible group is to be considered a separate group. In the pattern used in the sequential tests, other than the "least squares," those who fail a given item and those who pass the next easiest item take the same succeeding item. This pattern has no empirical test in this study.

Category--a class of scores or ability levels. The ability levels or the score values considered in the category must be designated.

Classification--assigning the individuals to two or more categories in an attempt to designate the level of ability of the individuals. Classification usually produces more than the pass or fail categories which are needed for the problem of selection.

Cumulative--test or scoring procedure in which every item is available to the individual taking the examination. Cumulative scoring unless otherwise designated will refer to the counting of the number of correct responses with or without a correction for those incorrect.

Difficulty of items--measure related to the number of people who would be assumed to be able to pass the item. A 50 per cent level of difficulty item is one that 50 per cent of the individuals in the referent group would pass. The standard score form of difficulty always refers to difficulty for the entire group, and is the value below which all those who fail the item would be on a normal curve.

Difficulty level--a point on the difficulty scale.

Difficulty, rule for--states the difficulty of the item considered most appropriate for a given group. The item to be taken by a group is decided by the rule for the arbitrary pattern described above. In this dissertation, if the difficulty is not empirically determined (i.e., by the "least squares" solution), then difficulty is the computed difficulty equal to the standard score for the ability level where the division is to be made regressed by the value of r_{bis}. The higher the r_{bis} the less the value is regressed toward the mean.

Discrimination--ability to rank individuals in the proper order or to classify individuals into a score category that reflects the ability level of the individuals. In this study discrimination is used as the ability to classify individuals into a score category that reflects the ability level. If there are many individuals at the extreme ability level, then these same individuals should be at the extreme score category.

Discrimination index--a measure of discrimination. In this dissertation the discrimination index is one developed by Lord which reflects the classification of individuals into score categories, which in turn reflect the ability levels of the individuals.

Distribution--the manner in which individuals are distributed, including both the number of individuals in a category and the ability levels of the individuals in a category.

Efficiency of test--effective production of test scores that serve a desired function. The two measures of efficiency used in this dissertation are (1) the variance of ability levels assigned to any one score category and (2) the variance of scores assigned to any one ability category. The lower the variance the more efficient the test. Efficiency could also be measured by the value of a product-moment correlation or correlation ratio (eta) between input levels and output scores.

Input--the variable that is to be measured by the test. The term "ability distribution" is used in this dissertation to indicate input distribution. In reality, the input variable could be ability, interest, or attitude. The input variable may also be referred to as the "criterion variable."

"Least squares" test--the test in the dissertation which was constructed with item difficulties such that the sum of the squared deviations of the individual's ability level from the mean ability level of the group into which the individual was classified would be a minimum. However, a restriction was placed on the values the difficulties could take. Even though individuals in each ability category were kept separate from individuals in other categories, individuals in different categories took the same difficulty item if the calculated difficulties were less than .20 standard deviation units apart.

Level--the point on the score (output) distribution or ability (input) distribution which represents the amount of ability (criterion) variable possessed by an individual or group of individuals. The higher the number assigned to a level the more of the variable the individual at that level is presumed to possess.

Output distribution--manner in which scores assigned to individuals are distributed. The output is a measure of the input variable rather than a correlate of the input variable.

Parameter values--arbitrary values assigned to measures of difficulty or precision. These are the two parameters of primary concern in this dissertation.

Parameter errors--values assigned to measures of difficulty or precision that do not agree with the values calculated to fit the "least squares" model.

Pattern of response--sequence of items taken by an individual. The pattern indicates which items the individual answers and whether his answers are correct or incorrect.

Precision of item--measure of how validly the item splits the group into two parts. In this dissertation the calculations use σ_d as the measure of precision. The value of σ_d is directly related to the value of the item-total biserial correlation.

Precision of score--measure of the variance of the ability levels of individuals assigned to a score. A precise score has low variance of ability level for individuals assigned that score--i.e., all of the individuals at the score are at approximately the same ability level.

Reflection--to accurately represent the position of each individual. If the score distribution reflects the input distribution, then the score distribution will have the individuals in the same relative location as the input distribution; i.e., those individuals at one extreme of the input distribution will likewise be at the extreme of the output or score distribution.

Selection--the process of picking out individuals for inclusion or exclusion. This process differs from classification where one needs to determine the level of the individual in order to assign him to one of several categories. In selection one is only concerned with whether an individual is above or below a specified level.

Sequential test--a testing procedure in which the examinees are directed to subsequent items on the basis of their responses to prior items.

Significance of difference--differences should be considered significant only if the difference could not arise by chance from the sampling of identical populations. The value necessary for significance depends upon the number of individuals included in the sample. As there is no sample used in this dissertation,when significant differences are reported they are really an indication of the probable size of the difference relative to the likely error rather than significance in the usual sense.